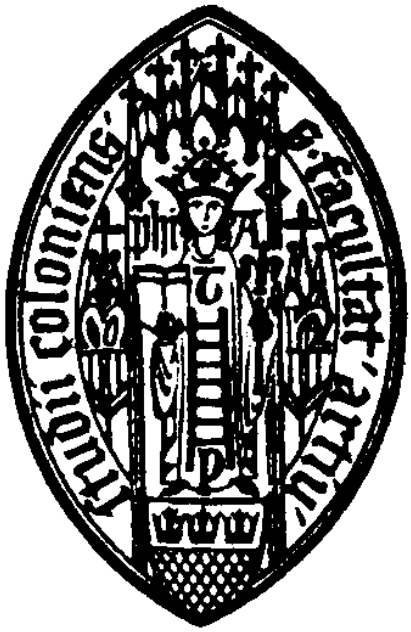
Fitness Landscapes, Adaptation and Sex on the Hypercube



Inauguraldissertation
zur Erlangung des Doktorgrades der
Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

Vorgelegt von

Johannes Neidhart

aus Bergisch Gladbach

Berichterstatter: Prof. Dr. Joachim Krug,
Universität zu Köln

Prof. Dr. Anton Bovier,
Rheinische Friedrich-Wilhelms-Universität Bonn

Tag der mündlichen Prüfung: 26.11.2014

Zusammenfassung

Diese Arbeit beschäftigt sich mit der theoretischen Betrachtung von Fitness-Landschaften im Kontext evolutionärer Prozesse. Diese verknüpfen das Genom eines Organismus mit seiner Fitness; sie sind ein wichtiges Werkzeug der theoretischen Evolutionsbiologie und seit wenigen Jahren auch im Blickpunkt experimenteller Studien. Im Folgenden werden Modelle von Fitness-Landschaften mit analytischen und numerischen Mitteln untersucht, mit der Zielsetzung, charakteristische Eigenschaften zu identifizieren, welche eine Zuordnung experimentell erschlossener Systeme ermöglichen. Desweiteren werden verschiedene adaptive Prozesse betrachtet; zum einen jene, welche mit Mutationen unter Selektion ablaufen, insbesondere sogenannte 'Adaptive Walks'. Zum anderen auch solche, bei denen Rekombination hinzukommt, was die Komplexität der verwendeten Modelle erheblich steigert. Vor allem die entstehende Nichtlinearität der Zeitentwicklung erschwert die analytische Betrachtung, weswegen hier verstärkt auf Computersimulationen zurückgegriffen wird.

Abstract

The focus of this thesis is on the theoretical treatment of fitness landscapes in the context of evolutionary processes. Fitness landscapes connect an organism's genome to its fitness. They are an important tool of theoretical evolutionary biology and in the recent years also experimental results became available. In this thesis, several models of fitness landscapes are analyzed with different analytical and numerical methods. The goal is to identify characteristics in order to compare the model landscapes to experimental measurements. Furthermore, different adaptive processes are examined. On the one hand such which run with mutations under selection, especially adaptive walks. On the other hand such which include recombination. Since these are non-linear in time development, an analytical approach is hindered which leads to an increasing use of computer simulations.

II

Contents

The fact that we live at the bottom of a deep gravity well, on the surface of a gas covered planet going around a nuclear fireball 90 million miles away and think this to be normal is obviously some indication of how skewed our perspective tends to be. Adams [1]

Chapter 1

Introduction

In the recent years the number of physicists working on problems from evolutionary biology grew steadily. Methods from theoretical physics can be applied, analogies found and models built.

This chapter shall give a brief introduction into the field and the used mathematics. Also the connection to theoretical physics will be made by the discussion of spin glasses. In ch. 2 the Rough Mt. Fuji model will be analyzed and certain properties will be calculated which lead to a suggestion of parameters to fit the model to experimental data. In this way it shall help to answer the question, which characteristics do such fitness landscapes have, and how can one tell if they are realistic? To extend this, Ch. 3 will present a way to reduce the number of data points to make statements about fitness landscapes of a certain family. This results in an algorithm which can also help to answer, how models can be fitted to experimental data. After these chapters about static landscapes, it will be asked, how does adaptation behave in certain evolutionary scenarios? Therefore dynamics are introduced and ch. 4 shall contain several analytic and numeric results on adaptive walks, amongst others a phase transition concerning the adaptive walk length in a Rough Mt. Fuji model. Finally, ch. 5 concerns the question: Why is sex? Sexual reproduction and genetic recombination will be discussed and various results presented, most notable perhaps on the transient benefit of sex. A list of some of the used symbols and probability distributions in the appendix (app. A) shall avoid confusion. For completeness, the used algorithms that are not presented in the main text are described in app. B.

Throughout this thesis, the style is chosen to be a mixture of "classical" mathematics literature and "modern" physics literature. This means, that important definitions and analytical results will be given separately with a reference number, and will be connected by running text. Results will usually be followed by a proof. Numerical results will not be stated in

this way but in the form of prose and figures, as these can usually not be formulated as exact and closed and need a description. The intention of this mixed presentation is to simplify reading and mark important passages in the inherent way of mathematical texts without exaggerated rigor. This means, that also calculations and intermediate results will be stated as proofs and results, which might not deserve the name from a mathematicians point of view. Nevertheless this drawback is accepted for the benefit of more clarity.

1.1 Biological foundation & mathematical description

When an organism proliferates, it gives *hereditary information* to its offspring. This information is coded into its genome into a molecule which is called *Deoxyribonucleic acid* (DNA). This molecule is arranged in such a way, that the information is written in an alphabet of four bases, guanine (G), adenine (A), thymine (T) and cytosine (C), and all is shaped in a double helix, where bases are paired, G always with C and A always with T. This information will be copied into every cell of the developing offspring, as it was in the parent organism. A helpful metaphor gives Dawkins:

> It is as though, in every room of a gigantic building, there was a book-case containing the architect's plans for the entire building. The 'book-case' in a cell is called the nucleus. The architect's plans run to 46 'volumes' in a man – the number is different in other species. The 'volumes' are called chromosomes. They are visible under a microscope as long threads, and the genes are strung out along them in order. It is not easy, indeed it may not even be meaningful, to decide where one gene ends and the next one begins. [...] 'Page' will provisionally be used interchangeable with gene, although the division between genes is less clear cut than the division between the pages of a book. [...] Incidentally, there is of course no 'architect'. The DNA instructions have been assembled by natural selection. Dawkins [2, p. 22].

In a simplified picture where all genomes are assumed to be of the same length and also only haploid organisms are present, the genome is written with an *alphabet* $\mathfrak{A}$ of four letters: $\mathfrak{A} = \{A, C, G, T\}$. The genetic information can be written as a sequence $\sigma = (\dots, A, T, C, T, G, \dots)$. All possible sequences form the *sequence space*. Every position of σ is called *locus*, and all elements of $\mathfrak{A}$ which can be present at a locus are called *alleles*. Each realization σ is called a *genotype*. Usually, minor changes to the genome, like a change at

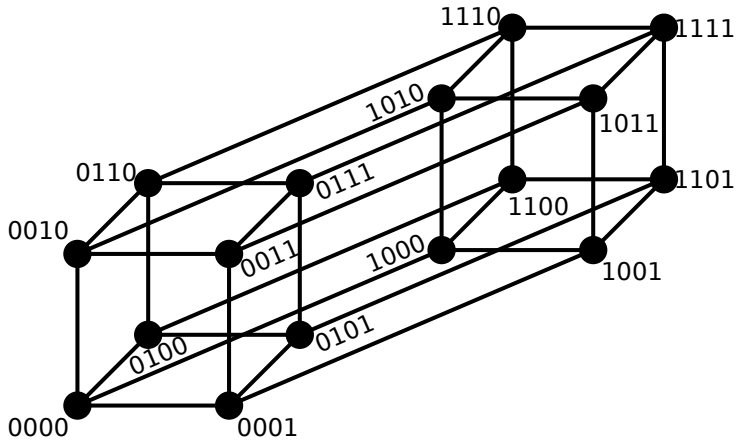


Figure 1.1: The binary sequence space is a hypercube of dimension L. Here, the projection of a binary hypercube of dimension $L = 4$ is shown with the binary alphabet $\mathfrak{A} = \{0, 1\}$. [3]

a single locus, do not have a large effect on the organism. Thus, within a species, there are many genomes, which are different, but yield an organism of a certain species. If the organism proliferates sexually, a subset of individuals which can reproduce by mating is called a *deme*.

Of course, with a different alphabet $\mathfrak{A}$, also different systems can be described, for example RNA has the same structure and alphabet size, but different letters, proteins have a larger alphabet with 20 letters, each corresponding to an amino acid. For the study of theoretical models it is convenient to restrict to a binary alphabet $\mathfrak{A} = \{+1, -1\} =: \mathfrak{B}$. Although this results in a lack of generality, and a generalization is often hard, it enables the use of a many different analytical and numerical methods. And although the reduction of biological information seems drastic, the binary system can still be interpreted in a variety of biological ways, for example the presence or absence of a gene, the state (active or passive) of a gene, or the absence or presence of a mutation. Also, it is very close to the physical systems of spin glasses. Following this tradition, this work is done using $\mathfrak{B}$, if not mentioned otherwise.

In the binary alphabet, all sequences of a fixed length L span the L–dimensional hypercube $\mathbb{H}^L$ as sequence space, see fig. 1.1. $\mathbb{H}^L$ is a metric space equipped with the *Hamming distance*:

Definition 1.1. The *Hamming distance* is a metric on the hypercube. It is

defined as

$$d : \mathbb{H}^L \times \mathbb{H}^L \to \{0, 1, \ldots, L\}$$
$$(\sigma, \sigma') \mapsto \sum_{i=1}^{L} 1 - \delta_{\sigma_i, \sigma'_i}$$

with the Kronecker Delta δ. This implies that the maximal Hamming distance between two sequences σ, σ' is $d(\sigma, \sigma') = L$. If $d(\sigma, \sigma') = L$, σ' is also called the *antipodal* of σ and will be labeled $\overline{\sigma}$.

$\mathbb{H}^L$ consists of 2^L sequences. Each sequence σ has L *neighbors* σ' with $d(\sigma, \sigma') = 1$ and in general the number of sequences at Hamming distance d is $\binom{L}{d}$ (see (A.4) for the convention on binomial coefficients). Thus, with respect to a given $\sigma \in \mathbb{H}^L$ most other $\sigma' \in \mathbb{H}^L$ lie at Hamming distance $d(\sigma, \sigma') = \frac{L}{2}$, where $\binom{L}{d}\big|_{d=L/2} = \frac{L!}{(L/2)!}$ is maximal. In graph theoretical notion, $\mathbb{H}^L$ is a regular graph (because every sequence has the same amount of neighbors), with 2^L vertices (neighbors) and $2^{L-1}L$ edges (connections between neighbors). The information about the structure of $\mathbb{H}^L$ is also contained in the *adjacency matrix* of $\mathbb{H}^L$.

Definition 1.2. The *adjacency matrix* A is a $2^L \times 2^L$ matrix, defined by

$$A_{\sigma,\sigma'} = \begin{cases} 1, & d(\sigma, \sigma') = 1 \\ 0, & \text{else.} \end{cases} \tag{1.1}$$

1.2 The struggle for existence

> The severe and often-recurrent struggle for existence will determine that those variations, however slight, which are favorable shall be preserved or selected, and those which are unfavorable shall be destroyed. This preservation, during the battle for life, of varieties which possess any advantage in structure, constitution, or instinct, I have called Natural Selection; and Mr. Herbert Spencer has well expressed the same idea by the Survival of the Fittest. Darwin [4]

"The Survival of the Fittest" is perhaps the most popular phrase connected to evolutionary biology. Although evolutionary biologists meanwhile refrain to use it due to its lack of generality and high potential for misunderstanding, the impact is remarkable. It has been used in and inspired works on early theories of evolution [5], economy, sociology and politics [6]. And still, in hindsight, set in the correct context it gives a very

nice description of evolutionary processes of many kinds, formulated in a time when the molecular basis of evolution was not even set.

Nevertheless, the phrase was not only misunderstood often but was also abused. The word 'fittest' can be interpreted in terms of physical or economical strength and toughness, which lead to the attempt to build a biological foundation and justification of the suppression of the weaker [6]. The real meaning of fitness in the biological context is probably 'best adapted'. The confusion about the term fitness is nevertheless, it seems, not only a problem outside evolutionary biology, but also for scientists in the field [7]. Therefore, this section is intended to clarify the used terms, introduce the mathematical language used to analyze evolutionary processes and give a few basic results.

There are at least two commonly used definitions of the term *fitness* on a molecular level.

Definition 1.3. If $n(\sigma, t)$ individuals carry the genotype σ at time t the *Malthusian fitness* $F(\sigma)$ is defined by

$$\frac{\mathrm{d}}{\mathrm{d}t} n(\sigma, t) = F(\sigma, t) n(\sigma, t). \tag{1.2}$$

Hence, $F(\sigma)$ is the growth rate of the organisms with genotype σ in continuous time.

Definition 1.4. Based on the last definition, the *Wrightian fitness* $w(\sigma)$ is defined by solving the differential equation (1.2) defining the Malthusian fitness:

$$n(\sigma, t+1) = w(\sigma, t) n(\sigma, t) \tag{1.3}$$

which implies the relation $w(\sigma, t) = e^{\int_0^t F(\sigma, t')\mathrm{d}t'}$, or for time independent fitness $w(\sigma) = e^{F(\sigma)}$. This fitness definition is particularly useful in a discrete time scenario, where the *generation time* t_g is set 1 for convenience. In the following, if not mentioned otherwise, time independent fitness is of interest.

Although mostly F will be used to denote fitness, concerning the properties of fitness landscapes the results do also apply to the Wrightian fitness w. Nevertheless, if using w it has to be ensured that for all $\sigma \in \mathbb{H}^L : w(\sigma) \geq 0$, which is not necessary for the Malthusian fitness F. Usually fitness is seen as a measure for reproductive success and thus the above definitions are very common. Nevertheless, in experiments it might be more convenient to measure a proxy for fitness. This might be for example the output of a certain protein [8] or the resistance to an antibiotic [9].

Defining fitness in dependence of the genotype, it is natural to think of F as a mapping from the sequence space into the real numbers.

Definition 1.5. A *fitness landscape* is a mapping

$$F : \mathbb{H}^L \to \mathbb{R}$$
$$\sigma \mapsto F(\sigma). \tag{1.4}$$

The idea of a fitness landscape was according to McCoy [10] first presented by Toulon in 1895 who used a different name. Nevertheless mostly Wright is credited and he also introduced the notion 'fitness landscape' [11]. Before experimental data became available in the recent years (see e.g. [12] for a review), the study of fitness landscapes was purely theoretical. Early population geneticists often used *additive* fitness landscapes for their mathematical analysis.

Definition 1.6. The *Mt. Fuji landscape* is an additive fitness landscape. Given an arbitrary reference sequence σ^*, the fitness is distributed as

$$F_{\sigma^*}(\sigma) = -cd(\sigma, \sigma^*).$$

Additive means here, that all loci are independent from each other, and the change in fitness resulting from the change of one locus never depends on the rest of the genome, which is also called the *genetic background*. An additive landscape only has one global *fitness maximum*: a point in the landscape at which all neighbors have lower fitness. Perhaps the most prominent arguments about the biological legitimacy of additive fitness landscapes is 'the beanbag genetics dispute' between the two friends Mayr and Haldane. While Haldane favored the simplicity of the additive model due to the mathematical possibilities, Mayr rejected it, calling genetics on such a model 'beanbag genetics' [13].

Opposed to the additive landscapes are the *epistatic* landscapes. Epistasis means nothing but the absence of additivity. One distinguishes basically two types. One is *magnitude epistasis*, where a beneficial (deleterious) mutation will be beneficial (deleterious), regardless of the genetic background, but its impact may vary. The other is *sign epistasis* [14] which means that not only the magnitude of the effect of a mutation may vary, but also its *sign*, thus a formerly beneficial mutation may become deleterious, depending on the genetic background. While additive and magnitude epistatic landscapes can only have one global fitness maximum, sign epistasis enables the possibility of multiple fitness maxima, see fig. 1.2. Due to this property, landscapes with a lot of sign epistasis are also called *rugged*. A model for such a rugged landscape was introduced by Kingman [15] as follows:

Definition 1.7. If all fitness values $F(\sigma)$ are identically and independently distributed (i.i.d.) random variables, F is called *House-of-Cards* (HoC) landscape.

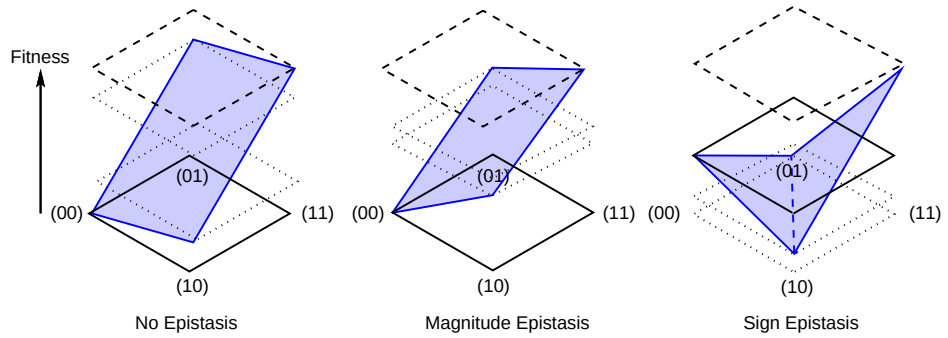


Figure 1.2: Illustration of the different types of epistasis in a two dimensional hypercube with alphabet $\mathfrak{A} = \{0, 1\}$. Only in the case with sign epistasis multiple fitness maxima are possible.

The HoC- and the Mt. Fuji-model describe in some sense contrary extremes of fitness landscapes. Although the HoC-landscape is more complex due to the high degree of epistatic interactions, it is still mathematically well feasible. A natural generalization is an interpolation between these two landscapes.

Definition 1.8. Let $\sigma^* \in \mathbb{H}^L$ be an arbitrary reference sequence, and let $\{\xi_i\}$ be a set of 2^L i.i.d. random variables. Then with $c \in \mathbb{R}$

$$F_{c,\sigma^*}(\sigma) = -cd(\sigma, \sigma^*) + \xi_\sigma$$

is called the *Rough-Mt.-Fuji* (RMF) landscape [16, 17, 18]. Note that for $c \to \infty$ and $c \to 0$ the Mt. Fuji and the HoC-landscapes are retrieved, respectively.

Definition 1.9. An alternative definition is given as follows. Let $\sigma^* \in \mathbb{H}^L$ be an arbitrary reference sequence, and let $\{\xi_i\}$ be a set of 2^L i.i.d. random variables. Then with $c \in \mathbb{R}$

$$F_{c,\sigma^*}(\sigma) = cd(\sigma, \sigma^*) + \xi_\sigma$$

is also called the *Rough-Mt.-Fuji* (RMF) landscape

Note that both definitions of the RMF-model are equivalent in the sense, that by a simultaneous change $c \to -c$ and $d \to L - d$ both can be transformed into one another.

The RMF-model provides the possibility to tune the correlations in the landscape. It is a more flexible model than the HoC which results in a more complicated mathematical analysis as trade-off.

Another popular model for fitness landscapes with epistasis is Kauffman's LK-model [19][1]. The idea is that each locus interacts with K other loci. The subsequence containing a locus and the loci it interacts with is called its LK-*neighborhood*. Since an LK-neighborhood has the size $K+1$, it is convenient to introduce $k = K + 1$.

Definition 1.10. Let $\{f_i : \mathfrak{B}^k \to \mathbb{R}\}$ be a set of i.i.d. random functions and let $\{\Xi(\sigma_i)\}$ be a set of LK-neighborhoods ($\Xi_i \in \mathfrak{B}^k$). The LK-*model* is defined by the fitness function

$$F(\sigma) = \frac{1}{\sqrt{L}} \sum_{i=1}^{L} f_i(\Xi(\sigma_i)).$$

Often used choices for LK-neighborhoods are for example the adjacent neighborhood $\Xi(\sigma_i) = (\sigma_i, \sigma_{i+1}, \sigma_{i+2}, \ldots, \sigma_{i+K})$, or the random neighborhood, where besides σ_i all K other loci are chosen at random. Note, that simple generalizations of the LK-model are achieved by altering the LK-neighborhoods to vary in size or to exclude σ_i.

The LK-model is, as the RMF-model, suited to tune between the Mt.-Fuji landscape ($k = 1$) and the HoC-landscape ($k = L$) [19, 20, 21, 22, 23, 24], although the k values in between do not yield a 'smooth' transition as in the RMF-landscape.

For convenience a fitness landscape with a random component which is distributed according to a distribution function P will be called P-distributed.

1.3 Evolutionary processes

The definitions of fitness 1.3 and 1.4 already imply an evolutionary process of *selection* which is one of the three *evolutionary forces:*

Definition 1.11. The three evolutionary forces are the mechanisms associated to the three parameters s (selection coefficient), N (population size) and μ (mutation rate).

- *Selection* $\sim s$ is a relative fitness measure which leads to focusing around particularly fit sequences. Its timescale is $\tau_s \propto \frac{1}{s}$.

[1] Kauffman called the model NK, referring to the sequence length as N, but because it is more common to call the population size N, the model is re-labeled here.

- *Genetic drift* $\sim \frac{1}{N}$ is caused by demographic stochasticity, fluctuations in the number of offspring. Its timescale is $\tau_N \propto N$.
- *Mutation* $\sim \mu$ leads to stochastic changes in the genome, the timescale is $\tau_\mu \propto \frac{1}{\mu}$.

Note that mutation and genetic drift introduce two different types of stochasticity.

Point mutations occur with probability μ. A point mutation is a randomly occurring change of allele at one locus, more precisely given two neighboring sequences $\sigma, \sigma' \in \mathbb{H}^L = \mathfrak{B}^L$ a point mutation $\sigma \to \sigma'$ at the ith locus is a transition $\sigma_i \to -\sigma_i$, comparable to a spin-flip in physics. If only two neighboring sequences are available and the mutation rate is the same in both directions, the following Langevin-equation is an extension of def. 1.3 and describes the evolution of $n(\sigma, t)$ under mutation, selection and reproductive fluctuations. In the situation where two genotypes σ, σ' with $d(\sigma, \sigma') = 1$ are present and thus $n(\sigma', t) = N - n(\sigma, t)$, it reads

$$\frac{\mathrm{d}}{\mathrm{d}t} n(\sigma, t) = F(\sigma, t) n(\sigma, t) + \chi(\sigma, t) + \mu(N - 2n(\sigma, t)), \tag{1.5}$$

with a random variable χ with $\langle \chi(\sigma) \rangle = 0$ and $\langle \chi(\sigma, t), \chi(\sigma, t') \rangle = \delta(t - t')$. It can be transformed into a Fokker-Planck-equation by Kramers-Moyale expansion [25], see e.g. [26] for the calculation. In population dynamics this is called *Kimura equation*. The *frequency* of σ is defined by $p(\sigma) = \frac{n(\sigma)}{N}$. The Kimura equation gives the change of the probability that a sequence σ has frequency p at time t [27, 28]:

$$\begin{aligned} \frac{\partial}{\partial t} \mathbb{P}(p, t) =& \frac{1}{2N} \frac{\partial^2}{\partial p^2} p(1-p) \mathbb{P}(p, t) \\ & - (F(\sigma, t) - F(\sigma', t)) \frac{\partial}{\partial p} p(1-p) \mathbb{P}(p, t) \\ & + \mu \frac{\partial}{\partial p} (1 - 2p) \mathbb{P}(p, t). \end{aligned} \tag{1.6}$$

In this context, the drift term is composed of mutation and selection, while the diffusion term is called *genetic drift*. Note, that in the presence of only two sequences $n(\sigma) + n(\sigma') = N$ and thus $p(\sigma) + p(\sigma') = 1$. The probability, that all individuals will carry only one of the two sequences is called *fixation probability* and was calculated by Kimura [29] to be

$$u(\sigma, \sigma') = \frac{1 - e^{-2(F(\sigma') - F(\sigma))}}{1 - e^{-2N(F(\sigma') - F(\sigma))}} \approx \begin{cases} 0, & F(\sigma') - F(\sigma) < 0 \\ 1 - e^{-2(F(\sigma') - F(\sigma))}, & F(\sigma') - F(\sigma) > 0 \end{cases} \tag{1.7}$$

where the approximation holds if $N|F(\sigma') - F(\sigma)| \gg 1$.

If more than two sequences are available, it is convenient to write the evolutionary equations as a matrix equation.

Definition 1.12. The *selection matrix* $\mathcal{S}$ is defined by $\mathcal{S}_{\sigma\sigma'} = w(\sigma)\delta_{\sigma\sigma'}$. The *mutation matrix* $\mathcal{M}$ is defined by $\mathcal{M}_{\sigma\sigma'} = (1-\mu)\delta_{\sigma\sigma'} + \frac{\mu}{L}A_{\sigma\sigma'}$. To formulate matrix equations it is necessary to understand n, p and w as vectors which contain as elements $n(\sigma), p(\sigma)$ and $w(\sigma)$ respectively.

The mutation-selection equations in a discrete time setting are then

$$\begin{aligned} n^{t+1} =& \mathcal{S}\mathcal{M}n^t && \text{unnormalized} \\ p^{t+1} =& \frac{\mathcal{S}\mathcal{M}p^t}{\sum_{\sigma\in\mathbb{H}^L}\left(\mathcal{S}\mathcal{M}p^t\right)(\sigma)} && \text{normalized} \end{aligned} \tag{1.8}$$

where the time t is measured in generations.

1.4 Evolutionary regimes

A common way to start the analysis of population genetic problems is to classify the problem according to its *evolutionary regime*, which is determined by the relative size of the evolutionary parameters

- $N\mu < 1$ is the *weak mutation* regime: on average not every generation a new mutant arises.
- $N\mu > 1$ is the *strong mutation* regime: on average every generation more than one new mutant arises. This can lead to a very diverse population.
- $|Ns| \ll 1$ is the *weak selection* regime: the fixation probability is low and a diverse population is probable.
- $|Ns| \gg 1$ is the *strong selection* regime: beneficial mutants are very likely to fix, while deleterious mutants will probably go extinct very fast.

A combination of those regimes which is of particular interest in this thesis is the *Strong-Selection-Weak-Mutation* regime (SSWM) first introduced by Gillespie [30]. As the name indicates, selection is strong ($|Ns| \gg 1$) while only few mutations arise ($N\mu \ll 1$). This leads to a situation, where double mutations are impossible in the sense, that selection either fixes beneficial mutations or kills individuals with deleterious mutations before a double

mutation arises. Thus, the population is *monomorphic*, i.e. all individuals have the same genome, all the time except for the short period of time when a new mutant was created and is evaluated by selection. This means, that besides the definition of the regime, mutation rate and population size do not affect the dynamics because the fixation probability is approximated by (1.7). The resulting evolutionary process is called *adaptive walk*, the population can be seen as a walker on the hypercube which can only make steps to neighboring sequences with a higher fitness. This implies, that the walker must stop as soon as a local *fitness maximum* is reached, which is defined by the absence of fitter neighbors. There are different kinds of adaptive walks, which differ in the transition probability to the next sequence. It is common to distinguish between three different processes:

Greedy adaptive walks (GAW): The next step is always taken to the fittest sequence in the neighborhood.

Random adaptive walks (RAW): The next sequence is chosen at random from the sequences of the neighborhood which are fitter than the current one.

Natural adaptive walks (NAW): The next sequence is chosen from the neighborhood with a probability proportional to the fitness difference

$$P_{\sigma_i \to \sigma_j} = \frac{(F(\sigma_j) - F(\sigma_i))\Theta(F(\sigma_j) - F(\sigma_i))}{\sum_{k=1}^{L}(F(\sigma_k) - F(\sigma_i))\Theta(F(\sigma_k) - F(\sigma_i))}. \tag{1.9}$$

This transition probability was derived from (1.7) by Gillespie [31].

Since all three processes are restricted to increase fitness in every step they will, for finite L, eventually stop after a finite number of steps. The mean walk length ℓ is the mean number of steps until a local optimum is reached, where the average is taken over runs and landscapes. ℓ is one of the most important properties of adaptive walks.

If a sequence σ' can be reached by an adaptive walk which started on another sequence σ, a path must exist from σ to σ' in which the fitness increases in every step. If such a path exist, it is called *accessible*.

1.5 Recombination

A reproduction strategy which is very common in nature is *recombination*. It leads to the genetic variation by introduction of (parts of) a second genome which is intertwined with the first one. This can for example happen by

sexual recombination, where the genome of two parents is recombined, but also by transformation, a process in which certain kinds of bacteria take up and incorporate exogenous genetic material. The success of recombination, especially sex, is one of the great mysteries of modern science. Although its apparent superiority in large parts of nature, it is not understood *why* sexual reproduction can be of benefit at all. Very simple economic reasoning gives rise to questions, e.g. the fact, that two parents are needed to reproduce without increasing the number of offspring has become famous as the *two-fold cost of sex* [32, 33]. But also without the costs, it is not clear whether recombination gives any benefit at all [34, 35, 36, 37], one point being, that recombination might break apart useful structures in the genome, which is known as *recombination load* [38].

Nevertheless, this paradox has inspired several attempts to explain the prevalence of recombination:

Muller's ratchet & deterministic mutation hypothesis: recombination enables a population to purge deleterious mutations [39, 40, 41] faster than asexuals, which have to wait for a back-mutation.

Fisher-Muller- & Hill-Robertson-effect: If two beneficial mutations occur in a population, both can spread fast, without the need of two double mutations [42, 43, 44, 40].

Weismann effect: More variation is created [45].

Fisher's fundamental theorem: Here, the benefit follows from the Weismann effect, stating that the mean fitness increase is proportional to the genetic variance [42]. Sadly, this result is not generally true but assumes the absence of epistasis [46], or it needs a certain kind of variance measure [47].

Red Queen Hypothesis: Recombination increases the *speed* of adaptation, not necessarily the *ultimate fitness*. This way, organisms can adapt quicker to the surroundings than other ever-evolving species [48].

The mathematical description of recombination will be done in a similar manner as mutation above, following Stadler and Wagner [49]. And as before, the sequence length is constrained to be constant.

Definition 1.13. The *recombination operator* $\mathfrak{R}$ is a mapping into the powerset of $\mathbb{H}^L$:

$$\begin{aligned}\mathfrak{R} : \mathbb{H}^L \times \mathbb{H}^L &\to \mathfrak{P}(\mathbb{H}^L)\\ (\sigma, \sigma') &\mapsto \{\sigma'' | {\sigma''}_i = \sigma_i \vee {\sigma''}_i = \sigma'_i\}.\end{aligned}$$

$\mathfrak{R}(\sigma, \sigma')$ is the set containing all possible recombinations of σ and σ' which have the same length L.

The adjacency matrix is not the matrix containing the most valuable information about the structure of the process anymore, as it was for mutation. It is replaced:

Definition 1.14. The *incidence matrix* H is to recombination what the adjacency matrix is for mutation. It is defined by

$$H_{\sigma,(\sigma',\sigma'')} = \begin{cases} 1, & \sigma \in \mathfrak{R}(\sigma', \sigma'') \\ 0, & \text{else.} \end{cases}$$

In terms of H, a recombinatorial transition matrix can be defined by

$$\mathcal{T}_{\sigma' \to \sigma} = \sum_{\sigma'' \in \mathbb{H}^L} p(\sigma'') p(\sigma') \frac{H_{\sigma,(\sigma',\sigma'')}}{|\mathfrak{R}(\sigma', \sigma'')|}.$$

There, $|\bullet|$ means the cardinality.

In the definition of $\mathcal{T}$ it becomes obvious, that recombination is quadratic in the population frequency p and not, as mutation, linear. There, recombination is incorporated by using a *uniform crossover*, i.e. the recombined sequence is put together in such a way, that at each locus the allele is taken from one of the corresponding parent genomes at random. Other possible crossovers are, e.g., *single point crossover*, where the parent genomes are cut at a certain point and the fronts and rears are exchanged. Or the *two-point-crossover* where the genomes are cut two times and the middle part is exchanged. At such a crossover two possible genomes are created of which one has to be chosen at random. In the following, only the uniform crossover will be analyzed.

1.6 Extremes: Fitness and probability

In 'everyday statistics' it is important to understand the *average* behavior of apparently random events. One of the most used tools to do so is the *central limit theorem*. Its statement is the following:
For every set of n i.i.d. random variables $\{X_i\}$ with expected value μ and variance σ^2, the random variable $\sqrt{n}\,(\sum_{i=1}^n X_i - \mu)$ converges in distribution to a Normal-distribution, more precisely

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \xrightarrow{d} \mathcal{N}(0, \mathrm{Var}(X_i)) \ (n \to \infty). \tag{1.10}$$

As mentioned above, the central limit theorem describes the average behavior, which is reflected by the fact, that it is a statement about a *sum* of random variables. Now, in some cases the average is not of particular interest. For example is an average tide nothing to worry about. But what is of critical interest is the height of *extreme* tides. If a city builds a levee, it has to know that it is high enough to withstand not only the everyday flood, but also the one-in-a-century flood. In evolutionary biology the extremes are important, too. The sequence space is so large, that probably most of the possible genomes are lethal. Only extraordinary fit sequences bear the possibility to be viable. Instead of the average behavior, it is now important to describe the *maxima* and the *extraordinary large* of a set of random variables. In the last century this part of mathematics has become very successful under the name *Extreme-Value-Theory* (EVT).

Definition 1.15. The *Generalized-Extreme-Value-Distribution* (GED) is defined by the distribution function

$$G(x;\mu,\sigma,\kappa) = \exp\left\{-\left[1+\kappa\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\kappa}\right\}.$$

It has a scale parameter σ, a location parameter μ and a shape parameter κ with the restriction $1+\kappa(x-\mu)/\sigma > 0$.

Theorem 1 (Fisher-Tippett-Gnedenko [50, 51])**.** *Let $\{X_i\}$ be a set of n i.i.d. random variables and $M_n = \max(X_1,\ldots,X_n)$. Suppose there exists a sequence of constants $a_n > 0$, b_n with*

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \to D(x)\,(n\to\infty)$$

with a non-degenerate distribution function D, then parameters can be found such that $D(x) = G(x;\mu,\sigma,\kappa)$.

The shape parameter κ governs the tail behavior of the GED. In terms of κ three *probability classes* can be defined, each containing the probability distributions which converge to a corresponding GED in terms of thm. 1:

- $\kappa = 0$: Defines the *Gumbel class* or type I distributions. This class contains distributions with a tail vanishing faster than power law.
- $\kappa > 0$: Defines the *Fréchet class* or type II distributions. This class contains all distributions with power law tail.

- $\kappa < 0$: Defines the *Weibull class* or type III distributions. GED has a support limited to the right by $\mu - \frac{\sigma}{\kappa}$. This class contains all distributions with power law tail and bounded support.

Another approach to the analysis of extreme values is via *Peaks-Over-Threshold* (POT). If P is some distribution function and u is a threshold, the *excess function* is the distribution function for values *over* the threshold:

$$\mathcal{E}_u(y) = \mathbb{P}(X - u \leq y | X > u) = \frac{P(u+y) - P(u)}{1 - P(u)}.$$

Definition 1.16. The *Generalized-Pareto-Distribution* (GPD) is defined by the distribution function

$$\mathrm{GPD}_{(\kappa,\mu,\sigma)}(x) = \begin{cases} 1 - \left(1 + \frac{\kappa(x-\mu)}{\sigma}\right)^{-1/\kappa} & \text{for } \kappa \neq 0, \\ 1 - \exp\left(-\frac{x-\mu}{\sigma}\right) & \text{for } \kappa = 0, \end{cases}$$

for $x \geqslant \mu$ when $\kappa \geqslant 0$, and $\mu \leqslant x \leqslant \mu - \sigma/\kappa$ when $\kappa < 0$, where κ has the same role as in the GED and gives the same information about the extreme value class. μ and σ are location and scale parameter, respectively.

Theorem 2 (Pickands-Balkema-deHaan [52, 53])**.** *Let $\{X_i\}$ be a set of n i.i.d. random variables. Then parameters can be found, such that:*

$$\mathcal{E}_u(y) \to \mathrm{GPD}_{(\kappa,\mu,\sigma)}(y) \ (u \to \infty).$$

Although both approaches have been proved to be very powerful in the past, usually for data analysis the limits $n \to \infty$ or $u \to \infty$ cannot be satisfied. Thus, as an alternative to the above discussed *ultimate EVT*, the *penultimate EVT* deals with the description of large but finite sets of i.i.d. random variables. It might in such a case be better to describe a given dataset with a distribution function which has a different κ than its limiting distribution function. Assuming, that a set $\{X_i\}$ of n i.i.d. random variables with distribution function P which has a limiting distribution with $\kappa = 0$ (one says it is in the domain of attraction of Gumbel type). Then the $(1 - \frac{1}{n})$th quantile $q(n) = P^{-1}(1 - \frac{1}{n})$ is an approximation for the typical largest value. The *hazard function* is defined by

$$h(x) = \frac{\frac{\mathrm{d}}{\mathrm{d}x} P(x)}{1 - P(x)}. \tag{1.11}$$

It has only a finite limiting value $\lim_{x\to\infty} h(x)$ for distributions with exponential tail. One choice of the shape parameter is then [54]

$$\kappa_n = \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{h(x)}\right)\bigg|_{x=q(n)}. \tag{1.12}$$

The fact that even distributions from the Gumbel class might behave like non-Gumbel-class distributions for finite sample sizes emphasizes one thing: it is always important to know how things are for distributions besides the domain of attraction of Gumbel type. All experimental data sets are finite, which implies the necessity to investigate the behavior of processes, correlations, etc. for all three probability classes.

1.7 Spin glasses & other fields of interest

> Spin glasses (also called amorphous magnets) are magnetic substances in which the interaction among the spins is sometimes ferromagnetic (it tends to align the spins; $J_{ik} > 0$), sometimes antiferromagnetic (it antialigns the spins; $J_{ik} < 0$). The sign of the interaction is supposed to be random. In some spin glasses the spins can take only two values ± 1 (Ising spins) [...]. Mezard et al. [55]

In the above quote, J_{ik} are the couplings, or interactions, between two spins i and k. Their supposed randomness leads to inherent *disorder*, more precisely *quenched disorder*. Considering spin glasses of Ising spins means that spin configurations σ are elements from the hypercube $\mathbb{H}^L$, and also, that $\mathbb{H}^L$ is the configuration space for this spin glass model. The spatial configuration of the spins (chain, lattice, etc.) enters through the interactions. The energy of each configuration is measured with help of the *Hamiltonian* $\mathcal{H}(\sigma)$ which corresponds to negative fitness. Special states in the system are those which minimize energy. Those are called *(meta)stable states*. This means, that the Hamiltonian corresponds to a negative fitness landscape in the picture of evolutionary biology where fitness maxima are (meta)stable. A stable state or ground state corresponds to a global fitness maximum, whereas a metastable state corresponds to a local fitness maximum. If a fitness landscape shows sign epistatis, the corresponding situation in spin glass analysis is called *frustration*. Both fields are in fact very similar. This is also expressed in the similarity of models. A toy model which was proposed by Derrida [56] is the *Random Energy Model* (REM) in which $\mathcal{H}(\sigma)$ are i.i.d. random variables. This corresponds to the House-of-Cards model for fitness landscapes. As a straight forward generalization Derrida [57] proposed the *p-spin-model* [58], in which subsets of p of the L spins are interacting and the Hamiltonian takes the form

$$\mathcal{H}(\sigma) = -\sum_{i_1 \dots i_p} A_{i_1 \dots i_p} \sigma_{i_1} \dots \sigma_{i_p}$$

with random contributions $A_{i_1 \ldots i_p}$. This model is the analogue of the Fourier expansion of a fitness landscape (see sec. 3.2 for details). Each coefficient $A_{i_1 \ldots i_p}$ introduces interactions between p spins. Since in the LK-model the number of interacting loci is determined by K it can be understood as a superposition of sparse p-spin-models since in general many coefficients are null. The RMF-model corresponds to a REM model in an external magnetic field, which has the Hamiltonian

$$\mathcal{H}(\sigma) = a\xi_\sigma - \mu \sum_{i=1}^{L} h\sigma_i$$

with a *magnetic moment* μ, field h and a constant a which ensures the correct dimension.

Adaptive processes are linked to spin glasses as well. For example are adaptive walks very similar to a Metropolis dynamics on a spin glass at zero temperature [59]. The relaxation dynamics of a spin glass in an exterior magnetic field have been studied in the context of spin-glass *aging* [60], which is related to adaptive walks on an RMF-landscape.

Spin glasses are nevertheless not the only area, where similar models apply. Basically in all fields, in which binary structures appear, similarities occur. A very prominent example for that are several problems in computer science (e.g. [61]).

The probability for the existence of accessible paths on fitness landscapes [24] is closely connected to *percolation* [62]. Recently, in the context of an RMF model on a Cayley tree (which is very similar to the hypercube for large L) the term *accessibility percolation* was coined [63].

1.8 Experiments and fitness proxies

Although the basic ideas of fitness and fitness landscapes are several decades old, the experimental tools for fitness measurements are quite new (or only recently technically available). There are basically two approaches in the experimental study of fitness landscapes [64].

For the first one, the evolution of organisms with a short generation time is observed and the fitness development is measured and compared to the ancestors (e.g. [65]). From the fitness changes, properties of the underlying landscape can be inferred on a qualitative level. Although the generated data sets may cover large regions of the genotype space, the resulting picture of the fitness landscape is incomplete. Additionally, it is biased by the evolutionary regime and the resulting adaptive dynamics.

Fitness values which are obtained in a similar way are compared to an RMF-landscape in sec. 2.3. The data from Miller et al. [66] are fitness values of the bacteriophage ID11. The measurements were performed in a bottleneck manner [67]: in several passages, the bacteriophage was allowed to grow in a bacterial host medium, here *Escherichia coli C*, for one hour. Then the process was stopped and the growth rate measured to calculate the fitness.

The second approach relies on the analysis of predefined mutations, which are created to observe the fitness in a tiny, but complete part of the hypercube. For experimental convenience, it is common to not measure Malthusian or Wrightian fitness itself, but a *proxy*, such as antibiotic resistance.

In sec. 3.5 empirical landscapes of the second type are compared to model-landscapes. One "real" fitness landscape is the $L = 6$ landscape from Hall et al. [68] of yeast, fitness is measured as growth rate. Another one is the $L = 8$ growth rate landscape of the fungus *Aspergillus niger* presented by Franke et al. [69]. Additionally two $L = 9$ landscapes which measure the output of certain enzymes in *Nicotiana Tobaccum* as a fitness proxy, specifically the enzymatic specificity of terpene synthases, that is, the relative production of 5-epi-aristolochene and premnaspirodiene, presented by O'Maille et al. [8] were analyzed. The last two landscapes are not complete, only 418 of the 512 fitness values are given. The missing data points are interpolated by fitting a linear model [70].

Chapter 2

The Rough-Mt.-Fuji model

In this chapter, several properties of the RMF-model (def. 1.8) will be calculated and it will be fitted to experimental data. It was introduced by Aita et al. [16] in the context of biopolymers in a slightly more general way. In the following, the calculations will be restricted to the def. 1.8 and the reasoning is close to [18].

2.1 Further remarks on the definition

The Rough Mt. Fuji Landscape will in this chapter be used as defined in def. 1.8. Because the mean fitness gradient is for $c > 0$ directed towards σ^*, RMF-landscapes are not isotropic. On average, fitness increases in one and decreases in another direction. These directions are defined by the change in Hamming distance to σ^*, $d(\sigma, \sigma^*)$.

Definition 2.1. The *neighborhood-set* ν of a sequence is defined by $\nu(\sigma) = \{\sigma' | d(\sigma, \sigma') = 1\} \cup \{\sigma\}$. This set is split in three parts ($d(\sigma, \sigma^*) =: d$):

- The *uphill–neighborhood* $\nu^{\uparrow}(\sigma) = \nu(\sigma) \cap \{\sigma' | d(\sigma', \sigma^*) = d - 1\}$ which contains all neighbors which are closer to σ^* and thus have an average fitness advantage of c,
- σ itself in $\nu(\sigma)^{\bullet} = \{\sigma\}$,
- and the *downhill–neighborhood* $\nu^{\downarrow}(\sigma) = \nu(\sigma) \cap \{\sigma' | d(\sigma', \sigma^*) = d + 1\}$ which contains all neighbors which are further away from σ^* and thus have an average fitness disadvantage of c.

Obviously $\nu(\sigma)^{\uparrow} \cup \nu(\sigma)^{\bullet} \cup \nu(\sigma)^{\downarrow} = \nu(\sigma)$.

Note that the fitness values in the RMF-landscape are not i.i.d. but correlated random variables.

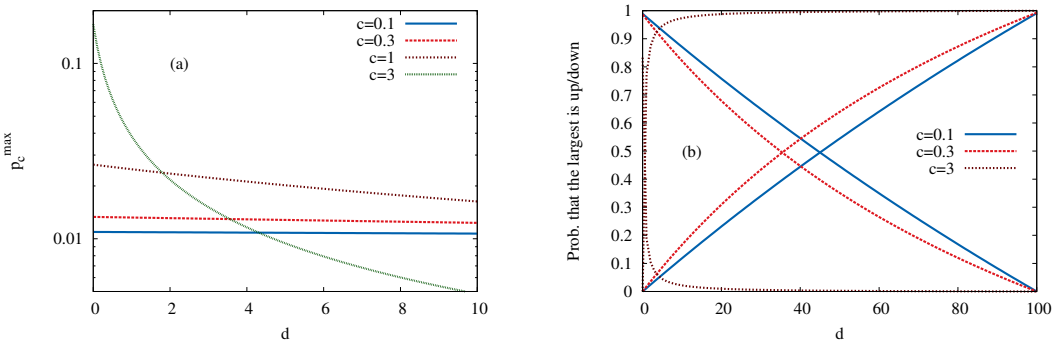


Figure 2.1: (a) The probability that a given sequence is a local fitness maximum is shown as a function of the distance d to the reference sequence for several values of c and $L = 100$. (b) The probability that the neighboring sequence of largest fitness is in the uphill (solid lines) or downhill (dashed lines) direction is shown as a function of d for different values of c and $L = 100$. Both plots show results of a Gumbel distributed RMF-landscape.

2.2 Fitness maxima & correlations

One of the characteristic properties of a fitness landscape is the number of fitness maxima. It gives information about the landscape's roughness and is also a so called 'epistasis measure' [12]. In the HoC-landscape, the probability $p^{\max}$, that $F(\sigma)$ is a fitness maximum, is equal to the probability, that it is the largest fitness value in $\{F(\sigma')|\sigma' \in \nu(\sigma)\}$ which is $\frac{1}{|\nu(\sigma)|} = \frac{1}{L+1}$. In the RMF-model the probability depends additionally on the parameters c and d as well as on the underlying probability distribution. If the distribution function of the random variables is P, the probability can be written as

$$p_c^{\max}(d) = \int \mathrm{d}x\, p(x) \left(P(x-c)\right)^d \left(P(x+c)\right)^{L-d}. \tag{2.1}$$

The sums in the argument of the distribution function prevent a general evaluation of the integral. A special case is the *Gumbel distribution* $P_G(x) = e^{-e^{-x}}$. It is a limiting distribution in the Gumbel class, $P_G(x) = \mathrm{GEV}_{0,1,0}(x)$, which comes with a very useful *shifting property*:

$$P_G(x+c) = e^{-e^{-x-c}} = e^{e^{-x}e^{-c}} = \left(e^{-e^{-x}}\right)^{-e^{-c}} = P_G(x)^{-e^{-c}} \tag{2.2}$$

Result 2.1 ($p_c^{\max}$ in the Gumbel case)**.** *In the Gumbel case, the probability that a given sequence is a fitness maximum is given by*

$$p_c^{max}(d) = \frac{1}{1 + de^c + (L-d)e^{-c}}.$$

Proof. Using the shifting property on (2.1) yields:

$$\begin{aligned} p_c^{\max}(d) = & \int \mathrm{d}x\; p(x)\left(P(x-c)\right)^d \left(P(x+c)\right)^{L-d} \\ = & \int \mathrm{d}x\; p(x)\left(P(x)^{-e^c}\right)^d \left(P(x)^{-e^{-c}}\right)^{L-d} \\ = & \int \mathrm{d}P_G\; P_G^{-e^c d - e^{-c}(L-d)} = \frac{1}{1+de^c+(L-d)e^{-c}}. \end{aligned}$$

□

The behavior of res. 2.1 for intermediate values of c is illustrated in fig. 2.1(a).

Besides the special case of the Gumbel distribution, it is possible to approximate (2.1) following [71, 72] by expanding in c:

$$p_c^{\max}(d) = \frac{1}{L+1} + c(L-2d)I_{L-1} + \mathcal{O}(c^2) \;\text{ with }\; I_{L-1} = \int \mathrm{d}x\;\; p(x)^2 P(x)^{L-1}. \tag{2.3}$$

Expressions for I_L for representatives of the three extreme value classes have been derived by Franke et al. [71]. For large L the integral behaves as [72]

$$I_L \sim L^{-(2+\kappa)}, \tag{2.4}$$

where κ denotes the extreme value index from def. 1.15. This implies a stronger effect of c on the number of maxima for Weibull class distributions than for Fréchet or Gumbel class distributions.

The anisotropy of the landscape also inspires the question, in which direction the fittest neighbor is positioned. Although the uphill neighbors have a fitness benefit of c, the number of neighbors in $\nu^{\uparrow}$ and $\nu^{\downarrow}$ varies with the position with respect to the reference sequence. The corresponding probabilities to find the fittest of the neighborhood uphill or downhill are $p_c^{\uparrow}(d)$ and $p_c^{\downarrow}(d)$. A modification of (2.1) restricted to $\nu^{\uparrow}(\sigma)$ and $\nu^{\downarrow}(\sigma)$ yields the general expressions

$$p_c^{\uparrow}(d) = d\int \mathrm{d}x\;\; p(x)P(x)^{d-1}P(x+c)P(x+2c)^{L-d}, \tag{2.5}$$

$$p_c^{\downarrow}(d) = (L-d)\int \mathrm{d}x\;\; p(x)P(x)^{L-d-1}P(x-c)P(x-2c)^{d}, \tag{2.6}$$

which can be explicitly evaluated for the Gumbel distribution.

Result 2.2 (Fittest is up/down in the Gumbel case)**.** *In the Gumbel case, the probability, that the fittest of the neighborhood is positioned uphill/downhill is given by*

$$p_c^{\uparrow}(d) = \frac{d}{d + e^{-c} + e^{-2c}(L-d)}, \qquad p_c^{\downarrow}(d) = \frac{L-d}{L-d + e^{c} + e^{2c}d}.$$

Proof.

$$\begin{aligned} p_c^{\uparrow}(d) = & \quad d \int \mathrm{d}x\; p(x) P_G(x)^{d-1} P_G(x+c) P_G(x+2c)^{L-d} \\ = & \quad d \int \mathrm{d}P_G\; p(x) P_G(x)^{d-1} P_G(x)^{-e^{-c}} (P_G(x)^{-e^{-2c}})^{L-d} \\ = & \quad d \int \mathrm{d}P_G\; P_G^{(d-1)+e^{-c}+e^{-2c}(L-d)} \\ = & \quad \frac{d}{d + e^{-c} + e^{-2c}(L-d)} \end{aligned}$$

and analogously for $p_c^{\downarrow}$. □

Note that $p_c^{\uparrow} + p_c^{\downarrow} + p_c^{\max} = 1$ and $p_c^{\uparrow} = de^{c}p_c^{\max}$, $p_c^{\downarrow} = (L-d)e^{-c}p_c^{\max}$. d induces a benefit to the largest sub-neighborhood which is $\nu^{\uparrow}(\sigma)$ ($\nu^{\downarrow}(\sigma)$) if $d < \frac{L}{2}$ ($d > \frac{L}{2}$) just because the larger number of random variables in it increases the expected largest value in it. This leads to a certain kind of asymmetry, even in the case $c = 0$. For $c > 0$ the crossing point where $p_c^{\mathrm{up}} = p_c^{\mathrm{down}}$ moves towards the reference sequence with increasing c and is generally located at $d = \frac{L}{1+e^{2c}}$, see fig. 2.1(b).

From the density (2.1) the total number of maxima M is calculated by averaging over the landscape:

$$\begin{aligned} M = & \quad \sum_{d=0}^{L} \binom{L}{d} p_c^{\max}(d) \\ = & \quad \sum_{d=0}^{L} \binom{L}{d} \int \mathrm{d}x\; p(x) \left(P(x-c)\right)^{d} \left(P(x+c)\right)^{L-d} \\ = & \quad \int \mathrm{d}x\; p(x) \sum_{d=0}^{L} \binom{L}{d} \left(P(x-c)\right)^{d} \left(P(x+c)\right)^{L-d} \\ = & \quad \int \mathrm{d}x\; p(x) \left(P(x-c) + P(x+c)\right)^{L}. \end{aligned} \tag{2.7}$$

Using the linearized expression (2.3), c drops out of the expression. This means, that c in linear order only influences the place, where maxima are probable, not M, which depends obviously only on higher order terms.

Result 2.3 (Higher order evaluation of M)**.** *Expanding* (2.7) *in* c *yields corrections up to third order terms:*

$$M = \frac{2^L}{L+1} - c^2 2^{L-2} L(L-1) J_L + \mathcal{O}(c^4)$$
$$\text{with } J_L = \int \mathrm{d}x\; p(x)^3 P(x)^{L-2}.$$

In terms of the GPD this yields $J_L \sim L^{(3+2\kappa)}$ *for large* L*.*

Proof. To calculate the expansion, the integrand of (2.7) is derived:

$$\frac{\partial}{\partial c}(P(x+c)+P(x-c))^L = L(P(x+c)+P(x-c))^{L-1}(p(x+c)-p(x-c))$$
$$\frac{\partial^2}{\partial c^2}(P(x+c)+P(x-c))^L =$$
$$L(L-1)(P(x+c)+P(x-c))^{L-2}(p(x+c)-p(x-c))^2$$
$$+L(P(x+c)+P(x-c))^{L-1}\left(\frac{\partial}{\partial c}p(x+c)+\frac{\partial}{\partial c}p(x-c)\right).$$

Since all terms $(p(x+c)-p(x-c))$ vanish for $c=0$ they do not contribute to the expansion, such that the integral up to $\mathcal{O}(c^2)$ takes the form

$$M \approx \int \mathrm{d}P(x)\;\; (2P(x))^L + \frac{c^2}{2} L \left((L-1)\,(2P(x))^{L-2} + L\,(2P(x))^{L-1}\, 2\frac{\partial}{\partial c}p(x) \right),$$

where the first part yields $\frac{2^L}{L+1}$ and the second arrives at $-c^2 2^{L-2} L(L-1) J_L$ after integration by parts. □

Result 2.4 (Number of maxima in the exponential case)**.** *In an exponentially distributed landscape (see* (A.5)*), the expected number of maxima is given by*

$$M = \frac{e^c}{L+1}\left((1-e^{-2c})^{L+1} - (1-e^{-c})^{L+1}\right) + \frac{2^L\left(1-(1-e^{-c}\cosh(c))^{L+1}\right)}{\cosh(c)(L+1)}.$$

Proof. The exponential distribution function is $P(x) = (1-e^{-x})\,\Theta(x)$. This leads for $x > c$ to

$$\begin{aligned} P(x-c)+P(x+c) &= 2 - e^{-x}\left(e^{-c}+e^{+c}\right) \\ &= 2\left(1-e^{-x}\cosh(c)\right) \\ &= 2\cosh(c)\left(\frac{1}{\cosh(c)} - 1 + P(x)\right) \end{aligned}$$

Includig the term for $x < c$ it follows from (2.7):

$$\begin{aligned} M = & \int_0^c \mathrm{d}x\; p(x)P(x+c)^L + \int_c^\infty \mathrm{d}x\; p(x)\left(P(x-c)+P(x+c)\right)^L \\ = & \int_{p(c)}^1 \mathrm{d}p(x)\;\left(1-e^{-c}p(x)\right)^L + 2^L\int_0^{p(c)} \mathrm{d}p(x)\;\left(1-p(x)\cosh(c)\right)^L \end{aligned}$$

which leads directly to the result. □

Result 2.5 (M for uniform distribution)**.** *In a uniform distributed landscape, the expected number of maxima is given by*

$$M = \begin{cases} \frac{(2-c)^{L+1}-2^L\left((1-c)^{L+1}+c^{L+1}\right)}{L+1}, & c < \frac{1}{2} \\ \frac{(2-c)^{L+1}-2^L(1-c)^{L+1}}{L+1}, & c \geq \frac{1}{2}. \end{cases}$$

Proof. Starting with (2.1) M is calculated as done in (2.7). The uniform distribution has $P(x) = x\Theta(1-x)\Theta(x) + \Theta(x-1)$ which leads to

$$\begin{aligned} M &= \sum_{d\geq 0}\binom{L}{d}\left[\Theta\left(\frac{1}{2}-c\right)\int_c^{1-c}(x-c)^d(x+c)^{L-d}\mathrm{d}x\; + \int_{1-c}^1 (x-c)^d\right] \\ &= \Theta\left(\frac{1}{2}-c\right)\int_c^{1-c} 2^L x^L + \int_{1-c}^1 (1+x+c)^L. \end{aligned}$$

Which arrives at the result after simple integrations. □

Result 2.6 (Number of maxima in the Gumbel case)**.** *In a Gumbel distributed landscape, the exact expression of M in terms of the hypergeometric function is*

$$M = (1+Le^{-c})^{-1}\;{}_2F_1(-L,\zeta;\zeta+1;-1) \quad \textit{with} \quad \zeta = \frac{1+Le^{-c}}{2\sinh(c)}.$$

Proof. The hypergeometric function is defined by [73]

$${}_2F_1(a,b;c;z) = \sum_{n\geq 0}\frac{(a)_n(b)_n}{(c)_n}\,\frac{z^n}{n!} = \sum_{n\geq 0} t_n$$

with the Pochhammer symbol

$$(x)_n = \begin{cases} 1 & \text{if } n = 0 \\ x(x+1)\cdots(x+n-1) & \text{if } n > 0. \end{cases}$$

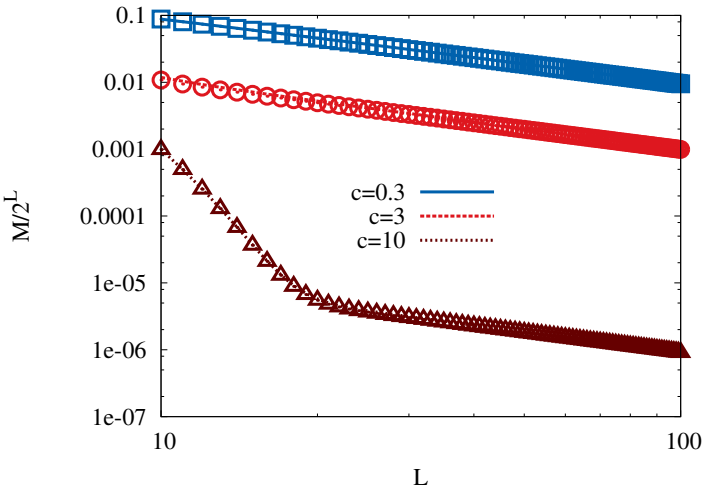


Figure 2.2: The density of local fitness maxima $\mathcal{M}/2^L$ in a Gumbel distributed RMF-landscape is shown as a function of the number of loci L. Symbols correspond to the large L approximation (2.9) and lines to the exact result 2.6.

The defining property of the hypergeometric function is that the terms t_n satisfy $t_0 = 1$ and

$$\frac{t_{k+1}}{t_k} = \frac{(k+a)(k+b)}{k+c}\frac{z}{k+1}.$$
$$M = \frac{1}{1+Le^{-c}}\sum_{d\geq 0}\binom{L}{d}p_c^{\max}(d)(1+Le^{-c}) =: \frac{1}{1+Le^{-c}}\sum_{d\geq 0}t_d,$$

ensuring that $t_0 = 1$, the fractions t_{d+1}/t_d have to be calculated according to

$$\frac{t_{d+1}}{t_d} = \left(\frac{L-d}{d+1}\right)\frac{1+Le^{-c}+2d\sinh(c)}{1+Le^{-c}+2(d+1)\sinh(c)} = \left(\frac{-1}{d+1}\right)\frac{(d-L)\left(d+\frac{1+Le^{-c}}{2\sinh(c)}\right)}{d+1+\frac{1+Le^{-c}}{2\sinh(c)}}.$$

By comparison the arguments a, b, c and z can be identified. □

For large L, the binomial coefficients become peaked around $d = \frac{L}{2}$ which yields the approximation

$$M \overset{L\gg 1}{\approx} 2^L p_c^{\max}(L/2) = \frac{2^L}{L\cosh(c)+1}. \tag{2.8}$$

Nevertheless the approximation violates $M \geq 1\,(c \to \infty)$, which can be fixed

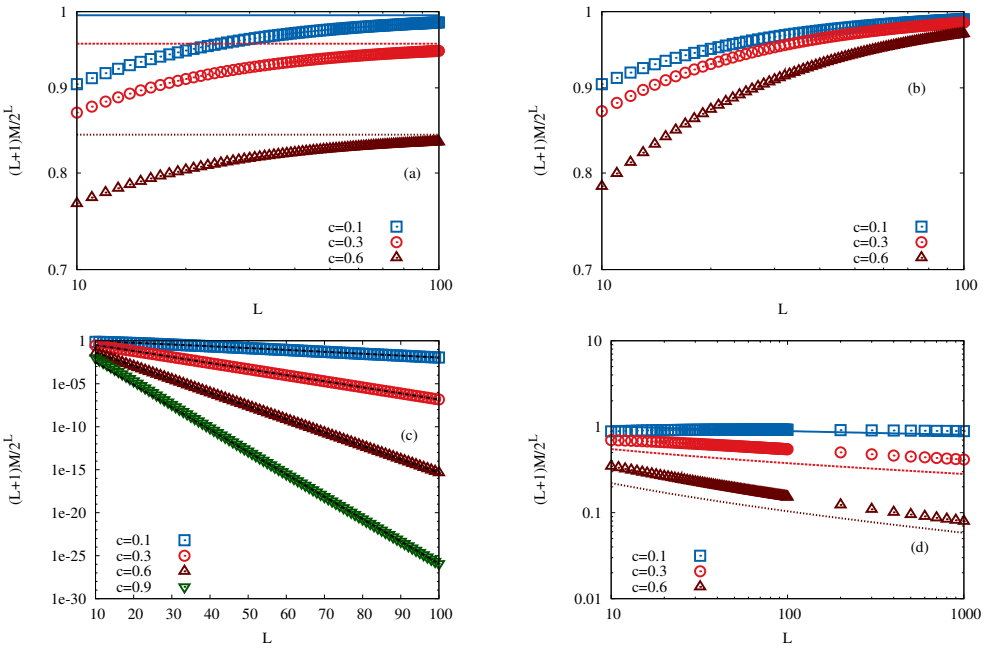


Figure 2.3: This figure illustrates the behavior of the number of Maxima M. The number of local maxima is normalized by its asymptotic value in the HoC-case $\frac{2^L}{L}$ and plotted in dependence on the sequence length L. Panel (a) shows res. 2.4 for an exponentially distributed RMF-landscape ($\xi \sim e^{-x}$, Gumbel class). The dashed horizontal lines show, that for large L the expression converges to $\frac{1}{\cosh(c)}$. Panel (b) shows the numerical evaluation of (2.7) for a Pareto distributed random contribution ($\xi \sim 2x^{-3}$, Fréchet class). In accordance with res. 2.8 (the distribution corresponds to $\kappa = \frac{1}{2}$) the expression converges to unity for large L. Panel (c) shows res. 2.5 in comparison to (2.7). Panel (d) shows results for a Weibull distributed RMF-landscape ($\xi \sim e^{-x^\beta}$, Gumbel class) for $\beta = 2$. The points are from numerical evaluation of (2.7) while the lines are obtained by fitting a factor to res. 2.7.

by

$$M_{\text{approx}} = 1 + \frac{2^L}{L\cosh(c) + 1}. \tag{2.9}$$

As can be seen in fig. 2.2 this approximation describes the data from numerical evaluation of res. 2.6 quite well.

As mentioned in sec. 1.6, the Gumbel class contains also distributions with a tail lighter or heavier than the exponential distribution. To study the behavior of such slight changes in the tail, the Weibull distribution $P(x) = \left(1 - e^{-x^\beta}\right)\Theta(x)$ (see also (A.6)) will be studied in the context of the number of maxima, particularly the behavior for $\beta \neq 1$.

Result 2.7 (Number of maxima for different tails in the Gumbel class)**.** *In a Weibull distributed RMF-landscape, the leading order of the number of maxima for large L is given by*

$$M \approx \begin{cases} \frac{2^L}{L} e^{-\beta c \log(L)^{1-\frac{1}{\beta}}}, & \beta > 1 \\ \frac{2^L}{L}, & \beta < 1. \end{cases}$$

Proof. In the calculations above, the occuring integrals were split into parts to respect the support of the distribution. One for $x \in [0, c]$ and one for $x \in [c, \infty)$. The determining part of M is the latter one, with $x \in [c, \infty)$. This means explicitly

$$M \approx 2^L \int_c^\infty \mathrm{d}x\, p(x) \left(1 - \frac{1}{2}\left(e^{-(x+c)^\beta} + e^{-(x-c)^\beta}\right)\right)^L \tag{2.10}$$

such that the different behavior for $\beta > 1$ and $\beta < 1$ is determined by $\frac{e^{-(x+c)^\beta}}{e^{-(x-c)^\beta}}$:

$$\lim_{x\to\infty} \frac{e^{-(x+c)^\beta}}{e^{-(x-c)^\beta}} = \begin{cases} 1, & \beta < 1 \\ 0, & \beta > 1. \end{cases}$$

This means that for $\beta < 1$ the shift by c does not contribute significantly which implies that $M \sim \frac{2^L}{L}$ if $L \gg 1$. For $\beta > 1$ on the other hand, the vanishing limit is induced by the fact that for large x: $e^{-(x-c)} \gg e^{-(x+c)}$ which justifies the following approximation of (2.10):

$$\begin{aligned} M &\approx 2^L \int_c^\infty \mathrm{d}x\, p(x) \left(1 - \frac{1}{2} e^{-(x-c)^\beta}\right)^L \\ &\approx 2^L \int_c^\infty \mathrm{d}x\, p(x) \exp\left(-\frac{L}{2} e^{-(x-c)^\beta}\right) =: 2^L \int_c^\infty \mathrm{d}x\, I(x, c) \end{aligned}$$

To get $I(x,c) > 0$ significantly, the exponent has to be of such order that $Le^{-(x-c)^\beta} \sim 1$, which leads to $x \approx c + (\log(L))^{\frac{1}{\beta}} =: x^*$. Since the integrand vanishes extremely fast for large L, the Laplace-like integral approximation at x^* is applicable and yields

$$M \approx 2^L I(x^*, c) = 2^L \left(1 - P(x^*)\right) \approx \frac{2^L}{L} e^{-\beta c \log(L)^{1-\frac{1}{\beta}}}.$$

□

A similar analysis of the leading order behavior in L can also be done for the other probability classes in dependence of the EVT index κ in addition to the above results. For the Fréchet class, the Pareto distribution $P(x) = (1 - x^{-\alpha})\,\Theta(x-1)$ (see also (A.8)) is chosen, where α corresponds to $\frac{1}{\kappa}$.

Result 2.8. *In the Fréchet class, the leading order behavior of the number of maxima is given by*

$$M \sim \frac{2^L}{L}.$$

Proof. With focus on the second integral in the calculation of M, as above, it reads

$$\begin{aligned} M &\approx 2^L \int_{1+c}^{\infty} \mathrm{d}x\; p(x) \left(1 - \frac{1}{2}\left((x+c)^{-\alpha} + (x-c)^{-\alpha}\right)\right)^L \\ &\approx 2^L \int_{1+c}^{\infty} \mathrm{d}x\; p(x) \exp\left(-\frac{1}{2} L x^{-\alpha} \left(\left(1 + \frac{c}{x}\right)^{-\alpha} + \left(1 - \frac{c}{x}\right)^{-\alpha}\right)\right). \end{aligned}$$

With a substitution $z = Lx^{-\alpha}$ the integral can be evaluated:

$$\begin{aligned} M &\approx \frac{2^L}{L} \int_0^{\frac{L}{(1+c)^\alpha}} \mathrm{d}y\; \exp\left(-\frac{y}{2}\left(\left(1 + c\left(\frac{y}{L}\right)^{\frac{1}{\alpha}}\right)^{-\alpha} + \left(1 - c\left(\frac{y}{L}\right)^{\frac{1}{\alpha}}\right)^{-\alpha}\right)\right) \\ &\overset{L \gg 1}{\sim} \frac{2^L}{L}. \end{aligned}$$

□

For the Weibull class, the *Kumaraswamy distribution* $P(x) = (1 - (1-x)^\nu)\,\Theta(x)\Theta(1-x) + \Theta(x-1)$ (see also (A.7)) with ν corresponding to $-\frac{1}{\kappa}$ is analyzed.

Result 2.9 (Number of maxima for Weibull class distributions)**.** *In a Weibull class distributed RMF-landscape, the number of maxima is asymptotically given by*

$$M \sim \frac{\left(2 - c^{-\frac{1}{\kappa}}\right)^L}{L^{-\frac{1}{\kappa}}}.$$

Proof. Assuming $c < \frac{1}{2}$, the occurring integrals will look similar to those in the calculation of the uniform case. And again, the main contribution is expected from the integral which has the upper support boundary as integration boundary. This enables the approximation

$$M \approx 2^L \int_{1-c}^{1} \mathrm{d}x\; p(x) \left(1 - \frac{1}{2}(1 - x + c)^\nu\right)^L = 2^L \int_0^c \mathrm{d}y\; \nu y^{\nu-1} \left(1 - \frac{1}{2}(y + c)^\nu\right)^L$$

in which the main contribution comes the region of small y. In order to transform to an integral which resembles the Gamma function (see (A.2)), $(2-c)^\nu$ is factored out, and then $(y+c)^\nu$ is expanded to linear order, i.e. $(y+c)^\nu \approx \nu c^{\nu-1} y + c^\nu$, leading to

$$M \approx (2 - c^\nu)^L \int_0^c \mathrm{d}y\; \nu y^{\nu-1} \exp\left(-\frac{\nu c^{\nu-1} L y}{2 - c^\nu}\right) \overset{L\gg 1}{\approx} \frac{\nu\Gamma(\nu)}{\left(\nu c^{\nu-1}\right)^\nu} \frac{\left(2 - c^\nu\right)^{L+\nu}}{L^\nu}.$$

This implies the asymptotic behavior $M \sim \frac{\left(2-e^{-\frac{1}{\kappa}}\right)^L}{L^{-\frac{1}{\kappa}}}$ for $c > 0$ since it is the same for $c > \frac{1}{2}$. □

To summarize, in the Fréchet class, the number of maxima is asymptotically independent of c and $M \sim \frac{2^L}{L}$, as it is for heavy tailed Gumbel class distributions. For light tailed Gumbel class distributions, $M \sim \frac{2^L}{L} e^{-\beta c \log(L)^{1-\frac{1}{\beta}}}$ for Weibull distributed landscapes with parameter $\beta > 1$. For the Weibull class, $M \sim \frac{\left(2-c^{-\frac{1}{\kappa}}\right)^L}{L^{-\frac{1}{\kappa}}}$. Illustrations and numerical results are presented in fig. 2.3.

In the LK-model (def. 1.10), the number of maxima depends on L in a different way. The K-parameter has to be scaled with L to resemble a similar situation as in the HoC case. Depending on the choice of the scaling, several exponential and algebraic connections between L and M can be found [74, 75, 76, 77, 78]. As shown above, for the RMF-model, various scalings can be achieved with the choice of the distribution, for example a behaviour proportional to the HoC case.

As mentioned above, the linear drift in the RMF-landscape introduces correlations between the fitness values, which is measured in a fitness correlation function [79, 80].

Result 2.10 (Correlation function of the RMF-model[1])**.** *The autocorrelation function (see sec. 3.2 for further details) of the RMF fitness landscape is given by*

$$R_d^{RMF} = \frac{\frac{c^2}{4}(L-2d) + v\delta_{d,0}}{\frac{c^2 L}{4} + v}.$$

where v denotes the variance of the random component of the landscape, which has to be finite.

Proof. Introducing angular brackets as average over the sequence space as well as over the random component ξ and additionally $\langle\bullet\rangle_d$ as the average over all sequence pairs at Hamming distance d, the correlation function reads

$$R_d = \frac{\langle(F(\sigma) - \langle F(\sigma)\rangle)(F(\sigma') - \langle F(\sigma')\rangle)\rangle_d}{\langle(F(\sigma) - \langle F(\sigma)\rangle)^2\rangle_d}$$

with a normalization which ensures $R_0 = 1$. The abbreviations $\eta(\sigma) = \xi(\sigma) - \mathbb{E}(\xi)$, $v = \mathrm{Var}(\xi) = \mathrm{Var}(\eta)$, $f(\sigma) = -cd + \eta(\sigma)$, $d = d(\sigma, \sigma^*)$ (and $d' = d(\sigma', \sigma^*)$) simplify the analysis of the correlation function:

$$\begin{aligned}\langle(F(\sigma) - \langle F(\sigma)\rangle)(F(\sigma') - \langle F(\sigma')\rangle)\rangle_r &= \langle(f(\sigma) - \langle f(\sigma)\rangle)(f(\sigma') - \langle f(\sigma')\rangle)\rangle_r \\ &= c^2[\langle dd'\rangle_r - \langle d\rangle^2] + v\delta_{\sigma,\sigma'}.\end{aligned}$$

The evaluation of $\langle dd'\rangle_r$ needs one sum over all σ and one over all σ' in distance r for each σ. Defining $k = |\{i|\sigma_i' \neq \sigma_i \wedge \sigma_i \neq \sigma_i^*\}|$, the correlator is

$$\begin{aligned}\langle dd'\rangle_r &= \frac{1}{2^L}\sum_{d=0}^{L}\binom{L}{d}\frac{d}{\binom{L}{r}}\sum_{k=0}^{r}\binom{L-d}{r-k}\binom{d}{k}(d+r-2k) \\ &= \frac{1}{2^L}\sum_{d=0}^{L}\binom{L}{d}d\left[d\left(1-\frac{2r}{L}\right)+r\right] \\ &= \frac{L^2}{4} + \frac{L}{4}\left(1-\frac{2r}{L}\right) = \langle d\rangle^2 + \frac{L}{4}\left(1-\frac{2r}{L}\right),\end{aligned}$$

where the combinatorial identities [73]

$$\sum_{k\geq 0}\binom{j}{l+k}\binom{m}{n-k} = \binom{j+m}{l+n} \quad \text{and} \quad k\binom{n}{m} = n\binom{n-1}{k-1}$$

have been used. Finally, the correlation function can be expressed as claimed. □

[1] Parts of this calculation were done by Ivan G. Szendro.

2.3 On the number of exceedances

As mentioned in the introduction, experimental fitness landscapes are quite a new tool in genetics. Since various kinds of theoretical works are present, the next step is now to identify the theoretical models which fit to the experimental results. But in the case of fitness landscapes this is not so easy due to the usually high dimension of the landscape. It is hence necessary to reduce to a few characteristic features of the fitness landscape for a comparison. One attempt is to compare the *Number of Exceedances* (NoE) [66].

Definition 2.2. Given a set of n random variables $\{X_i\}$, the *order statistics* $\{X_{(1)}, \ldots, X_{(n)}\}$ are a set of random variables defined by the sorting of the values $\{X_i\}$ in increasing order.

Definition 2.3. Given $\{X_i\}$ and $\{X_{(i)}\}$ as above, the *rank* of $X_{(i)}$ is $r(X_{(i)}) = n - i + 1$.

Definition 2.4. Let $\{x_i\}$ be a realization of n random variables. If $x_j = y$ has rank r in this set, let $\tilde{n}$ random variables be redrawn. The *number of exceedances* (NoE) is defined by $\mathcal{N}_r^{n,\tilde{n}}(y) = |\{\tilde{x}_j | \tilde{x}_j > y\}|$ where $\tilde{x}_j$ are realizations from the $\tilde{n}$ newly drawn random variables.

With these definitions the concept can be used on fitness landscapes. The random numbers of interest are the $L+1$ fitness values of the neighborhood of a sequence σ. The rank of σ, $r_\sigma(\sigma)$ is the rank of $F(\sigma)$ in the set $\{F(\sigma')|\sigma' \in \nu(\sigma)\}$. If a mutation $\sigma \to \sigma'$ occurs, it is important to distinguish between $r_\sigma(\sigma')$ which is the rank of σ' in $F(\nu(\sigma))$ and $r_{\sigma'}(\sigma')$, which is the rank of σ' in $F(\nu(\sigma'))$. Note that $\nu(\sigma) \cap \nu(\sigma') = \{\sigma, \sigma'\}$.

For i.i.d. random variables (HoC-model or RMF with $c = 0$), the distribution of the NoE can be calculated.

Result 2.11 (NoE-distribution for i.i.d. random variables)**.** *The distribution of the number of exceedances x over the mth largest among n observations in N future trials is given by [81]*

$$w(n, m, N, x) = \frac{\binom{n}{m} m \binom{N}{x}}{(N+n)\binom{N+n-1}{m+x-1}}.$$

Proof. A sketch of the proof by Gumbel and von Schelling [81] shall be given here: Consider a continuous variate ξ with distribution function P and let

ξ_m be the mth largest of n observations. The probability, that in N future trials ξ_m is exceeded x times is given by

$$w_1(P(\xi_m), N, x) = \binom{N}{x}(1 - P(\xi_m))^x P(\xi_m)^{N-x}.$$

The distribution of the frequency $P(\xi_m)$ of the mth largest among n values is given by

$$v(n, m, \xi_m)\mathrm{d}P(\xi_m) \ = \binom{n}{m} m P(\xi_m)^{n-m}(1 - P(\xi_m))^{m-1}\mathrm{d}P(\xi_m) \ .$$

To eliminate $P(\xi_m)$, the distribution of the number of exceedances is obtained by integrating:

$$w(n, m, N, x) = \int_0^1 w_1(P(\xi_m), N, x) v(n, m, F_m)\mathrm{d}P(\xi_m) \ .$$

□

Remark For $N = n$ large and m and x small, Stirling's formula yields a vast simplification:

$$w(n, m, n, x) \approx \binom{x + m - 1}{x} \left(\frac{1}{2}\right)^{m+x} .$$

This distribution is known as *negative binomial*, and has $\langle x \rangle = m$.

In the context of fitness landscapes the NoE appears, whenever a mutation occurs: If the fitness values are random variables, the *neighborhood change* discussed above leads to the following situation: a mutation $\sigma \to \sigma'$ leads to a rank change. Based on def. 2.4 the NoE can be written as

$$\mathcal{N}_{r_\sigma}^{L+1,L}(\sigma \to \sigma') = r_{\sigma'}(\sigma') - 1. \tag{2.11}$$

With the remark of res. 2.11, for the HoC-landscape (due to the i.i.d. property) the following can be shown.

Known result 2.1 (Mutational rank change [82])**.** *On a HoC-landscape, after a transition $\sigma \to \sigma'$, for the approximation $L \gg 1 \Rightarrow L \approx L + 1$ the NoE is*

$$\langle \mathcal{N} \rangle \approx r_\sigma(\sigma').$$

Proof. First mentioned by Rokyta et al. [82], Gumbel's distribution of the number of exceedances stands in direct connection to the neighborhood change problem. For $F(\sigma')$ the jth largest among the $L+1$ elements of $\nu(\sigma)$, the probability that $F(\sigma')$ is exceeded x times in $\nu(\sigma')$ is given by the probability $w(L+1, j, L, x)$. For large L, since $L \approx L+1$, the results from the remark of lemma 2.11 can be used, and the probability for exceeding $F(\sigma')$ x times in the new neighborhood is given by

$$w(L, j, x) \approx \binom{x+j-1}{x} \left(\frac{1}{2}\right)^{j+x}.$$

Since $\langle x \rangle \approx j$, $\langle r_{\sigma'}(\sigma') \rangle = j+1$. □

An application of the NoE was made on the evolution experiments with the microvirid bacteriophage ID11. Miller et al. [66] identified 9 beneficial second step mutations on the background of a mutation, named g2534t, that had been found to have the largest effect among 16 beneficial first step mutants. Assuming that the rank of g2534t among all beneficial first step mutations is at most 3, according to Gumbel and von Schelling [81] three beneficial second step mutations would have been expected if fitness values were identically and independently distributed. Thus, the observation of 9 beneficial second step mutations allowed Miller et al. [66] to reject the HoC hypothesis with high confidence ($P < 0.02$).

If the HoC-model does not describe the experimental data, the RMF-landscape might. Due to the non-isotropy of the RMF-model, the NoE will depend on the position, i.e. the distance to the reference state and on whether the adaptive step was taken in the uphill or downhill direction. Also, the mean slope c of the landscape will now matter as well as the underlying probability distribution.

In a correlated landscape, the calculation of the NoE in the evolutionary sense is much harder. Starting with the distribution equivalent to res. 2.11 yields:

Result 2.12 (NoE-distribution in an RMF-landscape [83])**.** *Let F_{c,σ^*} be an RMF-landscape on $\mathbb{H}^L$. Let its random component ξ be distributed with distribution function P and density p. The distribution of exceedances, the probability, that after a step to some sequence σ with rank m, it will be*

exceeded k times in the new neighborhood is given by:

$$\begin{aligned}
\mathbb{P}(\mathcal{N}_m = k) = \int_0^1 \sum_{i=0}^{k} &\binom{d(\sigma,\sigma_0)}{i}(1-P(\xi-c))^i P(\xi-c)^{d(\sigma,\sigma_0)} \\
&\times \binom{L-d(\sigma,\sigma_0)-1}{k-i}(1-P(\xi+c))^i P(\xi+c)^{L-d(\sigma,\sigma_0)-1-k+i} \\
&\times \left(m\sum_{i=0}^{k}\binom{d(\sigma,\sigma_0)-1}{i}(1-P(\xi-2c))^i P(\xi-2c)^{d(\sigma,\sigma_0)} \right. \\
&\times \binom{L-d(\sigma,\sigma_0)-2}{k-i}(1-P(\xi))^{i-1} P(\xi)^{L-d(\sigma,\sigma_0)-2-k+i} \\
&+ m\sum_{j=0}^{k}\binom{d(\sigma,\sigma_0)+1}{i}(1-P(\xi))^i P(\xi)^{d(\sigma,\sigma_0)+1} \\
&\left. \times \binom{L-d(\sigma,\sigma_0)}{k-i}(1-P(\xi+2c))^i P(\xi+2c)^{L-d(\sigma,\sigma_0)-k+i} \right) \mathrm{d}P(\xi)\ .
\end{aligned}$$

Since the above expression does not seem to yield a reasonable conclusion for the expected number of exceedances it might be a good idea to try another attempt.

As mentioned in def. 2.1, on an RMF-landscape F_{c,σ^*} the neighborhood $\nu(\sigma)$ with $d(\sigma,\sigma_*) = d$ is divided into the uphill neighborhood with the corresponding distribution function $P^{\uparrow}(x) = P(x + c(d-1))$, σ itself with distribution function $P^{\bullet}(x) = P(x+cd)$, and the downhill neighborhood with distribution function $P^{\downarrow}(x) = P(x+c(d+1))$. The full distribution function of fitness values is then given by [18]

$$\begin{aligned}
\Pi(x) =& \frac{1}{L+1}\left(dP^{\uparrow}(x) + P^{\bullet}(x) + (L-d)P^{\downarrow}(x)\right) \qquad (2.12)\\
=& \frac{1}{L+1}\left(dP(x+c(d-1)) + P(x+cd) + (L-d)P(x+c(d+1))\right)
\end{aligned}$$

and the expectation of the kth largest fitness value is obtained as [84]

$$\mu_k = (L+1)\binom{L}{L+1-k}\int_0^1 x\Pi(x)^{L-k+1}(1-\Pi(x))^{k-1}d\Pi(x). \qquad (2.13)$$

In general, the evaluation of this expression is complicated, because the different components of Π do not have the same support. For distributions with unbounded support, like a Gauß-distribution this is no problem, but the occurring integrals are very complicated. By using an exponential distribution $P(x) = 1 - e^{-x}$, an approximative expression can be found.

Result 2.13 (kth mean largest in an exponential RMF-landscape)**.** *In an RMF-landscape with exponential random component the mean kth largest in one neighborhood is approximately*

$$\mu_k = \log(\xi(c,d,L)) + H_{L+1} - H_{k-1} \approx \log\left(\frac{e^{-cd}}{k-1}\left(de^c + (L-d)e^{-c} + 1\right)\right), \tag{2.14}$$

with $\xi(c,d,L) = \frac{e^{-cd}}{L+1}\left(de^c + 1 + (L-d)e^{-c}\right)$.

Proof. Inserting the exponential distribution function $P(x) = 1 - e^{-x}$ into (2.12) yields

$$\Pi(x) = 1 - e^{-x+\log\left(\frac{1}{L+1}e^{-cd}\left(de^c+1+(L-d)e^{-c}\right)\right)} \tag{2.15}$$

$$=: 1 - e^{-x+\log(\xi(c,d,L))} = P(x - \log(\xi(c,d,L))) \tag{2.16}$$

with

$$\xi(c,d,L) = \frac{e^{-cd}}{L+1}\left(de^c + 1 + (L-d)e^{-c}\right). \tag{2.17}$$

The fact that (2.15) only holds on the intersection of the supports of $P^{\uparrow}$, $P^{\bullet}$ and $P^{\downarrow}$ is ignored. Instead the *common support* of $\Pi(x)$ is introduced by $[\log(\xi(c,d,L)), \infty)$, such that $\Pi(\log(\xi)) = 0$. The full distribution of fitness values defined in (2.12) is replaced by a simple exponential that is shifted in a d-dependent way (by $\log(\xi)$). The expected value $m_{k,n}$ of the kth largest out of n identically and independently exponentially distributed random variables is given by [84]

$$m_{n,k} = H_n - H_{k-1} \approx \log\left(\frac{n}{k-1}\right). \tag{2.18}$$

The integral in (2.13) can be solved after inserting the approximation (see also (A.1)). It follows that the mean of the kth largest fitness value in an RMF neighborhood is approximately given by

$$\mu_k \approx \log(\xi(c,d,L)) + H_{L+1} - H_{k-1} \approx \log\left(\frac{e^{-cd}}{k-1}\left(de^c + (L-d)e^{-c} + 1\right)\right).$$

□

In the RMF-landscape, the NoE have to be calculated for a step up and a step down separately.

Result 2.14 (NoE after a step up)**.** *After a step $\sigma \to \sigma'$ with $d(\sigma,\sigma^*) = d(\sigma',\sigma^*) - 1$ when the rank in the old neighborhood was r, the expected NoE is*

$$\begin{aligned}\mathcal{N}_r^{\text{up}} &\approx \min(k^\uparrow, d-1) + \min(k^\downarrow, L-d+1) \overset{d\gg 1}{\approx} 2 + (r-1)e^c \\ k^\uparrow &= 1 + \frac{(d-1)(r-1)e^{2c}}{de^c + 1 + (L-d)e^{-c}} \\ k^\downarrow &= 1 + \frac{(L-d+1)(r-1)}{de^c + 1 + (L-d)e^{-c}}.\end{aligned}$$

Proof. The kth largest of n i.i.d. random variables is labeled by $m_{n,k}$. For the exponential distribution $m_{n,k} = H_n - H_{k-1}$ [84]. The idea is now to compare the $m_{n,k}$ in both neighborhoods with μ_k from (2.13) to find out, how many sequences are expected to have a larger fitness in the new neighborhood. To obtain the mean Number of Exceedances, after a transition from a sequence at distance d to one at distance $d-1$ with rank r in the old NH has taken place, it has to be summed over Heaviside-theta-functions:

$$\begin{aligned}\mathcal{N}_r = & \sum_{k=1}^{d-1} \Theta(m_{k,d-1} - c(d-2) - \mu_r(L,c,d)) \\ & + \sum_{i=1}^{L-d+1} \Theta(m_{i,L-d+1} - cd - \mu_r(L,c,d)) \\ =: & \sum_{k=1}^{d-1} \Theta(a_k^\uparrow) + \sum_{i=1}^{L-d+1} \Theta(a_i^\downarrow).\end{aligned}$$

To make further simplifications, all harmonic numbers have to be approximated as logarithms. Then, the solution of

$$a_k^\uparrow = 0 \text{ and } a_i^\downarrow = 0$$

returns the values $k^\uparrow$ ($k^\downarrow$) which are the first ranks exceeding in the uphill (downhill) neighborhood. Thus, NoE is given by

$$\begin{aligned}\mathcal{N} = & \; k^\uparrow\Theta(d-1-k^\uparrow) + (d-1)\Theta(k^\uparrow - d + 1) \\ & + k^\downarrow\Theta(L-d+1-k^\downarrow) + (L-d+1)\Theta(k^\downarrow - L + d - 1) \\ k^\uparrow = & \; 1 + \frac{(d-1)(r-1)e^{2c}}{de^c + 1 + (L-d)e^{-c}} \\ k^\downarrow = & \; 1 + \frac{(L-d+1)(r-1)}{de^c + 1 + (L-d)e^{-c}}.\end{aligned}$$

The constraints $k^\uparrow \overset{!}{\leq} d-1$ and $k^\downarrow \overset{!}{\leq} L-d+1$ are not always satisfied by the approximate expressions. Incorporating these by hand leads to

$$\mathcal{N}^{\text{up}} = \min(k^\uparrow, d-1) + \min(k^\downarrow, L-d+1).$$

For $d \gg 1$ and e^c not too large the above expressions give $k^\uparrow + k^\downarrow = 2 + (r-1)e^c$ while the min-constraints can be ignored, such that $\mathcal{N}^{\text{up}} = k^\uparrow + k^\downarrow$. □

Result 2.15 (NoE after a step down)**.** *After a step $\sigma \to \sigma'$ with $d(\sigma, \sigma^*) = d(\sigma', \sigma^*) + 1$ when the rank of σ' in the old neighborhood was r, the expected NoE is*

$$\begin{aligned}
\mathcal{N}_r^{\text{down}} &= \min(k^\uparrow, d+1) + \min(k^\downarrow, L-d-1) \overset{d\gg 1}{\approx} 2 + (r-1)e^{-c} \\
k^\uparrow &= 1 + \frac{(r-1)(d+1)}{de^c + 1 + (L-d)e^{-c}} \\
k^\downarrow &= 1 + \frac{(r-1)(L-d-1)e^{-2c}}{de^c + 1 + (L-d)e^{-c}}.
\end{aligned}$$

Proof. The calculation of the NoE after a step is taken in the downhill direction is analogous to the previous one. In terms of Heaviside functions:

$$\begin{aligned}
\mathcal{N} = & \sum_{k=1}^{d+1} \Theta(m_{k,d+1} - cd - \mu_r(d,c)) \\
& + \sum_{i=1}^{L-d-1} \Theta(m_{i,L-d-1} - c(d+2) - \mu_r(d,c)).
\end{aligned}$$

Analyzing the sum, the first summand should vanish as $c \to \infty$. If now the approximation of the harmonic numbers as logarithms take place, this will not happen any more, as a $\log(0)$ term appears, which will keep the Θ function to be unity. Therefore, the first summand should be neglected, as further simplifications are not possible with the exact harmonic numbers. Again the arguments of the Theta functions give rise to linear equations which are solved for $k^\uparrow$ and $k^\downarrow$:

$$\begin{aligned}
\mathcal{N} = & \Theta(d+1-k^\uparrow) + (d+1)\Theta(k^\uparrow - d - 1) \\
& + k^\downarrow \Theta(L-d-1-k^\downarrow) + (L-d-1)\Theta(k^\downarrow - L - d - 1) - 2 \\
k^\uparrow = & \; 1 + \frac{(r-1)(d+1)}{de^c + 1 + (L-d)e^{-c}} \\
k^\downarrow = & \; 1 + \frac{(r-1)(L-d-1)e^{-2c}}{de^c + 1 + (L-d)e^{-c}}.
\end{aligned}$$

With the same arguments as before, a simplified approximation for small c and large d is found. □

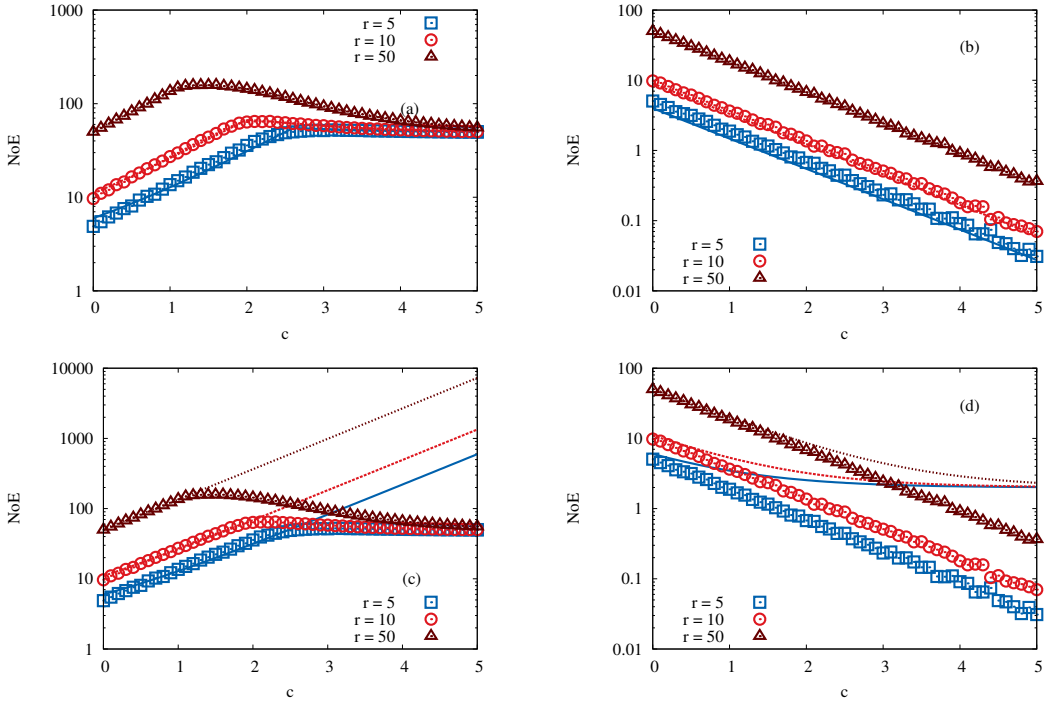


Figure 2.4: This figure shows the expected number of exceedances for a sequence which had rank r in its old neighborhood. $L = 1000$ and $d = 50$. Symbols show results from simulations and lines show plots of res. 2.14 and 2.15 for various r values. (a) shows results for a step taken uphill and the lines are a plot of the more complex approximation of result 2.14 which were evaluated numerically. (b) shows a similar plot for the situation after a step was taken down in comparison with the more complex approximation from res. 2.15. Although the analytic expression in (a) show an edge which is not visible in the data, in (a) and (b) the simulation results are well described by the analytic results. In (c) the same symbols are shown as in (a) but the lines shows the large d, intermediate c approximation from res. 2.14. (d) shows the symbols from (b) but with the large d, intermediate c approximation from res. 2.15. The limiting values of the NoE are $\mathcal{N}^{\text{up}} = \mathcal{N}^{\text{down}} = r$ for the HoC-model ($c = 0$) and $\mathcal{N}^{\text{up}} = d$, $\mathcal{N}^{\text{down}} = 0$ for a smooth landscape ($c \to \infty$). Obviously the approximation in (d) does not yield the large c limit, which can be fixed easily by omitting the 2 in the formula, at the cost of a bad small c description.

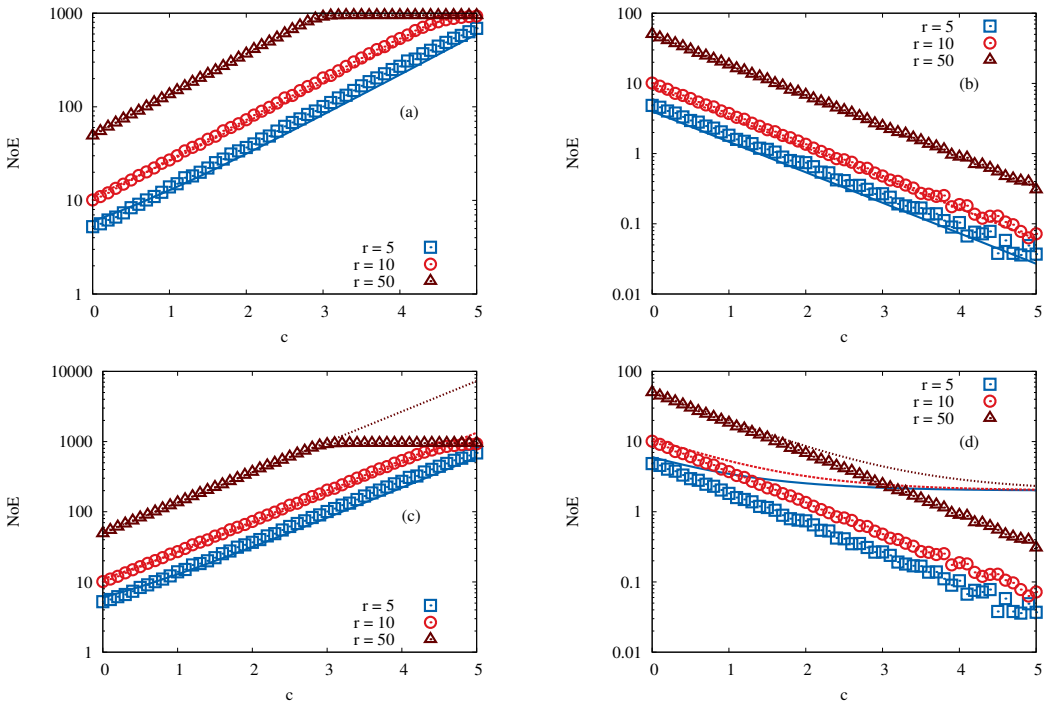


Figure 2.5: This figure shows the expected number of exceedances for a sequence which had rank r in its old neighborhood. $L = 1000$ and $d = 950$. Symbols show results from simulations and lines show plots of res. 2.14 and 2.15 for various r values. (a) shows results for a step taken uphill and the lines are a plot of the more complex approximation of result 2.14 which were evaluated numerically. (b) shows a similar plot for the situation after a step was taken down in comparison with the more complex approximation from res. 2.15. In (c) the same symbols are shown as in (a) but the lines show the large d, intermediate c approximation from res. 2.14. Obviously, there is no maximum in the data which makes the approximation better until $\mathcal{N} = L - d$ is reached. (d) shows the symbols from (b) but with the large d, intermediate c approximation from res. 2.15. Here the approximation is not as good, but for larger r the approximation becomes better for small c. In (d) the approximation converges to 2 for larger c, omitting the 2 in the approximation would lead to a better description of the data in this limit.

To summarize: For large d and small to intermediate c, the NoE is fairly well described by the simple form

$$\mathcal{N}^{\text{up}} \approx 2 + (r-1)e^{c}, \quad \mathcal{N}^{\text{down}} \approx 2 + (r-1)e^{-c}. \tag{2.19}$$

For the HoC-landscape ($c = 0$) (2.19) yield $\mathcal{N}^{\text{up}} = \mathcal{N}^{\text{down}} = r + 1$, which differs slightly from the exact result $\mathcal{N} = r$ (as shown in res. 2.1) as a consequence of the approximations involved in the derivation. Figure 2.4 compares the full expressions as well as the simple forms to numerical simulations, showing good agreement.

The evaluation of the NoE in the RMF-model shows that for the adaptive step uphill, there is a maximum for intermediate c. At this point, the NoE is considerably larger than 1, the HoC result. Hence it seems to be easy to choose a set of parameters to fit Miller et al. [66]'s data. But the problem is, that even if an RMF-model would describe the data well, it is neither known which distribution would fit, nor which distance to the reference sequence would be adequate nor if the step was taken uphill or downhill. Since the derivation of the above results does not seem to allow a generalization, the fit to the data was made with help of simulations. The distance to the reference sequence was chosen to be relatively small in an $L = 1000$, GPD RMF-landscape. The choice of the distance seems justified by the assumption, that the wild-type is already well adapted. The GPD distribution allows sampling through all probability classes. Simulations were performed by creating a landscape and choosing a sequence at $d = 50$ with rank 1 and 3 respectively. Then an adaptive step was taken with the transition probability (1.9). This was repeated while varying c until the NoE was 9, as in the experiment. The results are averaged over realizations of landscapes and sampled for c and κ. The results are shown in fig. 2.6. The parameters which are in accordance with the experimental data are $\kappa = -0.29$ (this was estimated in [66] from a maximum likelihood analysis of the fitness values of 16 first step beneficial mutations) and, depending on d, c between 0.86 and 1.04 for starting rank 1 and 0.44 and 0.76 for starting rank 3. Obviously the required c value depends strongly on the starting rank, d and κ. As seen in fig. 2.6 the appropriate c value varies from close to 0 in the Weibull regime to 10 in the Fréchet regime and seems to diverge towards $\kappa \to \frac{1}{2}$, where the variance of the GPD distribution diverges.

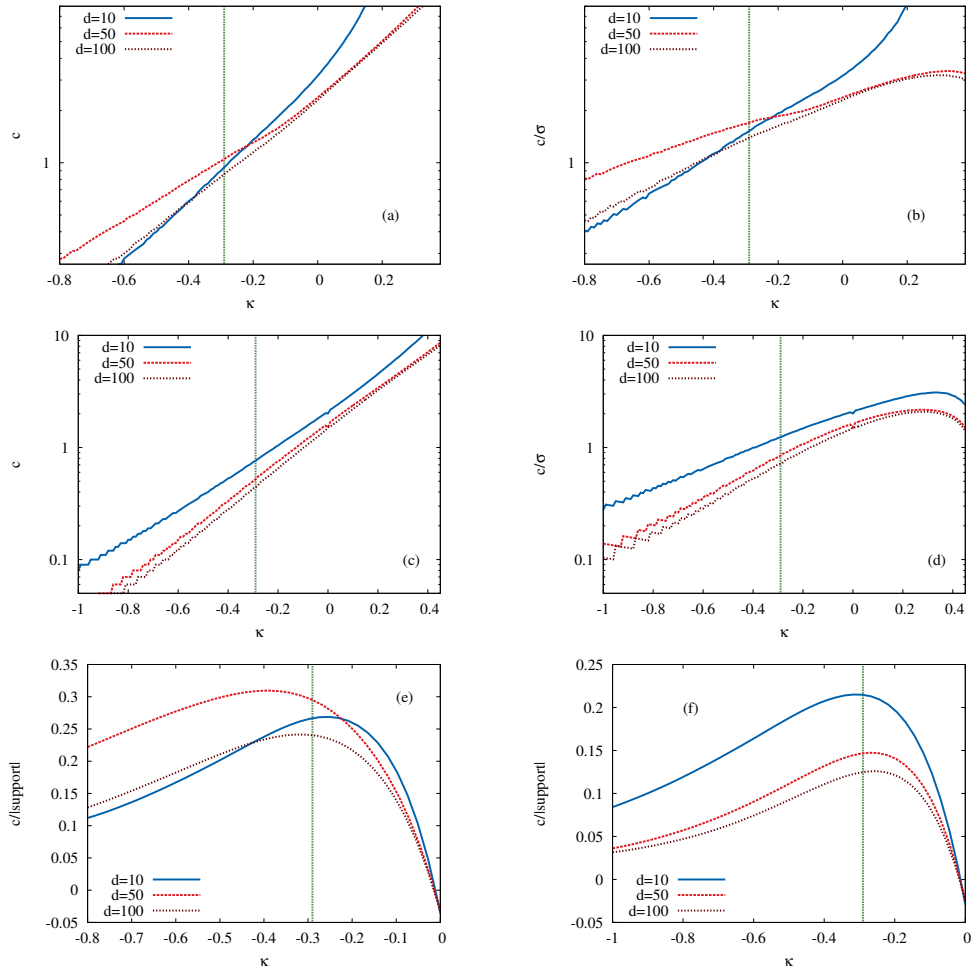


Figure 2.6: The figure shows an attempt to fit a GPD RMF-landscape to the experimental results of Miller et al. [66]. Therefore, by means of simulations, a minimal value of the RMF parameter c required to generate on average 9 exceedances after an adaptive step was found. Panel (a) and (b) show results for initial rank 1 and panel (c) and (d) for initial rank 3. Different curves correspond to different values of the initial distance d to the reference sequence. In (a) and (c) the parameter c is plotted versus κ, while in (b) and (d) c is also divided by the corresponding standard deviation $\sqrt{\mathrm{Var}(\kappa)} = \big((1-\kappa)^2(1-2\kappa)\big)^{-\frac{1}{2}}$ for a comparison between shift by c and the width of the distribution. In (e) and (f) the resulting c is divided by the corresponding support $1-\frac{1}{\kappa}$ to show, that the shift by c is still relatively small compared to it. The experimental estimate $\kappa \approx -0.29$ of the EVT index is indicated by a vertical line. The total number of loci is $L = 1000$.

Chapter 3

Amplitude spectra of fitness landscapes

"Can One Hear the Shape of a Drum?" is the title of a famous article by Kac [85]. If a drum of unknown shape is approximated as a membrane with a fixed boundary, then the domain D of the drum has Dirichlet eigenvalues λ_n. The corresponding eigenfunctions f_n can also be assumed to be unknown. Still, from the eigensystem

$$-\Delta f_n = \lambda_n f_n \tag{3.1}$$

information about the shape of the drum can be retrieved by investigation of the *spectrum* of the Laplace operator, the frequencies $\{\lambda_n\}$ at which the drumhead can vibrate and which can be heared. In the following, a similar approach is used to retrieve information about fitness landscapes and reduce the number of the used parameters. In the last part the eigensystems of experimentally measured fitness landscapes and theoretical models are used to compare and fit the both. The reasoning follows [86].

3.1 Fourier expansion and spectrum

Spectral theory can also be applied on discrete structures, like graphs. A *graph Laplacian* for L-regular graphs can be defined as

$$\Delta = A - L\mathbb{I} \tag{3.2}$$

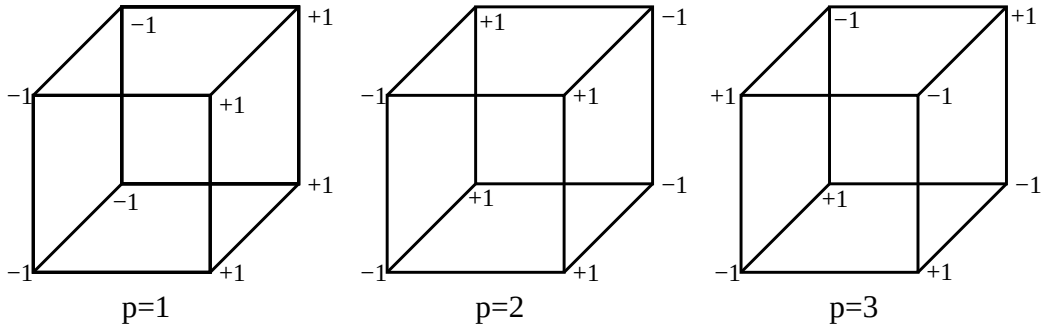


Figure 3.1: A fitness landscape can be decomposed in terms of eigenfunctions of the graph Laplacian. These so called Walsh-functions are visualized here for $L = 3$.

where $\mathbb{I}$ denotes the identity matrix. When it acts on a fitness landscape, it yields

$$\begin{aligned}\Delta F(\sigma) &= \sum_{\sigma' \in \mathbb{H}^L} A_{\sigma,\sigma'} F(\sigma') - LF(\sigma) \\ &= \sum_{\substack{\sigma' \in \mathbb{H}^L, \\ d(\sigma,\sigma')=1}} F(\sigma') - LF(\sigma). \end{aligned} \tag{3.3}$$

When σ_i denotes the i–th element of σ, the eigenfunctions of $-\Delta$ are given by the Walsh-functions $\phi_{i_1,\ldots,i_p}(\sigma) = 2^{-\frac{L}{2}} \sigma_{i_1} \ldots \sigma_{i_p}$ with $p \in \{1, \ldots, L\}$, $0 < i_1 \leq i_2 \cdots \leq i_p \leq L$ and $\phi_0 = 1$. The corresponding eigenvalues are $\lambda_p = -2p$ which are hence $\binom{L}{p}$ fold degenerate. The set of all eigenfunctions $\phi_i(\sigma)$ forms an orthonormal basis. See fig. 3.1 for the visualization of three eigenfunctions on an $L = 3$ hypercube [49].

Definition 3.1. Every fitness landscape on the binary hypercube can be decomposed in terms of Walsh-functions. This decomposition is called *Fourier expansion* [87], which reads

$$F(\sigma) = \sum_{p=0}^{L} \sum_{i_1 \ldots i_p} a_{i_1 \ldots i_p} \phi_{i_1 \ldots i_p}(\sigma). \tag{3.4}$$

The Fourier expansion of a fitness landscape motivates a change of variables. Because of the orthonormality of the Walsh functions, multiplication yields the coefficients $\sum_{\sigma \in \mathbb{H}^L} \phi_{i_1,\ldots,i_p} F(\sigma) = a_{i_1,\ldots,i_p}$. No information is lost, but it is now contained into the $a_{i_1,\ldots,i_p}$. Since the variability of the Walsh-functions increases with increasing order p, the

coefficients provide information about the fitness interactions in the genome. While the a_{i_1}'s contain the information about the relative influence of the non-epistatic contributions on fitness, the higher order coefficients $a_{i_1 \ldots i_p}$ with $p > 1$ describe the relative strength of the contributions of p–tupels of interacting loci. The zero order coefficient a_0 is proportional to the mean fitness of the landscape,

$$a_0 = 2^{-\frac{L}{2}} \sum_{\sigma \in \mathbb{H}^{\mathbb{L}}} F(\sigma),$$

where the pre-factor reflects the normalization of the Walsh functions. It is obvious, that for certain landscapes, the transformation into the coefficients can yield great insights into the structure of the landscape, e.g. if certain orders vanish completely. But generally, one is still left with 2^L parameters.

3.2 The amplitude spectra

It is useful to reduce the number of parameters in a clever way to remain with as much information as possible. One possibility of doing so is the introduction of the *amplitude spectra* as done by Stadler and Happel [88].

Definition 3.2. The *amplitude spectrum* of a fitness landscape is calculated by averaging and normalizing the coefficients of the Fourier expansion (def. 3.1). The order of normalization and averaging distinguishes between two possible choices of amplitude spectra. The first one reads

$$B_p = \left\langle \frac{\sum_{i_1 \ldots i_p} |a_{i_1 \ldots i_p}|^2}{\sum_{q \neq 0} \sum_{i_1 \ldots i_q} |a_{i_1 \ldots i_q}|^2} \right\rangle$$

for $p > 0$ and $B_0 = 0$, where the angular brackets denote an average over realizations of landscapes. The second one is

$$\tilde{B}_p = \frac{b_p}{b_0 + \sum_{q \neq 0} b_q}$$

with $b_p = \sum_{i_1 \ldots i_p} \langle |a_{i_1 \ldots i_p}|^2 \rangle$ for all $p \geq 1$. b_0 is not defined in terms of the Fourier coefficients a_i, but is proportional to the mean covariance,

$$b_0 = 2^{-L} \sum_{\sigma, \sigma' \in \mathbb{H}^L} \left[\langle F(\sigma) F(\sigma') \rangle - \langle F(\sigma) \rangle \langle F(\sigma') \rangle \right], \tag{3.5}$$

as defined[1] in [49].

[1] The prefactor of b_0 given in [49] appears to be incorrect.

For additive fitness landscapes $B_1 = 1$ and $B_{\text{sum}} = \sum_{i>1} B_i = 0$ while for a landscape with epistasis $B_{\text{sum}} > 0$. In [12] B_{sum} was used as a quantifier for epistasis in experimentally obtained fitness landscapes.

As shown by Stadler and Wagner [49] the amplitude spectra are closely connected to the correlations of the fitness landscapes. Similar to the definitions of the amplitude spectra they differ by the order of averaging. The *direct correlation function* describes correlations of sequence pairs of Hamming distance d as

$$\rho_d = \frac{1}{\binom{L}{d} 2^L} \sum_{\substack{\sigma,\sigma' \in \mathbb{H}^L \\ d(\sigma,\sigma')=d}} \frac{(F(\sigma) - \overline{F})(F(\sigma') - \overline{F})}{\overline{F^2} - \overline{F}^2}, \tag{3.6}$$

where "$\overline{\ }$" denotes the average over *one* landscape realization. It is connected to B_p according to

$$\langle \rho_d \rangle = \sum_{p \geq 0} B_p \omega_p(d) \tag{3.7}$$

where the ω_p are orthogonal functions in the sense, that $\sum_{d \geq 0} C_d \omega_p(d) \omega_q(d) \sim \delta_{pq}$ with a combinatorial weight C_d. The exact form depends on the underlying graph structure [88]. The *autocorrelation function* R_d is defined as

$$R_d = \frac{\langle F(\sigma) F(\sigma') \rangle_d - \left\langle \overline{F} \right\rangle^2}{\left\langle \overline{F^2} \right\rangle - \left\langle \overline{F} \right\rangle^2}, \tag{3.8}$$

where $\langle \ldots \rangle_d$ denotes a *simultaneous* average over all possible pairs (σ, σ') with $d(\sigma, \sigma') = d$ as well as over the realizations of the landscape. Note, that the original definition was restricted to landscapes with $\forall_{\sigma \in \mathbb{H}^L} : \langle F(\sigma) \rangle = a_0$. This is not fulfilled for all landscapes considered in this work. Nevertheless, the original definition was also working on *partitions* of the underlying graph and thus more general. Since here only the full graph is of interest, the underlying theorem 5 of Stadler and Happel [88] which deals with the spatial average of fitness values is still valid, although it is restricted to landscapes with constant mean fitness. This restriction is not needed if the only partition of interest is the full hypercube. R_d is linked to the amplitude spectrum $\tilde{B}_p$ according to [89]

$$R_d = \sum_{p \geq 0} \tilde{B}_p \omega_p(d). \tag{3.9}$$

On the hypercube $\omega_p(d)$ are closely related to the *Krawtchouk polynomials* K_{pd} [49]:

$$\omega_p(d) = \binom{L}{p}^{-1} K_{pd},$$

where [90, 91]

$$K_{pd} = \sum_{j\geq 0} (-1)^j \binom{d}{j}\binom{L-d}{p-j}. \tag{3.10}$$

The Krawtchouk polynomials can be understood as an $L \times L$ matrix K with elements K_{pd}. The amplitude spectrum and the correlation function can be written as vectors $\tilde{B}, R \in \mathbb{R}^L$ with elements $\tilde{B}_i$ and R_d respectively. Then (3.9) can be rewritten as

$$K\tilde{B} = R. \tag{3.11}$$

Here, the focus is on the calculation of analytical expressions of the $\tilde{B}_p$ for known R_d. Thus, an inversion is needed. Defining $\tilde{K}_{pd} = \binom{L}{d} K_{pd}$, multiplication from the left to (3.9) results in

$$\tilde{K}^T K\tilde{B} = \tilde{K}^T \cdot R. \tag{3.12}$$

Since the Krawtchouk Polynomials are known to be orthogonal in the sense that (see e.g. [92])

$$\langle K_p, K_q\rangle = \sum_{d\geq 0} \binom{L}{d} K_{pd}K_{qd} = 2^L \binom{L}{p}\delta_{pq}, \tag{3.13}$$

the last step yields in components:

$$\begin{aligned}\left(\tilde{K}^T K\tilde{B}\right)_q =& \sum_{d\geq 0}\sum_{i\geq 0} \binom{L}{d}(-1)^i \binom{d}{i}\binom{L-d}{q-i} \times \\ & \sum_{p\geq 0}\binom{L}{p}^{-1} \sum_{j\geq 0}(-1)^j \binom{d}{j}\binom{L-d}{p-j}\tilde{B}_p \end{aligned} \tag{3.14}$$

$$= \sum_{p\geq 0}\binom{L}{p}^{-1} 2^L \delta_{qp}\binom{L}{p}\tilde{B}_p = 2^L \tilde{B}_q = \left(\tilde{K}^T R\right)_q. \tag{3.15}$$

Thus, the application of $\tilde{K}^T$ results in an inversion. In general, it follows directly, that $2^{-L}\tilde{K}^T = K^{-1}$ and

$$\tilde{B}_q = 2^{-L}\sum_{d\geq 0} K_{qd}\binom{L}{d} R_d. \tag{3.16}$$

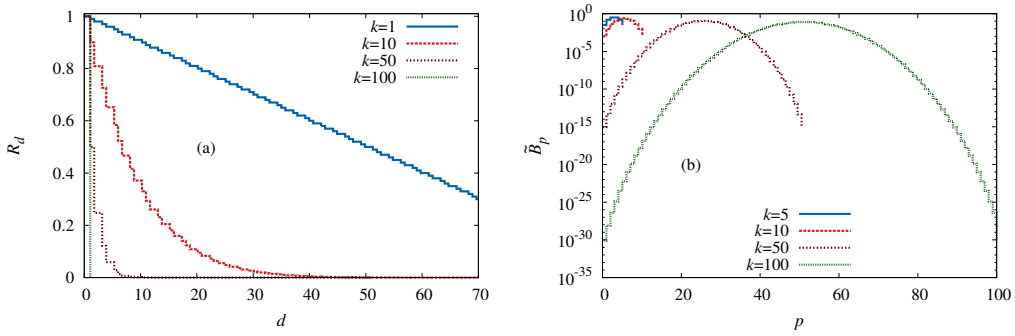


Figure 3.2: The autocorrelation function (a) and the amplitude spectrum (b) of the LK-model with $L = 100$ and different values of k.

Now, the calculation of amplitude spectra from the autocorrelation functions is possible and at least numerically any spectrum can be calculated from a given correlation function. But for some landscape models analytical solutions can be calculated, as will be shown in the following.

3.3 LK-model

The LK-model was defined in def. 1.10. Although the interpretation of the parameters is not as clear as in the RMF-case, its autocorrelation function has the remarkably simple form

$$R_d^{\mathrm{LK}} = \binom{L-k}{d}\binom{L}{d}^{-1}, \tag{3.17}$$

which was calculated by Campos et al. [93], see fig. 3.2. A very interesting point is that R_d neither depends on the underlying probability distribution, nor on the choice of the LK-neighborhood[2].

Result 3.1 (Amplitude Spectra of LK-landscapes)**.** *The amplitude spectrum of an LK-landscape is given by*

$$\tilde{B}_q = 2^{-k}\binom{k}{q}.$$

[2]This is in contrast to results found in the literature. Some incorrect expressions for the correlation functions have been reported [94]. The erroneus conclusion that the choice of the LK-neighborhood affects the amplitude spectra [49] is based on that.

Proof. Inserting (3.17) into (3.16) yields

$$\tilde{B}_q = 2^{-L} \sum_{d \geq 0} K_{qd} \binom{L-k}{d}. \tag{3.18}$$

Stoll [91] gives an alternative but equivalent formulation for the Krawtchouk polynomials:

$$K_{qd}^{(2)} = \sum_{i \geq 0} (-2)^i \binom{d}{i} \binom{L-i}{q-i}.$$

Inserting this into (3.18) yields

$$\begin{aligned} \tilde{B}_q &= \sum_{d \geq 0} K_q(d) \binom{L-k}{d} \\ &= 2^{-L} \sum_{i \geq 0} \sum_{d \geq 0} (-2)^i \binom{d}{i} \binom{L-i}{q-i} \binom{L-k}{d}. \end{aligned}$$

The identity [95]

$$\sum_{d \geq 0} \binom{d}{i} \binom{L-k}{d} = 2^{L-k-i} \binom{L-k}{i},$$

helps to calculate the sum:

$$\begin{aligned} \tilde{B}_q &= 2^{-L} \sum_{i \geq 0} (-2)^i \binom{L-i}{q-i} \binom{L-k}{i} 2^{L-k-i} \\ &= 2^{-k} \sum_{i \geq 0} (-1)^i \binom{L-i}{q-i} \binom{L-k}{i}. \end{aligned} \tag{3.19}$$

Although (A.4) claims positivity on the entries of the binomial coefficients, a very helpful trick is to neglect this during the calculation. This allows an 'upper negation' [73] in the first binomial factors in (3.19),

$$\binom{L-i}{q-i} = (-1)^{q-i} \binom{q-L-1}{q-i}.$$

Using the Vandermonde identity [73] the remaining sum can be calculated:

$$\begin{aligned} \tilde{B}_q &= 2^{-k} (-1)^q \sum_{i \geq 0} \binom{q-L-1}{q-i} \binom{L-k}{i} \\ &= 2^{-k} (-1)^q \binom{q-k-1}{q} \end{aligned}$$

A second upper negation leads to the final result

$$\begin{aligned}\tilde{B}_q &= 2^{-k}\sum_{i\geq 0}(-1)^i(-1)^{q-i}\binom{L-k}{i}\binom{q-i-(L-i)-1}{q-i}\\ &= 2^{-k}(-1)^q\sum_{i\geq 0}\binom{L-k}{i}\binom{q-L-1}{q-i}\\ &= 2^{-k}(-1)^q\binom{q-k-1}{q}\\ &= 2^{-k}\binom{k}{q}.\end{aligned}$$

□

As the autocorrelation, this result is characterized by remarkable simplicity (see fig. 3.2 for illustration). The amplitude spectrum vanishes for $q > k$ as expected [24, 96] and the known case of the HoC-model ($\tilde{B}_q = 2^{-L}\binom{L}{q}$) is reproduced for $k = L$. The spectra satisfy the symmetry $\tilde{B}_q = \tilde{B}_{k-q}$ and are maximal for $q = k/2$, as was conjectured in [49].

A straightforward generalization of LK-landscapes is achieved by construction of *superpositions* of LK-models, in the sense that LK-landscapes are added independently. Let $\{F_m(\sigma) = \frac{1}{\sqrt{L}}\sum_j f_j^{(m)}(\sigma_{j_1},\ldots,\sigma_{j_{k^{(m)}}})\}$ be a family of n LK fitness landscapes with neighborhood sizes $k^{(m)}$, $m = 1, ..., n$. Then its superposition $\mathcal{F}$ is defined by

$$\begin{aligned}\mathcal{F} : \sigma &\mapsto \sum_{m=1}^{n} F_m(\sigma)\\ &= \sum_{m=1}^{n}\frac{1}{\sqrt{L}}\sum_{j=1}^{L} f_j^{(m)}(\sigma_{j_1},\ldots,\sigma_{j_{k^{(m)}}}).\end{aligned} \tag{3.20}$$

Since the different LK-landscapes $\{F_m\}$ are independent, the correlation functions are additive,

$$R_d^{\mathcal{F}} = \frac{\sum_{m=0}^{n}\binom{L-k^{(m)}}{d}\binom{L}{d}^{-1}D_m}{\sum_{j=0}^{n}D_j} =: \sum_{i=1}^{L} C_i\binom{L-i}{d}\binom{L}{d}^{-1}, \tag{3.21}$$

with statistical weights

$$C_i = \sum_{\{m|k^{(m)}=i\}}\frac{D_m}{\sum_{j=0}^{n}D_j},$$

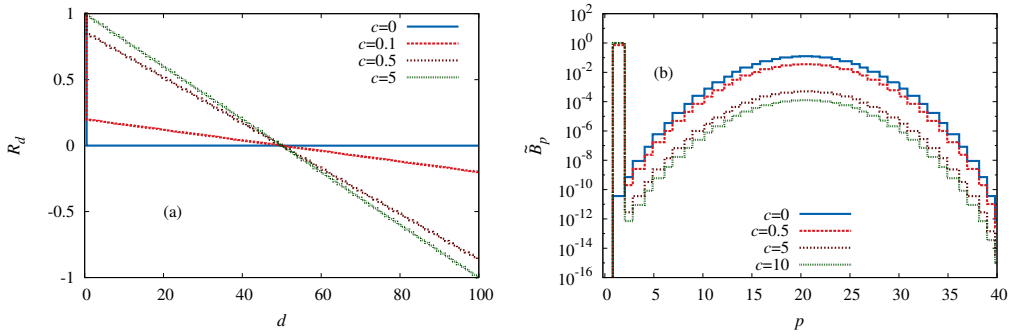


Figure 3.3: The autocorrelation function (a) and the amplitude spectrum (b) of the RMF-model with $L = 100$, $D_1 = 0$, $D_L = 1$ and various values of c.

where $D_m = \mathrm{Var}(f^{(m)})$ and the sum is over all landscapes with neighborhoods of size i. The amplitude spectrum of the superposition is thus of the form

$$\tilde{B}_p^{\mathcal{F}} = \sum_{i \geq 0} 2^{-i} C_i \binom{i}{p}. \tag{3.22}$$

The consistent interpretation of an empirical fitness landscape as a superposition of LK-landscapes requires all C_i to be positive. Nevertheless, it can be useful to consider superpositions containing negative C_i to calculate amplitude spectra of fitness landscapes constructed by different means (see section 3.4 for an example).

3.4 RMF-model

The RMF-model used here was introduced in def. 1.8 and its autocorrelation function was calculated in sec. 2.2. The amplitude spectrum of RMF-landscapes can be calculated by using the fact, that the RMF-model is a superposition of an additive with a HoC-landscape. Since an LK-model with $k = 1$ is additive and $k = L$ is a HoC-landscape, the RMF-landscape can be understood as a generalized LK-model in the sense of (3.20).

Result 3.2 (Amplitude spectrum of an RMF-landscape)**.** *The amplitude spectrum of an RMF-landscape is given by*

$$\tilde{B}_p^{\mathrm{RMF}} = \frac{\left(D_1 + \frac{c^2}{4}\right) L \delta_{p1} + D_L 2^{-L} \binom{L}{p}}{\left(D_1 + \frac{c^2}{4}\right) L + D_L},$$

with D_i as before.

Proof. Writing the autocorrelation function R_d^{RMF} (see res. 2.10) as a linear combination of correlation functions of the LK-model with different ks yields $R_d^{\mathrm{RMF}} = \sum_{k=0}^{L} C_k \binom{L-k}{d} / \binom{L}{d}$ with expansion coefficients

$$C_0 = -\frac{\left(D_1 + \frac{c^2}{4}\right) L}{\left(D_1 + \frac{c^2}{4}\right) L + D_L},\ C_1 = \frac{2\left(D_1 + \frac{c^2}{4}\right) L}{\left(D_1 + \frac{c^2}{4}\right) L + D_L},$$
$$C_L = \frac{D_L}{\left(D_1 + \frac{c^2}{4}\right) L + D_L}, \tag{3.23}$$

and $C_k = 0$ for all other ks. Exploiting the linearity of (3.16) the spectrum follows immediately:

$$\tilde{B}_p^{\mathrm{RMF}} = \frac{\left(D_1 + \frac{c^2}{4}\right) L\delta_{p1} + D_L 2^{-L}\binom{L}{p}}{\left(D_1 + \frac{c^2}{4}\right) L + D_L}. \tag{3.24}$$

□

See fig. 3.3 for an illustration of the autocorrelation functions and amplitude spectra for the RMF-model with $D_1 = 0$. The zeroth order coefficient is something like a shift in fitness. It does not contain epistatic information and can thus be chosen freely to fit the model. In the LK superposition picture, RMF-models always have $C_1, C_L > 0$ and $C_i = 0$ for $1 < i < L$.

3.5 Applications & experimental results

In sec. 2.3 the NoE was used to fit an RMF-landscape to experimental data. There, only isolated steps of adaptation where present. Now, the full fitness landscape will be fitted to experimentally obtained fitness landscapes with help of the amplitude spectra of the generalized LK-landscapes. Although in experiments it is not always Wrightian or Malthusian fitness which is measured, but some proxy of it as described in the introduction. The discussed landscapes are the $L = 6$ growth rate landscape from Hall et al. [68] of yeast, the $L = 8$ growth rate landscape of the fungus *Aspergillus niger* presented by Franke et al. [69] and additionally two $L = 9$ landscapes which measure the output of certain enzymes in *Nicotiana Tobaccum*.

In order to fit generalized LK-models to these fitness landscapes, it is helpful to use the following guidelines to achieve reasonable fits via the coefficients C_i[3]:

[3]The fitting was performed by Ivan G. Szendro

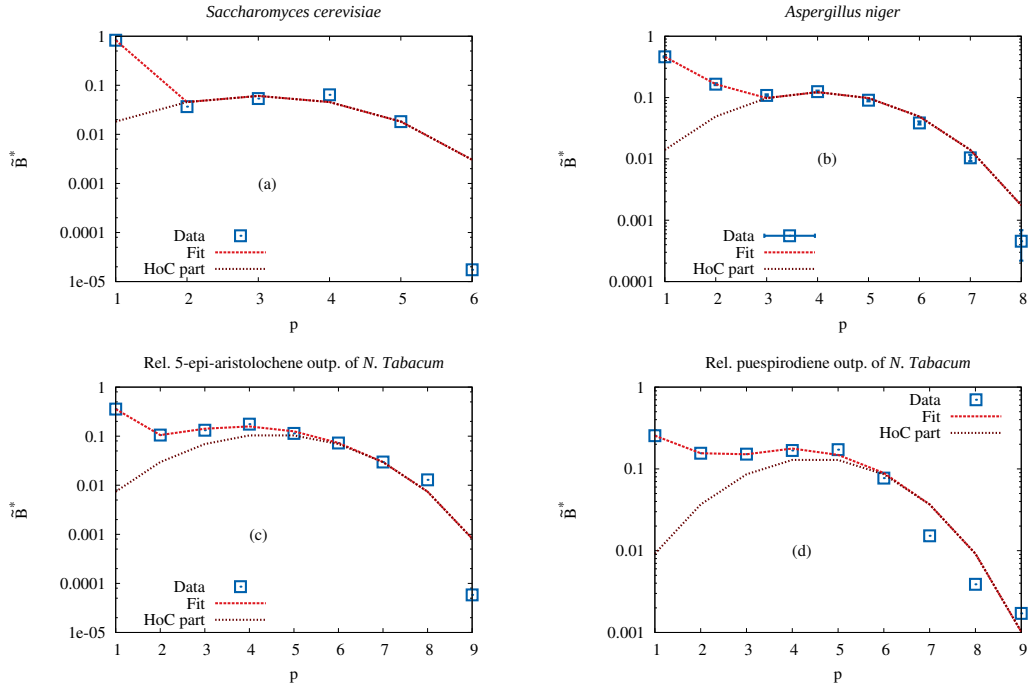


Figure 3.4: Spectra corresponding to various experimentally measured fitness landscapes. Blue boxes are retrieved from the experimental data. By fitting the spectrum of a superposition of LK-models to the data, a landscape was created which should resemble the properties of the measured landscapes, the fit is presented as a red line. To demonstrate the strength of the HoC component, the brown line, proportional to $\binom{L}{p}$, showing the spectrum expected for a HoC component, is plotted for comparison.

- for a good comparability, a renormalized amplitude spectrum

$$\tilde{B}_p^* = \frac{\tilde{B}_p}{\sum_{q>0} \tilde{B}_q} = \frac{b_p}{\sum_{q>0} b_q}. \tag{3.25}$$

  should be considered,

- the fit should have as few nonzero C_i as possible,
- $\forall_i : C_i \geq 0$,
- recursive fitting might be needed, in the sense, that a fit is done, the negative C_i and the very small C_i (compared to the rest) are constrained to 0 and another fit is done,
- C_0 is not needed in the fit as it can be fitted trivially (it is still just an additive constant).

Fig. 3.4 shows the data for the normalized amplitudes $\tilde{B}_p^*$ (blue points) with the fit (red curve), the HoC component $\propto \binom{L}{p}$ is separately shown (brown curve). It is obvious, that the HoC component seems to be very strong. For the *A. niger* landscape the errors are expected to be [26] too small to explain the strong HoC component as stemming from noise in the measurement.

As shown in fig. 3.4(a), the spectrum of the yeast landscape [68] is nicely fitted with only C_1 and C_L chosen different from zero. This is, as mentioned earlier the same spectrum as of an RMF-landscape. Only the value at $p = L$ seems too small to be fitted by this model. However, this value corresponds to a single component of the decomposition (3.4) and the large deviation may be due to the lack of averaging. A nice and sparse fit with nonzero coefficients C_1, C_2, and C_L is obtained for the *A. niger* landscape from Franke et al. [69] (see fig. 3.4(b)). The significant value of C_2 implies that there are important interactions between pairs of loci which rejects the modeling with an RMF type landscape, although the number of accessible pathways would allow this [69].

The spectrum of the 5-epi-aristolochene *N. tabaccum* landscape from [8] is fitted quite well with C_1, C_2, C_6 and C_L different from 0 (see fig. 3.4(c)). This might indicate, that there are additionally one or several groups consisting of 6 strongly interacting alleles. The premnaspirodiene landscape yields less convincing results, as the large p part of the spectrum seems to be poorly fitted (see fig. 3.4(d)). Introducing more components into the fitting Ansatz yields better results for this part of the spectrum, but such approaches can hardly be considered sparse anymore.

It is important, that a bad fit as above does not a priori mean, that the fitness landscape is not of the generalized LK type. For example, it might be, that the measured fitness or fitness proxy is inappropriately in the sense, that the measuring mechanism alters it. If $F'(\sigma)$ means the 'real' fitness (proxy) which would yield a sparse fit then the measurement procedure could act as a nonlinear transformation G, such that $F = G \circ F'$ does not fit sparsely.

Chapter 4

Adaptive walks

The basic concept of *adaptive walks* is explained in sec. 1.4. In the following, besides a brief recollection of known results, new results for adaptive walks on uncorrelated as well as on correlated landscapes will be given.

4.1 Previous work

Several properties of adaptive walks have been calculated, especially on a HoC-landscape. There, the GAW is known to have a constant walk length $\ell = e - 1 \approx 1.7$, independent of the sequence length L [97] and the underlying distribution function. While for the RAW a logarithmic dependency exists: $\ell \approx \log(L) + 1.09931$ [59]. This is also true for the NAW. *Gillespie's model* [31] was used for the calculation: The neighborhood change after every step is ignored and the transition probability (1.9) is used on the $L + 1$ random numbers of one neighborhood (see fig. 4.1). The result is $\ell \approx \frac{1}{2}\log(L) + \frac{1}{2}(\gamma + 1) + 0.44$ [30, 98] but is restricted to landscapes which are distributed according to a Gumbel class distribution. The independence of the distribution as long, as it is in the Gumbel class is due to the fact, that particularly high fitness values are preferred by the transition probability. This increases the importance of the tail of the distribution and motivates a treatment in terms of EVT.

It has also been shown, that for a more general underlying distribution, like the GPD, the walk length is given by

$$\ell_s \approx \frac{1-\kappa}{2-\kappa}\log(s) + \text{const.} \tag{4.1}$$

from which the RAW and GAW results are retrieved as limits $\kappa \to -\infty$ and $\kappa \to 1$ [99, 83, 100]. s is the rank of the sequence the process started on.

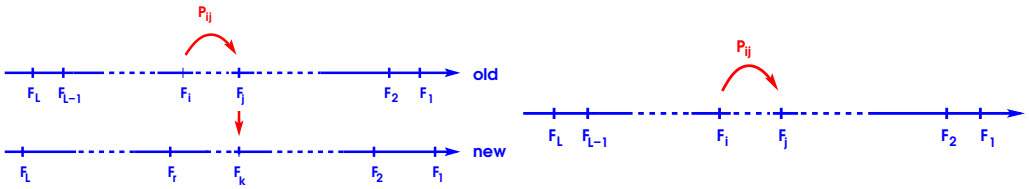


Figure 4.1: This figure visualizes the Gillespie approximation, in which the neighborhood change is ignored. Usually, a step is made from the ith to the jth fittest sequence in one neighborhood. Then the new neighbors occur, which will on average change the rank from j to j' (on the left). In Gillespie's approximation, only the step from i to j is taken into account, such that the process stops when it reaches rank 1 (on the right).

4.2 The GPD approach

In previous work, the results for other extreme value classes than Gumbel were calculated by using different distributions for each class and assemble them afterwards. Here the results shall be calculated directly from the GPD.

Following the calculation in [83], an approximative solution is yielded using Gillespie's model. The process starts at a certain *starting rank* s and stops when rank 1 is reached. ℓ is then the number of steps taken from s to 1. In the following, F_k will label the kth largest value in the neighborhood. The *spacings* are defined by $\Delta_k = F_k - F_{k+1}$. Gillespie [30] calculated the mean walk length in one realization of the landscape, which can be expressed in terms of $\lambda_n = \sum_{k=1}^{n-1} k\Delta_k$ as

$$l_s = \sum_{j=1}^{s-1} \left[\frac{1}{j} - \frac{\lambda_j}{\lambda_s} \frac{1}{s-1} - \sum_{n=j+1}^{s-1} \frac{\lambda_j}{\lambda_n} \frac{1}{n(n-1)} \right]. \tag{4.2}$$

Result 4.1 (Mean kth largest of i.i.d. GPD variates)**.** *The mean kth largest value of n i.i.d. random variables with GPD function $GPD_{\kappa,1,0}$ is given by*

$$\mu_{k,n} = n \binom{n-1}{n-k} \frac{1}{\kappa} \left(B(n-k+1, k-\kappa) - B(n-k+1, k) \right)$$

with the Beta-function $B(x, y)$ (see also (A.3)*).*

Proof. For simplicity, define $P_\kappa(x) = \mathrm{GPD}_{\kappa,1,0}(x) = 1 - (1+\kappa x)^{-\frac{1}{\kappa}}$. Then

the mean kth largest value is calculated straightforwardly [84]:

$$\begin{aligned}
\mu_{k,n} &= n\binom{n-1}{n-k}\int_0^1 xP_\kappa(x)^{n-k}(1-P_\kappa(x))^{k-1}\mathrm{d}P_\kappa(x) \\
&= n\binom{n-1}{n-k}\int_0^1 \frac{1}{\kappa}(1+\kappa x-1)P_\kappa(x)^{n-k}(1+\kappa x)^{(k-1)(-\frac{1}{\kappa})}\mathrm{d}P_\kappa(x) \\
&= n\binom{n-1}{n-k}\frac{1}{\kappa}\int_0^1 P_\kappa^{n-k}(1-P_\kappa(x))^{\kappa+k-1} - P_\kappa^{n-k}(1-P(x))^{k-1}\mathrm{d}P_\kappa(x) \\
&= n\binom{n-1}{n-k}\frac{1}{\kappa}\left(B(n-k+1,k-\kappa)-B(n-k+1,k)\right).
\end{aligned}$$

The appropriate combinatorial factor is $n\binom{n-1}{n-k}$. □

Result 4.2 (Spacings of i.i.d. GPD variates)**.** *The mean kth spacing* $\langle\Delta_k\rangle$ *of* n *i.i.d.* $GPD_{\kappa,1,0}$ *distributed random variables is given by*

$$\langle\Delta_k\rangle = \frac{\Gamma(n+1)\Gamma(k-\kappa)}{\Gamma(k+1)\Gamma(n-\kappa+1}.$$

Proof. The Beta-function can alternatively defined via the Γ-function: $B(x,y)=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$. With res. 4.1 this leads to

$$\begin{aligned}
\mu_{k,n} &= n\binom{n-1}{n-k}\frac{1}{\kappa}\left(B(n-k+1,k-\kappa)-B(n-k+1,k)\right) \\
&= \frac{n}{\kappa}\frac{\Gamma(n)}{\Gamma(n-k+1)\Gamma(k)}\left[\frac{\Gamma(n-k+1)\Gamma(k-\kappa)}{\Gamma(n-\kappa+1)}-\frac{\Gamma(n-k+1)\Gamma(k)}{\Gamma(n+1)}\right] \\
&= \frac{1}{\kappa}\left[\frac{\Gamma(n+1)\Gamma(k-\kappa)}{\Gamma(k)\Gamma(n-\kappa+1)}-1\right]
\end{aligned}$$

where the identities $n\Gamma(n)=\Gamma(n+1)$ and $\binom{n}{k}=\frac{\Gamma(n-1)}{\Gamma(k-1)\Gamma(n-k)}$ where used. From this, the spacings are calculated:

$$\begin{aligned}
\langle\Delta_k\rangle &= \mu_{k,n}-\mu_{k+1,n} \\
&= \frac{1}{\kappa}\left[\frac{\Gamma(n+1)\Gamma(k-\kappa)}{\Gamma(k)\Gamma(n-\kappa+1)}-1\right]-\frac{1}{\kappa}\left[\frac{\Gamma(n+1)\Gamma(k+1-\kappa)}{\Gamma(k+1)\Gamma(n-\kappa+1)}-1\right] \\
&= \frac{1}{\kappa}\frac{\Gamma(n+1)\Gamma(k-\kappa)k-\Gamma(n+1)\Gamma(k+1-\kappa)}{\Gamma(k+1)\Gamma(n-\kappa+1)} \\
&= \frac{\Gamma(n+1)\Gamma(k-\kappa)}{\Gamma(k+1)\Gamma(n-\kappa+1)}.
\end{aligned}$$

□

Following Gillespie [30] $\{\lambda_k\}$ will be a set of i.i.d. random variables if L is large. Using the approximation $\langle\frac{\lambda_i}{\lambda_j}\rangle \approx \frac{\langle\lambda_i\rangle}{\langle\lambda_j\rangle}$, the mean walk length can be calculated.

Result 4.3 (Mean walk length in a GPD distributed HoC-landscape)**.** *The mean natural adaptive walk length in a $GPD_{\kappa,1,0}$ distributed HoC-landscape from starting rank s is approximately*

$$\ell_s \approx \frac{1-\kappa}{2-\kappa}\log(s) + const.$$

Proof. The calculation needs lemmas 9, 10, 11, 12 and part of the proof of result 2.2 from [83]:

1. $\sum_{k=1}^{n}\frac{\Gamma(k+a)}{\Gamma(k)} = \frac{n\Gamma(n+1+a)}{(1+a)\Gamma(n+1)}$
2. $\sum_{k=1}^{n}\frac{\Gamma(k+a)}{\Gamma(k-1)} = \frac{(n-1)\Gamma(n+1+a)}{(2+a)\Gamma(n)}$
3. $B(x,y) \approx x^{-y}e^{y}\Gamma(y)$ for $x \gg y$.
4. $\sum_{j=1}^{i-1}\sum_{n=j+1}^{i-1}\left(\frac{j-1}{n-1}\right)^{\kappa}\frac{1}{n(n-1)} = \sum_{j=2}^{i-2}\frac{H_{j-1,-\kappa}}{j^{\kappa}j(j+1)}$ with the generalized harmonic numbers $H_{a,b} = \sum_{k=1}^{a}\frac{1}{k^b}$
5. $H_{j-1,k} \approx j^k\left(\frac{j}{k+1} - \frac{1}{2}\right)$

In the following, these expressions will be called by their number in this enumeration. For convenience all $\langle\bullet\rangle$ are omitted since all expressions are averaged.

$$\begin{aligned}\lambda_i &= \sum_{k=1}^{i-1}\frac{\Gamma(n+1)\Gamma(k-\kappa)}{\Gamma(k+1)\Gamma(n-\kappa+1)}k = \frac{\Gamma(n+1)}{\Gamma(n+1-\kappa)}\sum_{k=1}^{i-1}\frac{\Gamma(k-\kappa)}{\Gamma(k)}\\ &\overset{1.}{=} \frac{\Gamma(n+1)\Gamma(i-\kappa)}{(1-\kappa)\Gamma(n+1-\kappa)\Gamma(i-1)}\\ \Rightarrow \sum_{j=1}^{i-1}\frac{\lambda_j}{\lambda_i} &\overset{2.}{=} \frac{i-2}{2-\kappa}.\end{aligned}$$

After rewriting the fraction of λ_k's in terms of Beta-functions, the following approximation is needed:

$$\frac{\lambda_j}{\lambda_i}\frac{1}{n(n-1)} = \frac{B(n-1,1-\kappa)}{n(n-1)B(j-1,1-\kappa)} \overset{3.}{\approx} \left(\frac{j-1}{n-1}\right)^{1-\kappa}\frac{1}{n(n-1)}.$$

Using this on (4.2) yields:

$$\begin{aligned}
\ell_s = & \sum_{j=1}^{s-1}\left[\frac{1}{j} - \frac{\lambda_j}{\lambda_s}\frac{1}{s-1} - \sum_{n=j+1}^{s-1}\frac{\lambda_j}{\lambda_n}\frac{1}{n(n-1)}\right] \\
= & H_{s-1} - \frac{s-2}{2-\kappa}\frac{1}{s-1} - \sum_{j=1}^{s-1}\sum_{n=j+1}^{s-1}\frac{\lambda_j}{\lambda_n}\frac{1}{n(n-1)} \\
\overset{4.}{=} & H_{s-1} - \frac{s-2}{2-\kappa}\frac{1}{s-1} - \sum_{j=2}^{s-2}\frac{H_{j-1,-\kappa}}{j^\kappa j(j+1)} \\
\overset{5.}{=} & \frac{1-\kappa}{2-\kappa}H_{s-1} - \frac{1}{2-\kappa}\left(\frac{s-2}{s-1} - \frac{3}{2}\right) + \frac{1}{4} \\
\overset{s\gg 1}{\approx} & \frac{1-\kappa}{2-\kappa}\log(s) + \text{const.}
\end{aligned}$$

□

Note, that the underlying Markov process of the NAW in a HoC-model is similar to other processes, notably the dynamics in a *quasispecies model* and the one dimensional *Jepsen gas*. In the quasispecies model, the sequence space is distributed with a certain population distribution. Selection will alter it until the fittest sequence is also the most populated one. Until then, the most populated sequence is not necessarily very fit. The number of times, the most populated sequence changes until the fittest one is most populated, behaves exactly as the number of adaptive steps of a NAW [101, 102]. The one dimensional Jepsen gas on the other hand is a gas in which particles can move freely without interactions with a velocity, which is distributed at random at the beginning of the process. The number of times, the leading particle is overtaken behaves also like the number of adaptive steps of a NAW [103]. For the Jepsen gas the variance of the number of overtaking processes $\mathrm{Var}(\ell_s)$, where s is the starting position of the fastest particle, is known to be

$$\mathrm{Var}(\ell_s) = \frac{(\kappa-1)\,(2+\kappa(\kappa-2))}{(\kappa-2)^3}\log(s) + \text{const.}[\kappa] \text{ for } \kappa \leq 1. \tag{4.3}$$

Recently, also results for the NAW beyond the Gillespie approximation became available. Starting with an approach similar to the one used by Flyvbjerg and Lautrup [59], Jain and Seetharaman [100] arrived at a similar expression as res. 4.3. Further calculations following the same approach even resulted in an expression for higher moments, with variance equal to (4.3) [104].

4.3 Adaptation in correlated fitness landscapes

As indicated in the previous sections, experimental results suggest fitness landscapes to be correlated (see sec. 2.3 and 3.5). Despite this fact, the results in this field are scarce. Most results cannot a priori be transferred from a HoC- to an RMF- or LK-model, since they depend heavily on the i.i.d. property. Starting with single adaptive steps and proceeding to adaptive walks in this section the focus is on adaptation in the SSWM-limit on correlated fitness landscapes of the RMF type.

4.3.1 Single adaptive steps

Before studying adaptive walks, single steps of adaptation shall be investigated. For the HoC-model this has been done by Orr [98] and Joyce et al. [105] who arrived at results for all three probability classes in terms of the GPD with $\kappa < \frac{1}{2}$ (distributions with existing second moment). Starting from the transition probability of NAWs (1.9), the question was how the rank changes on average in each step, if the next sequence is chosen under selective pressure. Given, that the populated sequence has rank i in the old neighborhood before the step and thus rank j after the step is taken, in the new neighborhood expected value and variance of j are given by [105]

$$\begin{aligned} \langle j \rangle &= 1 + \frac{i-2}{2}\left(\frac{1-\kappa}{2-\kappa}\right), \\ \mathrm{Var}(j) &= \frac{(1-\kappa)(i-2)[(\kappa^2-4\kappa+7)i+6(1-\kappa)]}{12(3-\kappa)(2-\kappa)^2}. \end{aligned} \tag{4.4}$$

To be precise, only the calculation of the variance needs $\kappa < \frac{1}{2}$, while the expectation of j can be calculated up to $\kappa < 1$, until the first moment of the GPD diverges. The biological meaning and significance, as well as the validity of (1.9) (in the derivation s is assumed to be relatively small) is heavily discussed, see e.g. [9]. The equations (4.4) reduce to the previously known [98] results in the Gumbel class for $\kappa \to 0$:

$$\langle j \rangle = \frac{i+2}{4}, \qquad \mathrm{Var}(j) = \frac{(i-2)(7i+6)}{144}. \tag{4.5}$$

Since these results are based on the transition probability (1.9) which does only depend on the spacings Δ_k, their validity in the RMF model should depend on the form of the spacings. For an exponentially distributed RMF-landscape as in def. 1.8, an expression was derived for the mean kth largest

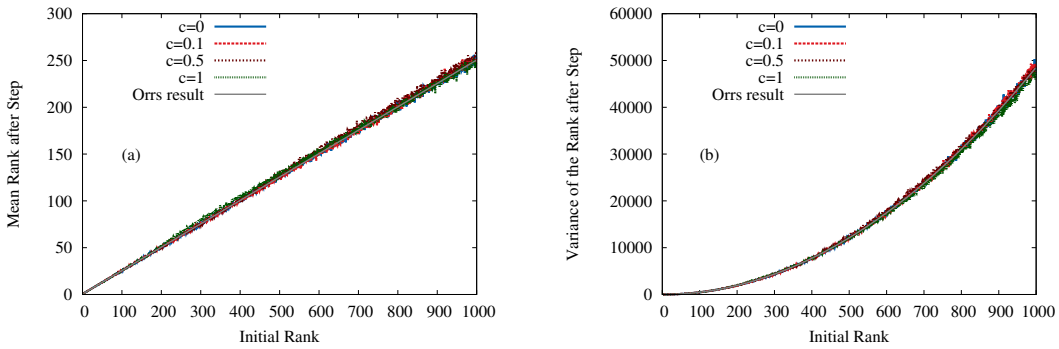


Figure 4.2: (a) Mean and (b) variance of the fitness rank after an adaptive step in the exponentially distributed RMF-landscape is shown as a function of the initial rank. Simulation results for different values of c are compared to the analytical expressions (4.5) for the HoC- landscape ($c = 0$). Here $L = 1000$ and $d = 50$.

in the neighborhood, μ_k, see (2.14). Obviously, in this expression of μ_k the parts which depend on the rank k and those which depend on the landscape properties c and d are connected with a '+'. This means, that the spacings are independent of the landscape properties. The remaining expressions are equal to those from an exponentially distributed HoC-landscapes. Thus, the results from the Gumbel class (4.5) should be the same for an exponentially distributed RMF-model. Of course, this is only valid in the regime where (2.14) is valid, which is reasonably small c. Simulations were performed for various values of c and the results are presented in fig. 4.2 with a comparison to (4.5).

To check the behavior of adaptive steps on non-exponentially distributed RMF landscapes simulations were performed as well for GPD landscapes and the results are presented in fig. 4.3 with a comparison to (4.4). In both cases, exponential distribution and GPD, the analytical results from the HoC-landscape fit the simulations data very well, although for larger $|\kappa|$ deviations increase. Nevertheless, it is quite astounding, that the statistics of single adaptive steps are hardly influenced by the introduction of a fitness gradient c.

4.3.2 Adaptive walks: Numerical results

Because results of the form of res. 4.3 are not at hand, simulations were performed to investigate the behavior of NAWs on GPD RMF-landscapes, according to def. 1.8. For these simulations the sequence length $L = 2000$ was chosen, while d, c and the starting rank r were varied. Results for an exponentially distributed RMF-model are shown in fig. 4.4. For small c the

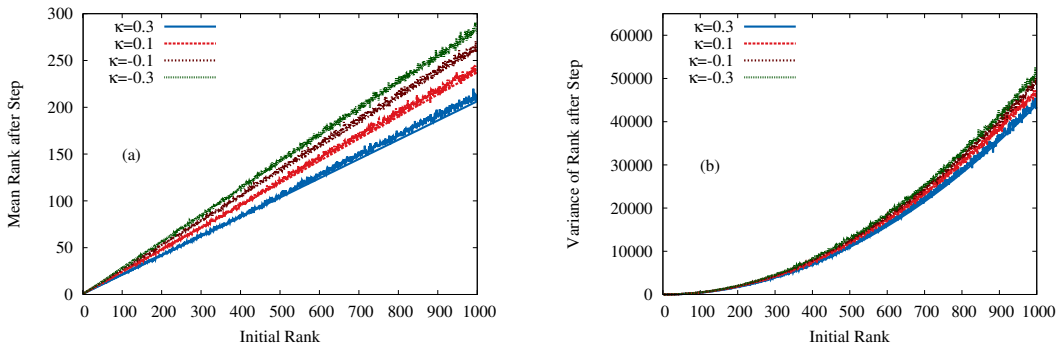


Figure 4.3: (a) Mean and (b) variance of the fitness rank after an adaptive step in the GPD RMF-landscape is shown as a function of the initial rank. Simulation results for different values of κ and $c = 0.5$ are compared to the analytical expressions (4.4) for the HoC- landscape ($c = 0$). Here $L = 1000$ and $d = 50$.

walk length seems to be independent of d and shows logarithmic behavior in the starting rank and can be fitted by res. 4.3. If c gets a bit larger, fitting is not possible anymore, but the walk length seems to be still logarithmic in r. A d-dependence emerges for the values $c > 0.3$. The walk length becomes linear in d for $c = 1$, where it also becomes independent of the starting rank. Simulations were also performed for a GPD RMF-model with κ different than zero. The results are presented in fig. 4.5. For $c = 0.5$ the d-dependence seems to be linear for all analyzed κ values. The slope of the fitting functions depends on κ and increases with negative κ, as the support of the distribution decreases, which leads to a stronger influence of c. The dependence on the starting rank is described by res. 4.3, although the deviations from this expression increase also for negative κ. From these results, the form of an analytical expression comparable to res. 4.3 can be conjectured, which would lead to the obtained results:

$$\ell(r, c, d, \kappa) = \frac{1-\kappa}{2-\kappa} \log(r) + \alpha(c, \kappa)d + \beta(c, \kappa) \tag{4.6}$$

with so far unknown, nonlinear functions α, β with $\alpha(0, \kappa) = 0$ and $\beta(0, \kappa) > 0$.

Closely connected to the adaptive walk length, especially considering the very long walks for large d and c values, is the question of the *crossing probability* $P_{\mathrm{Cr}}(c)$. $P_{\mathrm{Cr}}(c)$ is the probability, that a walk which starts on the antipodal of σ^*, i.e. $d = L$, is completed on σ^* and thus *crosses* the complete sequence space. Such a crossing depends crucially on the existence of adaptively accessible paths in the sequence space. A path of length ℓ is a sequence of elements from the hypercube $\mathfrak{P} = (\sigma_1, \ldots, \sigma_\ell)$. It is called

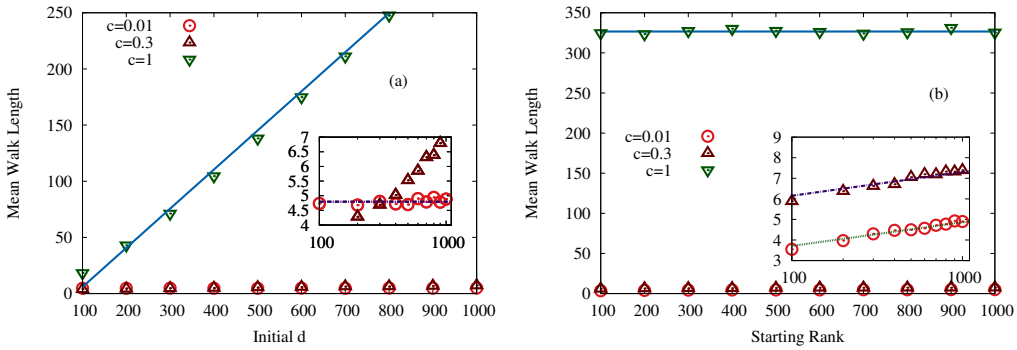


Figure 4.4: Mean length of adaptive walks in Gauß-distributed RMF-landscapes. (a) Mean walk length for randomly chosen starting rank versus initial Hamming distance d to the reference sequence. Straight line illustrates the linear dependence of the walk length on d for large c. (b) Mean walk length for constant initial Hamming distance $d = 1000$ versus starting rank r. The horizontal line connecting the data points for $c = 1$ illustrates that walk length becomes independent of starting rank for large c. Insets show the data for small c on logarithmic scales for d and r, respectively. Horizontal line in the inset in panel (a) illustrates that the walk length is independent of initial distance d for $c = 0.01$, but acquires such a dependence with increasing c. Straight lines in the inset of panel (b) illustrate the logarithmic dependence of the walk length on initial rank for small c. The number of loci is $L = 2000$.

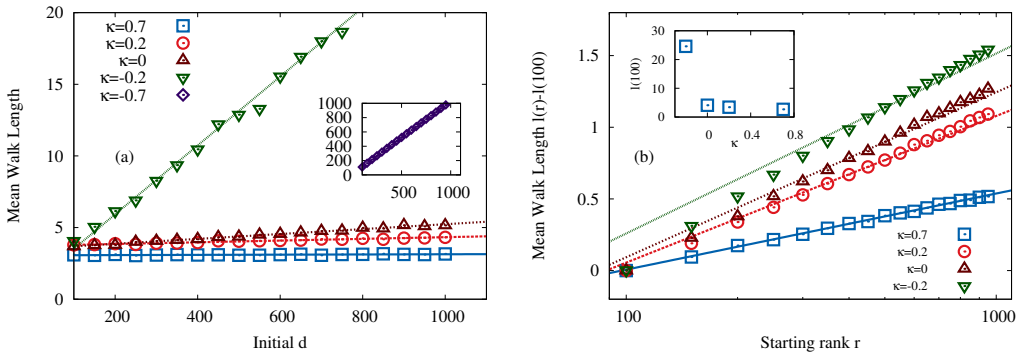


Figure 4.5: Mean length of adaptive walks in GPD RMF-landscapes. In both figures $c = 0.5$. (a) Walks with randomly chosen starting rank and varying initial Hamming distance d to the reference sequence. Inset shows results for $\kappa = -0.7$. (b) Walks starting at constant Hamming distance $d = 1000$ and varying starting rank r. In the main panel the walk length for $r = 100$ has been subtracted for clarity, and the corresponding values of ℓ_{100} are shown in the inset. Simulations were carried out for various choices of the EVT index κ. The lines in (a) correspond to fits assuming a linear d-dependence, lines in (b) show the HoC result (res. 4.3). The number of loci is $L = 2000$.

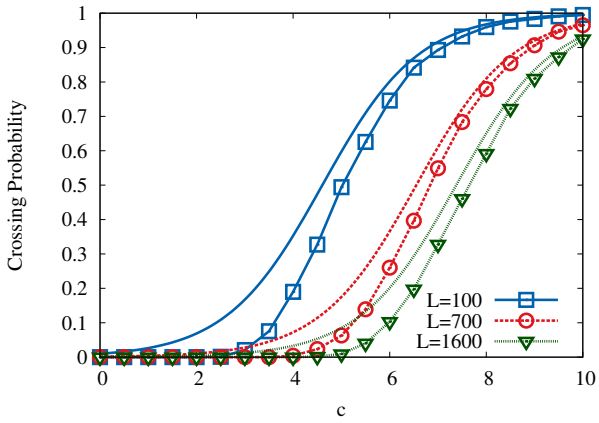


Figure 4.6: The probability for an adaptive walk starting from the antipodal sequence $\overline{\sigma^*}$ to reach the reference sequence σ^* and terminate there. Results are shown for Gumbel-distributed random fitness components. Numerical results are displayed by symbols connected with lines, while the corresponding lines without symbols show the upper bound given in Eq.(4.7).

accessible, if $\forall_{i,j\in[1,\ell]}\forall_{\sigma_i,\sigma_j\in\mathfrak{P}} : i < j \Rightarrow F(\sigma_i) < F(\sigma_j)$. For a crossing event an accessible path has to exist from $\overline{\sigma^*}$ to σ^*. The existence of such paths has been of recent interest [26, 24, 63]. For a special case of the RMF-model in which the global maximum is constrained to be at σ^*, Hegarty and Martinsson [106] found that the probability, that at least one accessible path from $\overline{\sigma^*}$ to σ exists converges to one for $L \to \infty$ and $c > 0$. On a general RMF-model, additionally to the existence of such a path, σ^* has to be a maximum, which it is with probability

$$p_c^{\max}(0) = \frac{1}{1 + Le^{-c}} \tag{4.7}$$

for a Gumbel distributed RMF-landscape (see res. 2.1). Since $p_c^{\max}(0) \to 0$ ($L \to \infty$), it is unclear, how much insight the results from Hegarty and Martinsson [106] give to the problem in this setting. Nevertheless, $p_c^{\max}(0)$ works as an upper limit of P_{Cr} as can be seen in fig. 4.6 where it is compared to results obtained by simulations.

4.3.3 Greedy walks and correlations

The previous chapters dealt with adaptation with the 'natural' transition probability (1.9). Now the focus is on greedy adaptive dynamics an an RMF-model from def. 1.8, where the step is always taken to the fittest sequence of

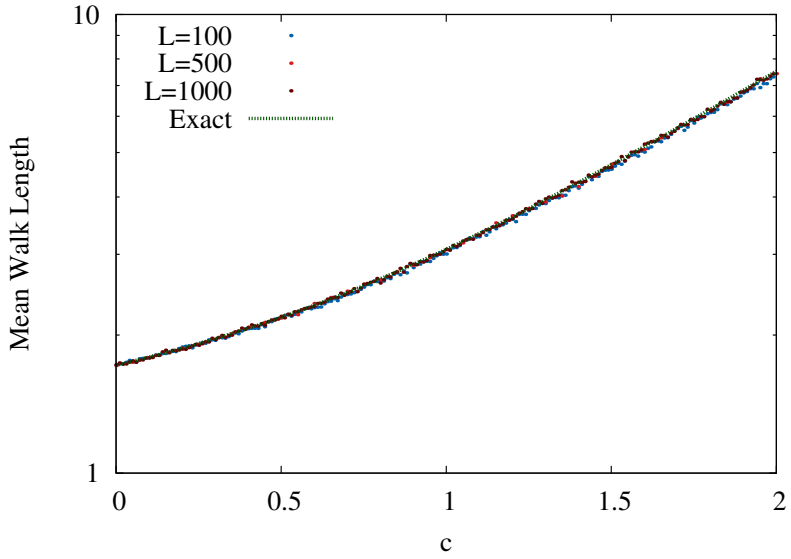


Figure 4.7: The figure compares the analytic results from the Gumbel distributed dRMF-model in the $L \to \infty$ limit (res. 4.5) with results retrieved from simulations on a Gumbel distributed RMF-landscape with finite L. Obviously for the c values analyzed here, the analytic expression describes the data very well.

the neighborhood. To get analytical results, the sequence length is assumed to be so large, that the change of the number of neighbors from uphill and downhill are negligible on the (relative to L) short adaptive walks, which start at $\overline{\sigma^*}$. This motivates the introduction of the *diminishing RMF-model* (dRMF-model) [107]: based on an RMF-model, only the uphill neighborhood is seen from each sequence, which diminishes each step. For later convenience, the q-analogue of a function is defined:

Definition 4.1. The *q-analogue* [108] of a known function is a generalization with parameter q, which converges to the known function for $q \to 1$. Starting with the q-number (or q-bracket) $[n]_q = \frac{1-q^n}{1-q} \to n\,(q \to 1)$, various q-functions can be defined straightforwardly:

- q-Pochhammer symbol $(a;q)_n = \prod_{k=0}^{n-1}(1-aq^k)$
- q-factorial $[n]_q! = \prod_{k=1}^{n}[k]_q = \frac{(q;q)_n}{1-q}$
- q-Gamma function $\Gamma_q(n+1) = [n]_q!$ for $n \in \mathbb{N}$
- q-exponential $e_q(z) = \sum_{n=0}^{\infty} \frac{z^n}{[n]_q!} = \sum_{n=0}^{\infty} \frac{z^n(1-q)^n}{(q;q)_n}$

- and many more.

Result 4.4 (Probability for l steps in the Gumbel case)**.** *The probability, that a GAW takes at least ℓ steps on a Gumbel distributed dRMF-landscape in the limit $L \to \infty$ is given by*

$$Q_\ell = \left([\ell]_{e^{-c}}!\right)^{-1}.$$

Proof. An approach similar to the one used in the HoC landscape for greedy walks by Orr [97] is used here. The idea is to develop the expression for the lth step as the product of the probability to have taken $l-1$ steps (Q_{l-1}) times the probability to make the lth step, which equals the probability that the largest value of all visited neighborhoods is among the $L-l$ fitness values of the lth neighborhood.

$$\begin{aligned}
Q_1 =& \mathbb{P}(\text{don't start on a maximum}) = \int_0^1 \left(1 - P(x-c)^L\right) \mathrm{d}P(x) \\
=& \int_0^1 \left(1 - P(x)^{e^c L}\right) \mathrm{d}P(x) \ = 1 - \frac{1}{L + e^c} \\
Q_2 =& \mathbb{P}(\text{one step is made}) \times \\
& \mathbb{P}(\text{largest from the neighborhoods with benefit 0 and c is in this neighborhood}) \\
=& \int_0^1 Q_1 (L-1) P(x+2c) P(x+c)^L \mathrm{d}P(x) \ = Q_1 \frac{L-1}{e^{-2c} + L e^{-c}} \\
Q_\ell =& \mathbb{P}(\ell \text{ steps are taken}) \times \\
& \mathbb{P}(\text{largest from the } \ell \text{ visited neighborhoods is among these } L - \ell \text{ values}) \\
=& Q_{\ell-1}(L-\ell) \int_0^1 P(x+\ell c) \left(\prod_{n=1}^{\ell-1} P(x + (\ell - n)c)^{L-n} \right) \mathrm{d}P(x) \\
=& Q_{\ell-1}(L-\ell) \int_0^1 P(x)^{-\ell c} \left(\prod_{n=1}^{\ell-1} P(x)^{(L-n)e^{-(\ell-nc)}} \right) \mathrm{d}P(x) \\
=& Q_{\ell-1} \frac{L-\ell}{1 + e^{-\ell c} + \sum_{n=1}^{\ell-1} (L-n) e^{-(\ell-n)c}} \\
\to& Q_{\ell-1} \frac{1}{\sum_{n=1}^{\ell-1} e^{-nc}} \ \ (L \to \infty) \\
=& Q_{\ell-1} \frac{1 - e^{-c}}{1 - e^{-c\ell}} = \frac{Q_{\ell-1}}{[\ell]_{e^{-c}}} = \frac{1}{[\ell]_{e^{-c}}!}.
\end{aligned}$$

□

Result 4.5 (GAW length on a Gumbel RMF)**.** *On a Gumbel distributed dRMF-landscape, the mean GAW length in the limit $L \to \infty$ is given by*

$$\ell = e_{e^{-c}}(1) - 1.$$

Proof. If Q_l is the probability, that at least l steps are taken, then the probability, that *exactly* l steps are taken, P_l is given by the probability to take at least l steps, minus the probability to go $l+1$ steps, thus:

$$P_l = Q_l - Q_{l+1}.$$

The mean walk length is then the expectation of l:

$$\begin{aligned}
\ell = \langle l \rangle &= \sum_{l>0} l P_l = \sum_{l>0} l\,(Q_l - Q_{l+1}) = \sum_{l>0} lQ_l - (l+1)Q_{l+1} + Q_{l+1} \\
&= \sum_{l>0} lQ_l - \sum_{l>1} lQ_l + \sum_{l>1} lQ_l = Q_1 + \sum_{l>1} lQ_l - \sum_{l>1} lQ_l + \sum_{l>1} lQ_l \\
&= \sum_{l\geq 0} lQ_l - Q_0 = \sum_{l\geq 0} \frac{1}{[l]_{e^{-c}}!} - \frac{1}{[0]_{e^{-c}}!} = e_{e^{-c}}(1) - 1.
\end{aligned}$$

□

Note, that for $c \to 0$ the HoC result $\ell = e - 1$ [97] is reached as the q-exponential converges to the exponential function. A comparison between this result and data from simulations on a finite hypercube with RMF-landscape can be found in fig. 4.7. It shows clearly, that the simplifications made to arrive at the analytic results do not compromise the validity of res. 4.5 as it describes the data very nicely for all tested L and c values.

For the uniform distribution, another interesting effect has been found. The largest value in the uphill neighborhood is approximated by $\int_{F_{\max}}^{1} \mathrm{d}P(x) = \frac{1}{L} \Rightarrow F_{\max} \approx 1 - \frac{1}{L}$, where the diminishing by the walk length is neglected. This leads to the idea, that a shift of $c = \frac{1}{L}$ will always ensure, that seen from the momentary sequence, the uphill neighborhood has the largest value $F_{\max} + c = 1 - \frac{1}{L} + \frac{1}{L} = 1$ which is the largest possible value in a uniform distribution. It is thus plausible, that the greedy walk will be of close to maximal length if $c \geq \frac{1}{L}$ and rather short ($\mathcal{O}(1)$) for $c < \frac{1}{L}$ which can be verified by simulations, see fig. 4.8.

If the walk does not start at the antipodal of the reference sequence but somewhere closer, the approximation by the dRMF-model is not justified, because the sheer number of downhill neighbors increases the possibility of a downhill step to a non-negligible level. Here, it is convenient to introduce $\alpha = \frac{d(\sigma^0,\sigma^*)}{L}$, for a starting sequence σ^0, which is kept constant while $L \to \infty$.

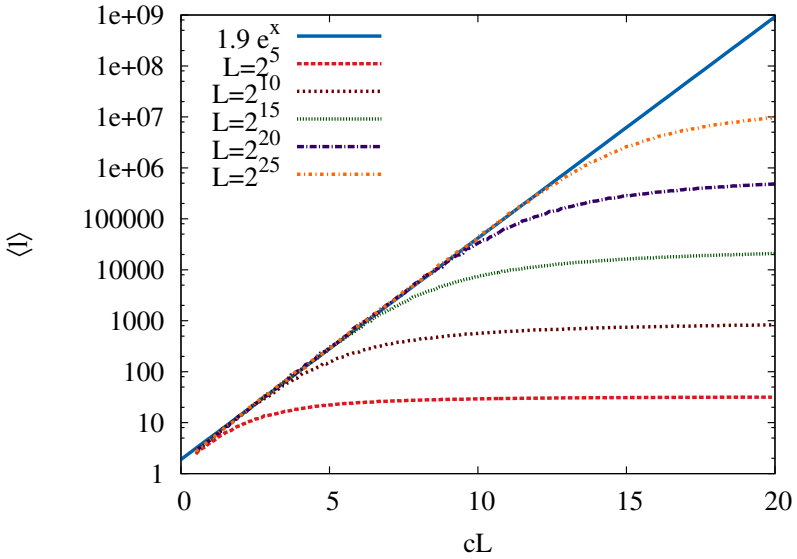


Figure 4.8: This figure shows, how the plots of simulation data for different values of L collapse, if the mean walk length of a GAW on a uniformly distributed dRMF is plotted versus Lc, which indicates a transition from short to long walks at $c \sim \frac{1}{L}$. The curves flatten if $\ell \sim L$. The $1.9e^{cL}$ behavior for small walk lengths compared to L is a numeric result by Su-Chan Park.

Result 4.6 (GAW length for $\alpha < 1$)**.** *In a Gumbel distributed RMF-model, the GAW length with respect to α and c is approximately given by*

$$\ell \approx e\left(1 + \frac{2\alpha - 1}{4}c + \frac{123 + 596\alpha(1-\alpha)}{864}c^2\right) - 1.$$

Proof. A detailed derivation[1] of this expression is given in Park et al. [107], here only a sketch is presented, since the detailed calculations are not very illustrative. To analyze the walk length, it is necessary to keep track over the steps taken up and down. Therefore, $d_i \in \{\pm 1\}$ is introduced as the change in the Hamming distance to the reference sequence at the ith step, i.e. if the sequence visited at the ith step is σ^i, then $d(\sigma^i, \sigma^*) - d(\sigma^{i-1}, \sigma^*) = d_i$. The complete walk up to step l in terms of distance changes can be written as a sequence $\mathfrak{d}_l = (d_1, \ldots, d_l)$. The total change in distance after the lth step is given by $M_l = \sum_{i=1}^{l} d_i$. If $J(\mathfrak{d}_l)$ is the probability, that the path $\mathfrak{d}_l$ has been chosen, after the walk started at σ_0, the probability, that at least l steps are taken in an adaptive walk is

$$Q_l = \sum_{\text{all } \mathfrak{d}_l} J(\mathfrak{d}_l).$$

After some algebra and the assumption, that the walks are so small, that the number of uphill and downhill neighbors is essentially constant, this expression arrives at

$$Q_l = \sum_{\text{all } \mathfrak{d}_l} \prod_{k=1}^{l} \frac{s_{d_k}}{1 + \sum_{m=1}^{k-1} e^{-cM_m}}, \tag{4.8}$$

where $s_{+1} = \frac{\alpha e^c}{\alpha e^c + (1-\alpha)e^{-c}}$ and $s_{-1} = 1 - s_{+1}$. For $c \ll 1$, Q_l can be expanded to the second order of c which leads to the mean walk length. Note, that this result retrieves the HoC result $e - 1$ for $c \to 0$. □

Note, that the behavior of ℓ is non-monotonic and has a minimum in c for small α which was first observed in numerical studies in [83]. The minimum becomes more pronounced for decreasing α, i.e. for walks starting closer at σ^*. If α is small enough, the expected walk length should be close to the minimal value 1, which implies, that it can be well approximated by

$$\ell - 1 \approx Q_2 = \left(\frac{s_{+1}}{1 + e^{-c}} + \frac{s_{-1}}{1 + e^{c}}\right). \tag{4.9}$$

Note, that $\ell > 1$ since the probability to start on a maximum (i.e. $l < 1$) vanishes for $c > 0$ and $\alpha > 0$ as $\frac{1}{\alpha L}$. See fig. 4.9 for an illustration of the above expression with a comparison to simulation data.

[1] Most of the calculations were done by Su-Chan Park

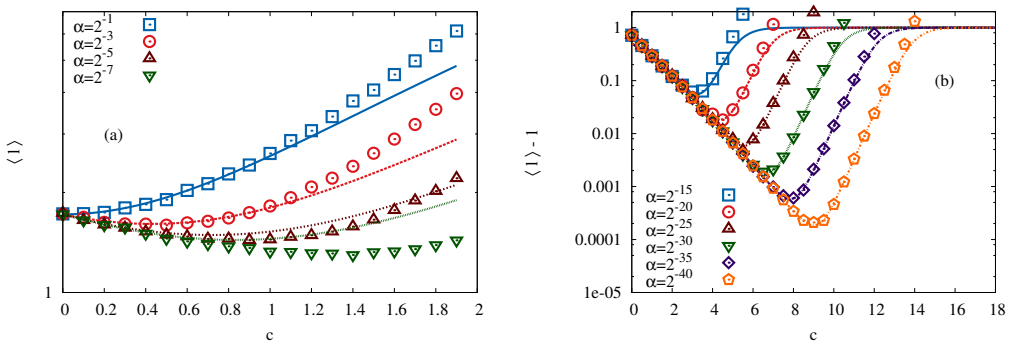


Figure 4.9: Mean length of a GAW in a Gumbel distributed RMF-landscape. (a) Illustration of res. 4.6. (b) Illustration of the approximation of the mean walk length (dots are simulation results for $\ell - 1$) by Q_2 (lines, see (4.9)). The simulations were done with $L = 2^{60}$.

4.3.4 Phase transition in the random adaptive walk

Now the focus shall be on 'random' adaptation where the next sequence is chosen at random from the beneficial neighbors and the transition probability is

$$P_{\sigma_i \to \sigma_j} = \frac{\Theta(F(\sigma_j) - F(\sigma_i))A_{ij}}{\sum_k \Theta(F(\sigma_k) - F(\sigma_i))A_{ik}}.$$

On HoC-landscapes, this process was analyzed by Flyvbjerg and Lautrup [59]. On the RMF-landscape, it is convenient here to use def. 1.9 of the RMF-model, where the average global minimum is located at the reference sequence which is also the starting sequence of the walk. It is also helpful to consider only walks in the uphill direction, away from σ^*. This way, the mean fitness after l steps is $cl\langle\xi\rangle$.

Result 4.7 (Phase transition on the RMF-model)**.** *In an exponentially distributed RMF-model with $\lambda = 1$, a phase transition exists for the mean adaptive walk length:*

$$\ell \sim \begin{cases} \frac{\log(L)}{1-c}, & c < 1 \\ \log(L)^2, & c = 1 \\ \mathcal{O}(L), & c > 1. \end{cases}$$

Proof. While a detailed derivation is given in [109] and the supporting material[2], here only a sketch shall be presented, since the details are rather long and not very illustrative. Analogously to Flyvbjerg and Lautrup [59],

[2]Most of the calculations were done by Su-Chan Park

the probability, to go at least l steps and reach fitness $y + lc$ satisfies the recursion relation

$$q_l(y) = \int_0^\infty \mathrm{d}y'\; q_{l-1}(y') \frac{1 - P(y' - c)^{L-l}}{1 - P(y' - c)} p(y).$$

The integrand can be understood as the probability that $l-1$ steps are taken and fitness y' is reached (i.e. $q_{l-1}(y')$), times the probability, that the next fitness value is y, which has to be larger than $y' - c$ (i.e. $\frac{p(y)}{1-P(y'-c)}$) times the probability, that there is one value larger than $y' - c$ among the $L - l$ neighbors (i.e. $1 - P(y' - c)^{L-l}$). The probability to walk at least l steps is then given by

$$Q_l = \int_0^\infty \mathrm{d}y\; q_l(y).$$

As in the GAW case, the probability to go exactly l steps is given by $D_l = Q_l - Q_{l+1}$. The mean walk length is $\ell = \sum_{l \geq 0} l D_l$. To evaluate these expressions it is helpful to have a look at the $L \to \infty$ limit of the recursion relation of q_l, which simplifies due to the fact, that $P(x) < 1$ and thus $\forall_x \forall_l : P(x)^{L-l} \to 0$ $(L \to \infty)$:

$$q_l^\infty(y) = \int_0^\infty \mathrm{d}y'\; \frac{q_{l-1}^\infty(y')}{1 - P(y' - c)} p(y). \tag{4.10}$$

If an approximation of $q_l(y)$ by $q_l^\infty(y)$ is justified or not, is decided by the absolute value of $q_l(y) P(x-c)^{L-l}$ which has to stay much smaller than unity for a good approximation. To find a regime, where this is fulfilled, the mean fitness after l steps with respect to $q_l^\infty(y)$ is defined as

$$z_l = \int_0^\infty y q_l^\infty(y) \mathrm{d}y\ .$$

To estimate values of l for which the infinite L approximation is valid, the maximum of z_l with $q_l(y) P(x - c)^{L-l} \ll 1$ is calculated and the equation $P(z_l - c)^{L-l} = \frac{1}{e}$ is solved for $q_l^\infty(y)$ which arrives after some lengthy calculations at the relations

$$q_l^\infty(y) = -\frac{\mathrm{d}}{\mathrm{d}y} \sum_{n=0}^{l} y \frac{(y + cn)^{n-1}}{n!} e^{-y-cn}$$

$$\Rightarrow z_l = 1 + \sum_{k=1}^{l} \alpha_l,$$

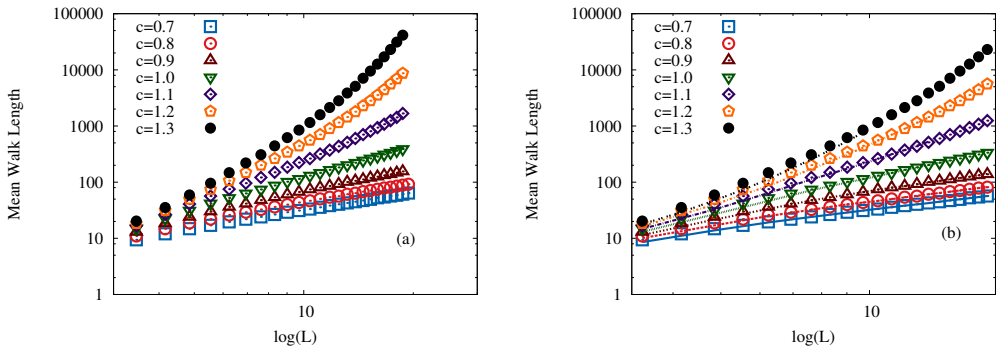


Figure 4.10: Mean length of adaptive walks in an exponentially distributed RMF-landscape. (a) Mean walk length in a dRMF-landscape. The phase transition can be seen nicely at $c = 1$. For smaller c the behavior becomes linear while for larger c faster increase is visible. Note the unusual double logarithmic plot with additional $\log(L)$ on the x-axis. (b) shows the mean walk length in an RMF-landscape where back-steps are allowed (lines) compared to a dRMF-landscape (points). The plot shows, that both data sets are very close and that the relative deviation gets smaller for larger L.

where $\alpha_l = z_l - z_{l+1} = \frac{(cl)^{l+1}e^{-cl}}{l!}\int_0^\infty \mathrm{d}t\ te^{-ct}e^{(l-1)(\log(1+t)-ct)}$. Using Laplace' method of integral approximation and Stirling's formula, in the regime of $l \gg 1$, α_l can be simplified to take the form

$$\alpha_l = \begin{cases} \frac{l^l e^{-l}}{l!} \sim \frac{1}{\sqrt{2\pi l}}, & c = 1 \\ \max(1-c,0)\frac{e^{-l(c-1-\log(c))}}{\sqrt{2\pi l}}\frac{c}{(c-1)^2 l}, & c \neq 1 \end{cases}$$

$$\Rightarrow z_l = 1 + \sum_{k=1}^{l}\alpha_k \sim \begin{cases} (1-c)l, & c < 1 \\ \sqrt{\frac{2l}{\pi}}, & c = 1 \\ \text{finite}, & c > 1. \end{cases}$$

The standard deviation s_l of Q_l has been calculated and behaves as follows:

$$s_l \sim \begin{cases} \mathcal{O}(\sqrt{l}), & c \leq 1 \\ \mathcal{O}(1), & c > 1. \end{cases}$$

This means Q_l will be sharply peaked for large L and can thus be approximated as $\delta(y - z_l)$ if $c > 1$. If $c \leq 1$, it can be assumed, that D_l gives a significant contribution when $\log(F(z_l + s_l)) \sim -\log(L)$. □

The effect from res. 4.7 can be seen in fig. 4.10. Obviously the differences between dRMF and RMF-landscapes concerning the adaptive walk distance

are negligible on a qualitative level and surprisingly small on a quantitative level.

Following Park et al. [109], it can be argued, that the phase transition will only occur for RMF-landscapes, which are distributed by a distribution with exponential tail. Starting at (4.10), this equation can be modulated into

$$z_{l+1} - z_l = \int_{-\infty}^{\infty} \frac{q_{l+1}^{\infty}(y)}{h(y)} \mathrm{d}y \; - c \tag{4.11}$$

with $h(y)$ defined in (1.11), for distributions with unbounded support in both directions. As mentioned in sec. 1.6, h has only a finite limit for distributions with exponential tail. Based on (4.11), it will be assumed, that z_l diverges and checked, whether this is consistent, using the asymptotics of the hazard function.

For the exponential distribution, that is $h(y) \to \frac{1}{\lambda}$ $(y \to \infty)$, which leads to

$$z_{l+1} - z_l \approx \lambda - c \Rightarrow z_l \approx (\lambda - c)l.$$

This means linear growth if $c < \lambda$ and inconsistencies with a large positive z_l for $c > \lambda$, which is a reproduction of the properties stated in the proof of res. 4.7.

For a distribution which has a tail of the form $\log(p(y)) \sim -y^{\alpha}$, the hazard function behaves like $h(y) \sim y^{\alpha-1}$, which means, that z_l only diverges for $\alpha < 1$ as $z_l \sim l^{\frac{1}{\alpha}}$. This implies that the mean walk length is logarithmic in L. If $\alpha > 1$ and thus the tail thinner as exponential, the integral in (4.11) stays small and z_l is dominated by c, if $c > 0$. This yields a walk length linear in L for $c > 0$ and logarithmic in L for $c = 0$ due to the then occurring $z_l \sim l^{\frac{1}{\alpha}}$.

For a distribution with power law tail $p(y) \sim y^{-(\nu+1)}$ the hazard function behaves as $h(y) \sim \frac{1}{y}$. This leads to an exponentially growing z_l, implying logarithmic walk length in L.

Chapter 5

Recombination

Apart from the biological importance of an understanding of the benefits of recombination the fact, that it leads to quadratic terms of p in the unnormalized evolutionary equations makes it a very interesting but complex process. It is not easy to find exact results for sequence lengths $L > 2$. Nevertheless, if the focus is not on the complete evolutionary solution, but only on parts, exact analytical expressions can be found. This approach will be used in the first part of this chapter. For more detailed studies, numerical simulations can be performed to give a more general understanding of the mechanisms contributing to the evolution with and without recombination. This will be done in the second part of this chapter.

5.1 Exploration of the sequence space

In the limit $N \to \infty$ the evolutionary equations become deterministic. For every starting condition, there is exactly one solution. Since the size of the population is infinite, once a sequence became populated, it will never go extinct, even if it has a very low fitness. This is due to the fact, that the Wrightian fitness is a positive number. If fitness is set aside for now, the evolution of a population on the hypercube reduces to the question whether a sequence is populated or not.

One of the advantages recombination holds is the supposedly faster exploration of the sequence space. For infinite populations, with single mutations the complete L dimensional hypercube will be explored by a mutation only process in L steps. In the following it will be shown, that recombination leads to an exponential exploration of the sequence space and is thus vastly faster.

5.1.1 Properties of Hamming balls

In the following, the infinite population limit, $N \to \infty$, with single mutations only will be analyzed. When a population starts monomorphic, mutations will occur equally distributed in all directions which leads to an isotropic exploration of the sequence space. It is thus worth to analyze the properties of *Hamming balls*

$$B_\sigma(r) = \{\sigma' \in \mathbb{H}^L | d(\sigma, \sigma') \leq r\} \tag{5.1}$$

which are the natural, discrete equivalent to a sphere. The maximal radius of a Hamming ball in $\mathbb{H}^L$ is thus L, and $\forall_{\sigma \in \mathbb{H}^L} : B_\sigma(L) = \mathbb{H}^L$. The volume of B is given by $|B_r(\sigma)| = \sum_{d=0}^{r} \binom{L}{d}$.

Result 5.1 (Intersection of Hamming balls)**.** *For two Hamming balls with radius r around sequences σ, σ' with $d(\sigma, \sigma') = d \leq 2r$, the number of sequences in the union is given by*

$$|B_\sigma(r) \cup B_{\sigma'}(r)| = 2 \sum_{i=0}^{r} \binom{L}{d} - \sum_{n=0}^{[r-\frac{d}{2}]} \binom{L-d}{n} \sum_{m=d-(r-n)}^{r-n} \binom{d}{m},$$

where the Gauß brackets $[x]$ denote the closest integer number below x.

Proof. W.l.o.g., choose $\sigma = (0, 0, \ldots, 0)$ and $\sigma' = (1, \ldots, 1, 0, \ldots, 0)$ with $d(\sigma, \sigma') = d$. Then, define

$$\Xi^d_{m,n} = \{\nu \in \mathbb{H}^L | m \text{ times 1 in the first } d \text{ loci and } n \text{ times 1 in the last } L - d \text{ loci}\}.$$

Now choose $\tilde{\sigma} \in B_\sigma(r) \cap B_{\sigma'}(r)$ s.t. $\tilde{\sigma} \in \Xi^d_{m,n}$. Then $d(\tilde{\sigma}, \sigma) = m + n \leq r$ and $d(\tilde{\sigma}, \sigma') = d - m + n \leq r$ and thus $n \leq r - \frac{d}{2}$ and $d + n - r \leq m \leq r - n$, which leads to

$$\begin{aligned}
|B_\sigma(r) \cap B_{\sigma'}(r)| &= \sum_{n=0}^{[r-\frac{d}{2}]} \sum_{m=\max\{0,d+n-r\}}^{r-n} |\Xi^d_{m,n}| \\
&= \sum_{n=0}^{[r-\frac{d}{2}]} \binom{L-d}{n} \sum_{m=d-(r-n)}^{r-n} \binom{d}{m} \\
|B_\sigma(r) \cup B_{\sigma'}(r)| &= |B_\sigma(r)| + |B_{\sigma'}(r)| - |B_\sigma(r) \cap B_{\sigma'}(r)|
\end{aligned}$$

□

If $d(\sigma, \sigma') > 2r$, the two balls do not intersect and hence $|B_\sigma(r) \cup B_{\sigma'}(r)| = |B_\sigma(r)| + |B_{\sigma'}(r)|$.

Result 5.2 (Pairs at distance d in a Hamming ball)**.** *The number of pairs at a given distance d in a Hamming ball $B_{\sigma_0}(r) \subset \mathbb{H}^L$ is given by*

$$n(d) = \frac{1}{2}\sum_{k=0}^{r}\binom{L}{k}\sum_{n=0}^{[\frac{1}{2}(r-k+d)]}\binom{d-n}{k}\binom{n}{L-k}.$$

Proof. W.l.o.g., choose $\sigma_0 = (0,0,\ldots,0)$ and $\sigma = (1,\ldots,1,0,\ldots,0)$ with $d(\sigma_0,\sigma) = k$. Then, define

$$\zeta_{m,n}^k = \{\nu \in \mathbb{H}^L | m \text{ times } 0 \text{ in the first } k \text{ loci and } n \text{ times } 1 \text{ in the last } L-k \text{ loci}\}.$$

Now choose $\tilde{\sigma} \in B_{\sigma_0}(r)$ s.t. $\tilde{\sigma} \in \zeta_{m,n}^d$ and $d(\tilde{\sigma},\sigma) = d$. Then $d(\tilde{\sigma},\sigma) = m+n = d$ and $k-m+n \leq r$ and thus $n \leq \frac{1}{2}(r-k+d)$, hence, the number of pairs including σ with distance d is

$$n_d(k_\sigma) = \sum_{n=0}^{[\frac{1}{2}(r-k(\sigma)+d)]} |\zeta_{d-n,n}^k| = \sum_{n=0}^{[\frac{1}{2}(r-k_\sigma+d)]}\binom{d-n}{k_\sigma}\binom{n}{L-k_\sigma}.$$

When summing over all pairs, every pair is counted twice, such that for the total number of pairs in the hypercube at distance d:

$$n(d) = \frac{1}{2}\sum_{k=0}^{r}\binom{L}{k}n_d(k).$$

□

Note, that the total number of pairs in the sequence space is 2^{2L-1}.

5.1.2 Recombination on the hypercube

To analyze recombination on the hypercube it is helpful to have a look on recombination in Hamming balls. If a population starts monomorphic in a flat landscape for large population size, mutations lead to a populated Hamming ball in the first step, on which recombination can act. To begin with, the total number of parents is calculated a given sequence can have on the hypercube.

Result 5.3 (Total number of possible parent pairs in $\mathbb{H}^L$)**.** *The total number of possible parent pairs for a given sequence is $\mathcal{E} = \frac{3^L-1}{2} + 1$.*

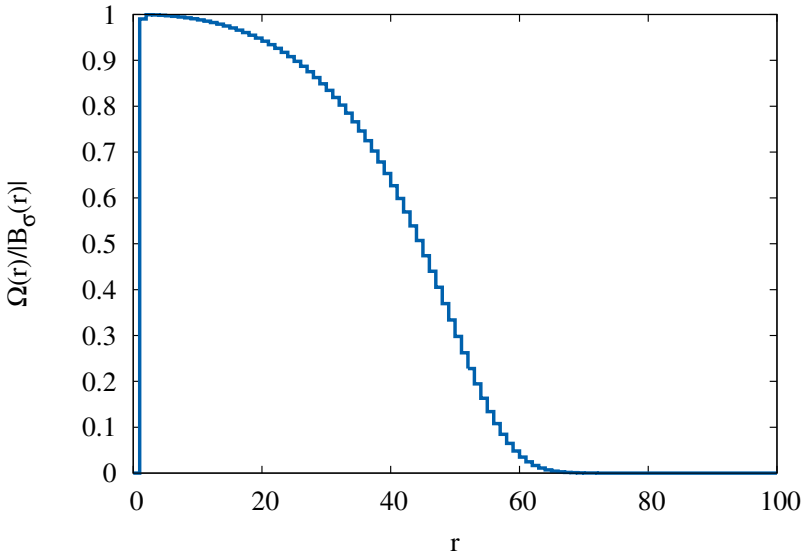


Figure 5.1: The number of pairs in the ball which can recombine out of the ball divided by the total number of pairs in the ball in dependence of the radius. L=100.

Proof. Let (σ', σ'') be a possible pair of parents to σ. This means, that if $d(\sigma, \sigma') = L - \kappa$, then $d(\sigma, \sigma'') \leq \kappa$. It has to be considered, that all pairs are counted twice but (σ, σ). Thus this pair is omitted in the sum (-1) to add it afterwards manually $(+1)$ and a factor $\frac{1}{2}$ is multiplied to the sum:

$$\begin{aligned}\mathcal{E} &= \frac{1}{2}\left(\sum_{\kappa=0}^{L}\sum_{i=0}^{L-\kappa}\binom{L}{\kappa}\binom{L-\kappa}{i} - 1\right) + 1 \\ &= \frac{1}{2}\left(\sum_{\kappa=0}^{L}\binom{L}{\kappa}2^{L-\kappa} - 1\right) + 1 = \frac{3^L - 1}{2} + 1.\end{aligned}$$

□

If a Hamming ball is populated the number of jumps and the exploration of the sequence space will depend crucially on the number of sequence pairs which can create offspring outside the populated ball.

Result 5.4 (Pairs which can recombine out of the ball)**.** *The number of pairs in $B_{\sigma_0}(r) \subset \mathbb{H}_L$ which can recombine to sequences not included in this ball is given by*

$$\Omega(r) = \frac{1}{2}\sum_{k=1}^{r}\binom{L}{k}\sum_{m=0}^{k}\sum_{n=r-k}^{r-k+m}\binom{k}{m}\binom{L-r}{n}.$$

Proof. W.l.o.g., choose $\sigma_0 = (0, 0, \ldots, 0)$ and $\sigma = (1, \ldots, 1, 0, \ldots, 0)$ with $d(\sigma_0, \sigma) = k$. Then, define $\zeta^k_{m,n}$ as before. Choose $\sigma' \in B_{\sigma_0}(r)$ s.t. $\sigma' \in \zeta^k_{m,n}$. Let $\mathfrak{R}(\sigma, \sigma') \notin B_{\sigma_0}(r) \Rightarrow n > r - k$ and $k - M + n \leq r$, which yields for the partners, with which σ can produce recombinations out of the Hamming ball:

$$\Omega(r; k(\sigma)) = \sum_{m=0}^{k} \sum_{n=r-k}^{r-k+m} |\zeta^{k(\sigma)}_{m,n}|$$

and as before

$$\Omega(r) = \frac{1}{2} \sum_{k=1}^{r} \binom{L}{k} \sum_{m=0}^{k} \sum_{n=r-k}^{r-k+m} |\zeta^{k(\sigma)}_{m,n}|.$$

□

Figure 5.1 shows the fraction of sequence pairs, which can recombine out of the ball. The number of pairs in a Hamming ball is given by $\frac{1}{2}|B_{\sigma_0}(r)|^2$. If recombination is considered, the Hamming distance alone is often not sufficient in calculations concerning recombination. This inspires the following generalizations.

Definition 5.1. The mapping

$$\mathcal{D} : \mathbb{H}^L \times \mathbb{H}^L \ni (\sigma, \sigma') \mapsto \{i_1, \ldots, i_{d(\sigma,\sigma')}\} \tag{5.2}$$

$$= \{i | \sigma_i \neq \sigma'_i\} \tag{5.3}$$

is designed to give information about the location of differences of two sequences. Note, that $|\mathcal{D}(\sigma, \sigma')| = d(\sigma, \sigma')$.

Now it can be shown, how exactly recombination increases the populated area in this setting.

Result 5.5 (Recombination doubles the radius in each step)**.** *As long as $r < \frac{L}{2}$, the radius of the populated ball is doubled in each generation:*

$$\bigcup_{(\sigma,\sigma') \in (B_{\sigma_0}(r))^2} \mathfrak{R}(\sigma, \sigma') = B_{\sigma_0}(2r).$$

Proof. 1. "$\supseteq$":
Choose $\sigma \in B_{\sigma_0}(r)$ arbitrarily. Choose $\delta \subset \mathcal{D}(\sigma, \sigma_0)$ s.t. $|\mathcal{D}(\sigma, \sigma_0) \setminus \delta| \leq r$ and $|\delta| \leq r$. $\exists_{\sigma' \in B_{\sigma_0}(r)} : \mathcal{D}(\sigma', \sigma_0) = \delta$; $\exists_{\sigma'' \in B_{\sigma_0}(r)} : \mathcal{D}(\sigma'', \sigma_0) = \mathcal{D}(\sigma'', \sigma) \setminus \delta$

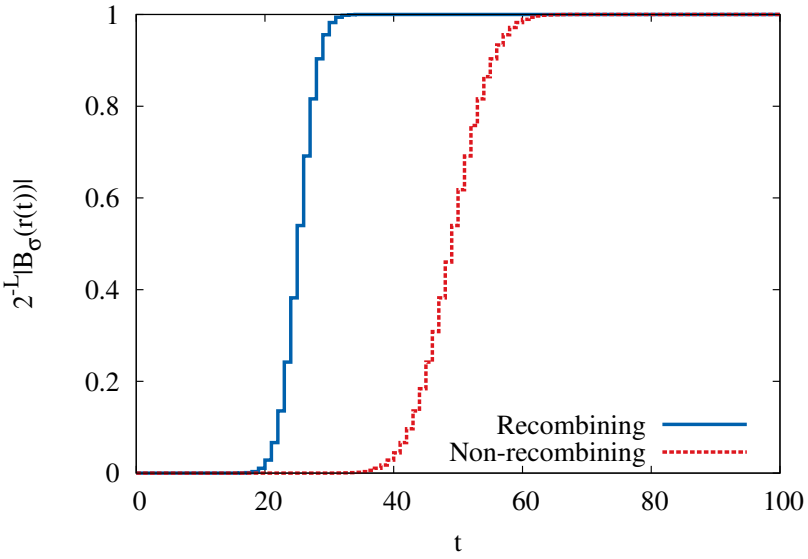


Figure 5.2: The fraction of explored sequences in dependence of time (5.4) for mutation only and mutation with recombination. L=100.

per definition of $B_{\sigma_0}(r)$. Thus, $\sigma \in \mathfrak{R}(\sigma', \sigma'')$.✓

2. "$\subseteq$":
Choose $(\sigma, \sigma') \in (B_{\sigma_0}(r))^2$ arbitrarily. Thus $d(\sigma, \sigma_0) + d(\sigma', \sigma_0) \leq 2r \Rightarrow \forall_{\tilde{\sigma} \in (\sigma, \sigma')} : d(\tilde{\sigma}, \sigma_0) \leq 2r$.✓ □

This means, that with recombination, a population can explore the sequence space exponentially in time, while with mutations only, exploration is linear in time. If now $B_\sigma(r_t)$ denotes the populated ball at time t with radius r_t around σ, this means:

$$\begin{aligned} \text{mutations only} &: B_\sigma(r_{t+1}) = B_\sigma(r_t + 1), \\ \text{mutations and recombination} &: B_\sigma(r_{t+1}) = B_\sigma(2(r_t + 1)), \\ \Rightarrow \text{mutations only} &: B_\sigma(r(t)) = B_\sigma(r_0 + t), \\ \text{mutations and recombination} &: B_\sigma(r(t)) = B_\sigma(2^t(r_0 + 2) - 2). \end{aligned} \tag{5.4}$$

A comparison of both processes is given in fig. 5.2. To get a better understanding of how recombination can lead to jumps in Hamming distance and thus increase the speed at which the hypercube is explored, it is necessary to understand how many sequence pairs from inside a populated ball map outside of this ball in dependence of the target distance. This way, it will

be easier to estimate, how many pairs will in fact produce offspring which is outside the ball and hence help to populate unpopulated areas of the genome space.

Result 5.6 (Pairs to recombine to doubled radius)**.** *The number of pairs* $(\sigma', \sigma'') \in (B_{\sigma_0}(r))^2 \subset (\mathbb{H}_L^2)^2$ *which can recombine to* $\sigma \in B_{\sigma_0}(2r)$ *with* $d(\sigma, \sigma_0) = D > r$ *is given by*

$$m_L(D) = \sum_{n=D-r}^{r} \binom{D}{n} \sum_{m=0}^{r-n} \binom{L-D}{m} \times \sum_{n'=D-n}^{r} \binom{D}{n'-(D-n)} \sum_{m'=0}^{r-n'} \binom{L-D-m}{m'}.$$

Proof. W.l.o.g., choose $\sigma = (0, 0, \ldots, 0)$ and $\sigma' = (1, \ldots, 1, 0, \ldots, 0)$ with $d(\sigma, \sigma') = d$. Then, define

$$\Xi_{m,n}^{d} = \{\nu \in \mathbb{H}_L | m \text{ times 1 in first } d \text{ loci and } n \text{ times 1 in last } L-d \text{ loci}\}.$$

Now choose $\sigma' \in \Xi_{m,n}^{D}, \sigma'' \in \Xi_{m'n'}^{D}$ such that $\sigma \in \mathfrak{R}(\sigma', \sigma'')$. That means, that $n > D - r$ and $n' > D - n$ to get D 1s in the first D loci. At all the places, where σ' has 1s in the last $L-D$ loci, σ'' has to have 0s. The rest can be chosen anywhere from the remaining loci. For each choice of n, m, n', m', there are m chosen of $L-D$, then $n'-(D-n)$ chosen of D, the rest is needed to compensate the 0s in the first D. And m' can be chosen anywhere, where the partner has 0s, thus, out of $L-D-m$. The first n can be chosen out of D. If now the sum is taken over all possible values, the result is reached. □

A plot of this result is given in fig. 5.3. In comparison to the total number of pairs the number of parents which are capable of exploring sequences outside the ball is very small, but the absolute value is quite large. Since the number decreases with the distance from the center of the ball, it is of interest, how many sequences would populate the maximum distance sequences at double radius.

To do so, it will be assumed, that the ball is populated uniformly, i.e. that every sequence is populated with same frequency, while outside the ball no sequence is populated. This assumption is not biological but enables an estimate and analytical methods. From the previous result follows that $\binom{2r}{r}$ pairs of parents are present which could produce offspring with $D = 2r$. Those parents have Hamming distance $2r$ and thus they can produce 2^{2r} different sequences. If all sequences in the ball are populated with equal

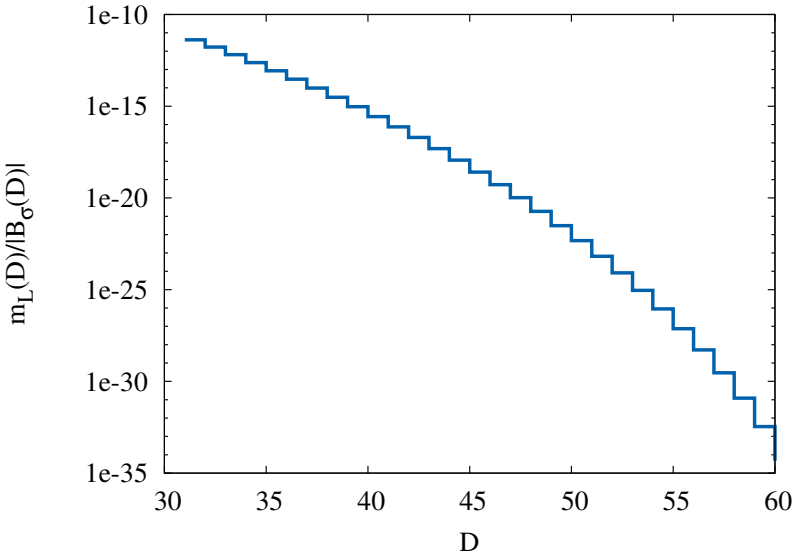


Figure 5.3: A ball with radius $r = 30$ is populated. The normalized number of parents for a sequence outside the ball is plotted in dependence of its distance from the center of the ball. L=100.

frequency p, this leads to a rate at which sequences in distance $D = 2r$ are populated by recombination of

$$2^{-2r}\binom{2r}{r}p^2. \tag{5.5}$$

With the correct combinatorial factor the rate with which sequences recombine from inside of the ball to any sequence at distance $2r$ from the center is given by

$$T_{2r} = \binom{L}{2r}\binom{2r}{r}4^{-r}p^2. \tag{5.6}$$

When all sequences are populated with equal frequency, $p^2 = |B_{\sigma_0}(r)|^{-2}$. A plot of this expression is given in fig. 5.4. It is obvious, that this rate decreases extremely fast with the balls radius, which means, that these long jumps are very unlikely in the case of finite populations. But of course, then the ball will not be populated uniformly and population distributions might occur which enable long jumps with high probability. Nevertheless on average it can be expected that a realistic frequency distribution enables even fewer long jumps than the uniformly populated Hamming ball. Recombination seems to have the ability to increase the speed, at which the sequence space

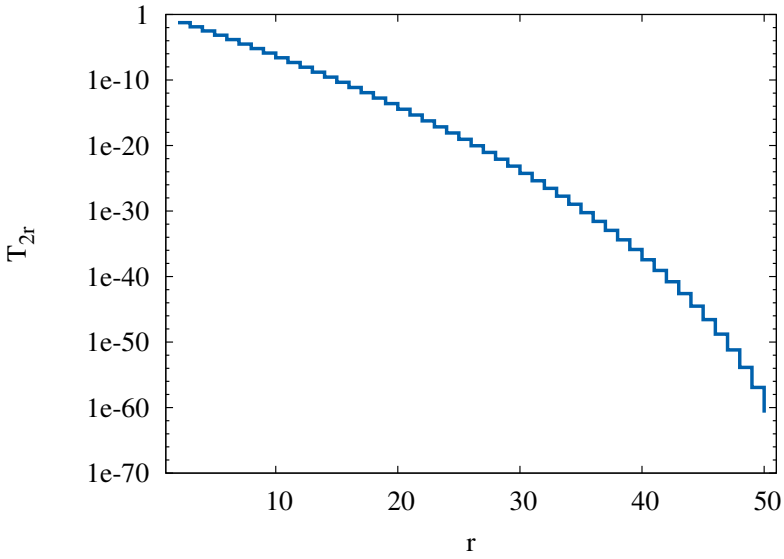


Figure 5.4: The rate with which sequences at $D = 2r$ are being populated from a uniformly populated ball of size r in dependence of r. L=20.

is populated, but the long jumps it can lead to are not the key phenomena of the benefit of sex.

5.2 Recombination in rugged fitness landscapes

If the simplifying conditions from above are relaxed and mutation and selection are included, the analytical description of processes with recombination becomes very complicated. This leads to the situation that there are far more numerical studies at hand than theoretical, starting from early works with few loci up to L dimensional hypercubic models with epistasis (e.g. [110, 111, 112, 113, 114, 115]). The focus will now be on the time series of fitness differences between recombining and non-recombining populations, both under selection and with mutation. As will be shown in the following, analysis of the time series reveals a common, transient benefit of sex, which can be prolonged indefinitely if time varying fitness landscapes are taken into account.

5.2.1 The simulations

The following was published in [116]. Simulations were performed on an exponentially distributed Wrightian RMF-landscape, which uses def. 1.9 to ensure positivity of the fitness values. It comes with parameters c (the mean slope) and λ (parameter from the exponential distribution). If not mentioned otherwise the hypercube dimension in this chapter's simulations is $L = 16$. The results are averaged over runs and landscape realizations. All runs start at σ^* which has on average the smallest fitness value in the landscape. The analyzed dynamics is of the Wright-Fisher type: for finite N, first the population frequencies evolve with the mutation matrix and the selection matrix:

$$p_{t+1} = \frac{\mathcal{S}\mathcal{M}p_t}{\sum_{\sigma' \in \mathbb{H}^L} (\mathcal{S}\mathcal{M}p_t)(\sigma')}.$$

Then, for each sequence σ the number of individuals carrying this sequence is drawn from a Poisson distribution with mean $p(\sigma)N$. After this, the actual number of individuals deviates on average by $\sqrt{N}$ from N and thus a renormalization is carried out. Recombination is performed by replacing an individual by a recombinant created after choosing two parents at random.

In the limit $N \to \infty$, the random sampling is not necessary any more (in fact it is also impossible, obviously). Here, $\mathcal{M}$ is exchanged by $\mathcal{M}^\infty$ for the infinite population limit (see sec. 5.2.4) and recombination is realized with the transition matrix $\mathcal{T}$ (see def. 1.14):

$$p_{t+1} = \frac{\mathcal{T}\mathcal{S}\mathcal{M}^\infty p_t}{\sum_{\sigma' \in \mathbb{H}^L} (\mathcal{T}\mathcal{S}\mathcal{M}^\infty p_t)(\sigma')}.$$

From the already existing studies several results and effects are known:

- In smooth fitness landscapes, recombination yields a faster adaptation towards the global optimum due to Fisher's fundamental theorem [117]. If the global optimum is reached, more diversity is produced around it by recombination load, this can results in a small disadvantage in mean fitness on the long run, compared to non-recombining populations. Nevertheless, if selection is weak, recombination can be beneficial for long times, especially on additive landscape . This can be explained with the Hill-Robertson-effect, see e.g. [118].

- A setting which has been considered numerous times is quite similar to this one, but includes an additional *modifier allele* [119] which determines whether the individual proliferated sexually or asexually.

This way selection for or against recombination can be studied in one population. In a population with a modifier allele, arriving at the global optimum can lead to a selection against recombination to achieve higher mean fitness. This is called the *reduction principle* [120].

- On a two locus landscape ($L = 2$) with sign epistasis, it has been shown, that for a population which is concentrated at the local fitness peak, recombination can be helpful to cross the fitness valley, if r is very small [14, 121, 122]. But the time to cross the valley grows for larger r and diverges in the limit to a critical recombination probability $r \to r_c < 1$ [123, 124] for $N \to \infty$. This means, that for larger, rough landscapes it may be expected that recombining populations show increased incidents of *trapping* at local maxima.

- Two regimes have been identified [125]: if selection is weak recombination can lead to a *delocalization* by creating diversity, if selection is strong and recombination is weak (r is small) the population becomes condensed at particularly fit genotypes. In the following the second regime is analyzed.

5.2.2 The observables and parameters

In the simulations, several parameters control the general behavior of the process. In brevity, those are:

N**:** Population size.

μ**:** Mutation rate, N and μ are often combined to the parameter $N\mu$ which is the average number of mutants per generation for finite N if double mutations can be neglected.

r**:** Fraction of the population which recombines.

λ**:** Parameter from the exponential distribution $P(x) = 1 - e^{-\frac{1}{\lambda}x}$.

c**:** Mean slope of the RMF-landscape.

The following are the used observables of the simulations:

- $\Delta w(t) = \langle w_{\mathrm{r}}(t)\rangle - \langle w_{\mathrm{nr}}(t)\rangle$ is the average of the difference between the fitness of the recombining population (subset 'r') and the fitness of the non-recombining population (subset 'nr') at generation t.

- The entropy $S(t) = -\sum_{\sigma \in \mathbb{H}^L} p_t(\sigma) \log(p_t(\sigma))$ is used to measure diversity in a population. In the same way as above the difference of recombining and non-recombining population is denoted by $\Delta S(t) = \langle S_{\mathrm{r}}(t) \rangle - \langle S_{\mathrm{nr}}(t) \rangle$.

5.2.3 Finite populations

At first, populations of finite size shall be analyzed. The results from simulations[1] are plotted in fig. 5.5. These plots show that the typical behavior seems to be as follows: first $\Delta w(t)$ has a minimum below zero, followed by a maximum which can be above zero and corresponds to the most beneficial time of recombination just before the slope becomes negative again and results in a disadvantage of recombination for later generations. The negative $\Delta w(t)$ at short times can be explained with the monomorphic starting condition: at $t = 0$ recombination cannot provide more diversity because no mutants exist. After the first few generations, several mutations might have been created, but in this case it is very probable that either both parents, or at least one parent carries the wildtype genome. This increases the probability, that the child has again the wildtype genome to 1/2. This is not the case for the non-recombining population, where adaptation can start faster because the chance for a back-mutation is $1/L$. First mutations have to create a more diverse population, before recombinations can increase the number of populated genotypes, see sec. 5.1.2. This means, that in the very beginning, recombination can decrease the adaptive process by reducing diversity, although a look at the slightly positive $\Delta S(t)$ at short time scales seems to indicate that this effect is not of importance here. More important may be a variant of recombinational load. If enough mutants did arise, recombination will create more diversity, this can again lead to a disadvantage for the recombining population because it will decrease the mean fitness since also very unfit sequences are populated. Especially compared to a population which proliferates without recombination: strong selection will concentrate the population around fit genotypes with a comparably small diversity and thus a small amount of unfit populated sequences which could decrease the mean fitness of the population. This can be validated by the plot of $\langle S_{\mathrm{nr}}(t) \rangle$. The deficit for the more diverse population with recombination lies here in the fact, that more diversity means also more less fit genotypes which decrease the mean fitness. But obviously this starting deficit turns into a benefit at intermediate times when the more diverse population can adapt faster in direction of σ^* which can result in a positive $\Delta w(t)$. This transitory benefit

[1] Most of the simulations for finite populations were performed by Stefan Nowak

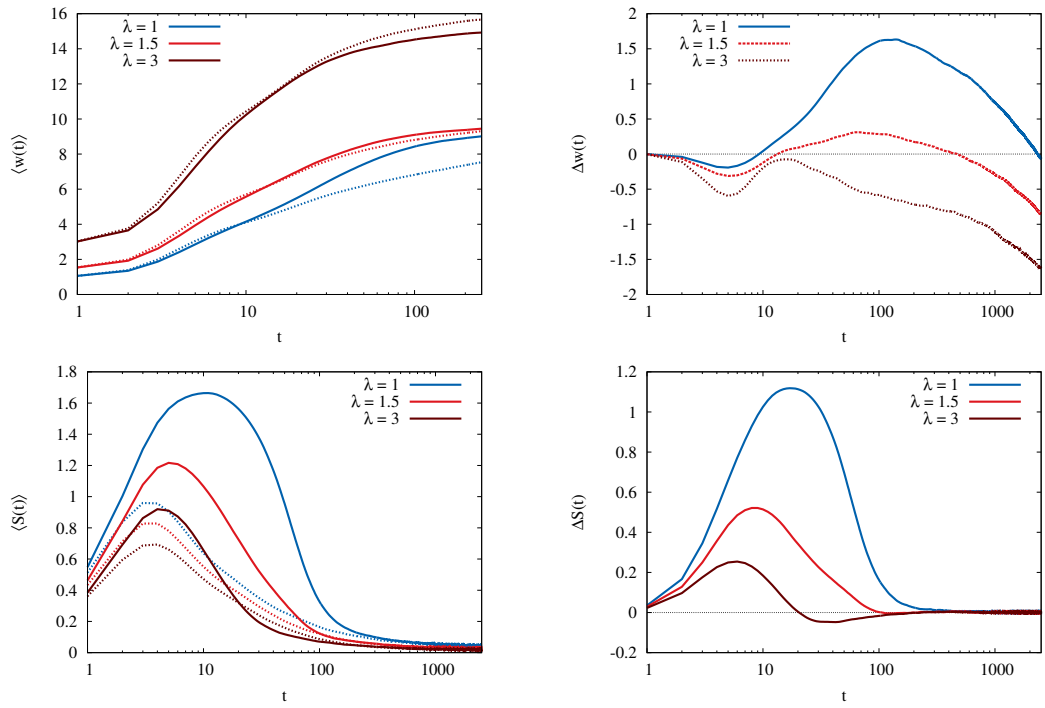


Figure 5.5: Plots of simulation results with $N = 2000$, $N\mu = 4$ and $c = 1$. Dashed lines correspond to non-recombining populations and solid lines recombining ones. The plots concerning fitness time series show the very typical behavior which can produce a transient benefit of recombination on intermediate timescales. The ones concerning entropy show that recombining populations produce much more diversity than non-recombining ones.

of recombination follows after $\Delta S(t)$ becomes positive. The maximum of $\Delta w(t)$ follows the maximum of $\Delta S(t)$.

The Weismann effect thus seems to make a large contribution to the dynamics. Due to the larger diversity, at which a benefit follows in the recombining population, Fisher's fundamental theorem might be involved, too. $\Delta S(t)$ and $\Delta w(t)$ seem to be correlated, as the population with the largest $\Delta S(t)$ yields the largest fitness advantage and the one with the smallest $\Delta S(t)$ yields the smallest. Another benefit might stem from the Fisher-Muller effect, which predicts a benefit of recombination from the gathering of multiple beneficial alleles. On an RMF-landscape, the Fisher-Muller effect allows recombining populations to make jumps in direction of σ^*. This might be connected to a reason why increasing λ reduces the height of the $\Delta w(t)$ maximum, because the landscape becomes rougher and more fitness maxima can arise, which decreases the additivity of fitness effects and

thus increases the probability that two fit parents recombine to a less fit child. Another reason might be, that stronger selection induced by increased λ leads to a decreased diversity which suppresses the Weismann as well as the Fisher-Muller effect. This argument seems to be supported by the data which shows smaller values of $\Delta S(t)$ if λ is increased.

If diversity is important, it is straightforward to consider $N\mu$, the *mutation supply*. If the supply of mutants is on average constant $N\mu =: m$ it is not obvious what happens if N is increased. If m_σ mutants are newly created at some genotype σ, the corresponding frequency is given by $\tilde{p}(\sigma) = \frac{m_\sigma}{N}$. As shown in sec. 5.1 recombination can make the largest steps in the sequence space, if two sequences from the edge of the populated area recombine. The sequences with $\tilde{p} > 0$ thus are sequences, which could lead to a great advantage if recombined with one other. But the probability for this to happen is

$$\begin{aligned}\mathbb{P}(\sigma, \sigma' \text{are newly populated, recombine}) &= 2\tilde{p}(\sigma)\tilde{p}(\sigma') \\ &= \frac{2m_\sigma m'_\sigma}{N^2} < 2\left(\frac{m}{N}\right)^2 \to 0 \ (N \to \infty).\end{aligned}$$

This means, that increasing N can decrease the possibilities of a recombining population as can be seen in fig. 5.6. There plots are shown which demonstrate, that $\Delta S(t)$ does decrease for larger N. Interestingly, although on the smoother landscape ($\lambda = 1$) $\Delta w(t)$ reaches a higher maximum, on the rougher landscape ($\lambda = 1.5$) $\Delta w(t)$ increases with increasing N, albeit the decreasing $\Delta S(t)$. If N is kept constant and μ is increased, recombining and non-recombining populations will benefit from a larger μ. This might be related to section 5.1 where it was shown, that a population spreads faster, if the initially populated Hamming ball is larger.

Fig. 5.7 shows data from simulations with different parameter choices for fixed N and c. The plots show that the diversity for both reproduction types benefits from large $N\mu$. But for a larger λ even though $\Delta S(t)$ is positive at most times, $\Delta w(t) < 0$ in most cases at all times. Additionally $\Delta w(t)$ reduces with increasing μ. Nevertheless, the peak height of $\Delta S(t)$ is increased with rising $N\mu$, and even if $\Delta w(t)$ is always negative as for $\lambda = 2$ and $N\mu = 1$, at around $t = 10$ there is a positive peak in ΔS. But it decreases much faster as in the setting with $\lambda = 1$. At around $t = 30$, only 20 generations later, $\Delta S(t) \approx 0$. This happens in the smoother landscape with $\lambda = 1$ not before $t = 100$. So in the smoother landscape with this particular setting, a diversity advantage of the recombining population is maintained for over 100 generations, roughly four times longer than in the rougher landscape.

Obviously, recombination can under various circumstances lead to an advantage in adaptation. But in all simulations considered in this section

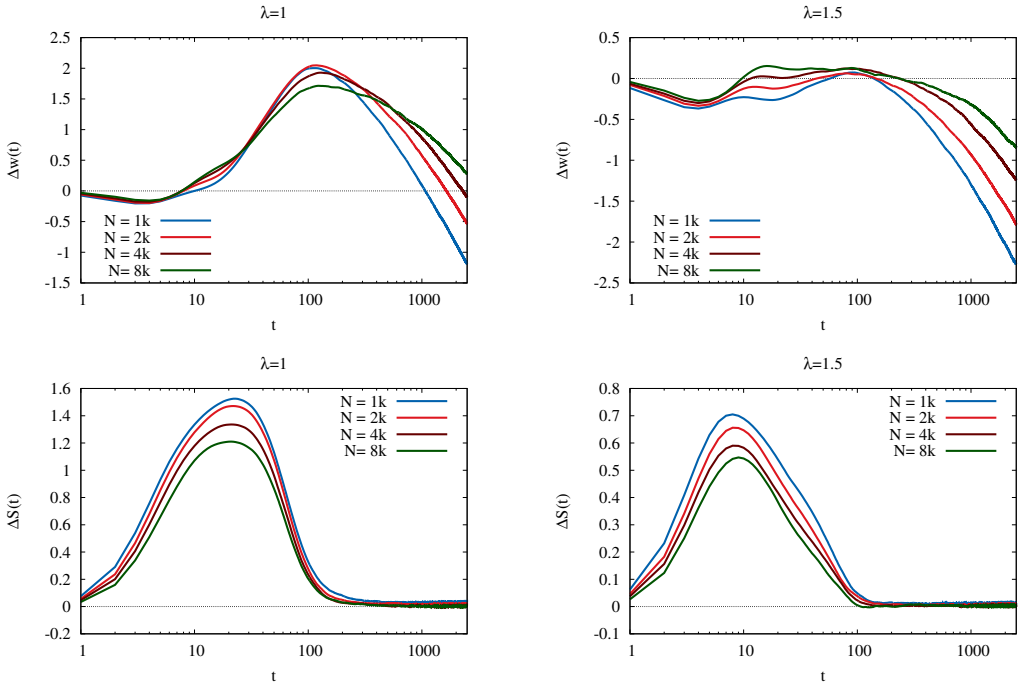


Figure 5.6: Plots of simulation results with $N\mu = 8$ and $c = 1$. Parameters λ and μ are varied.

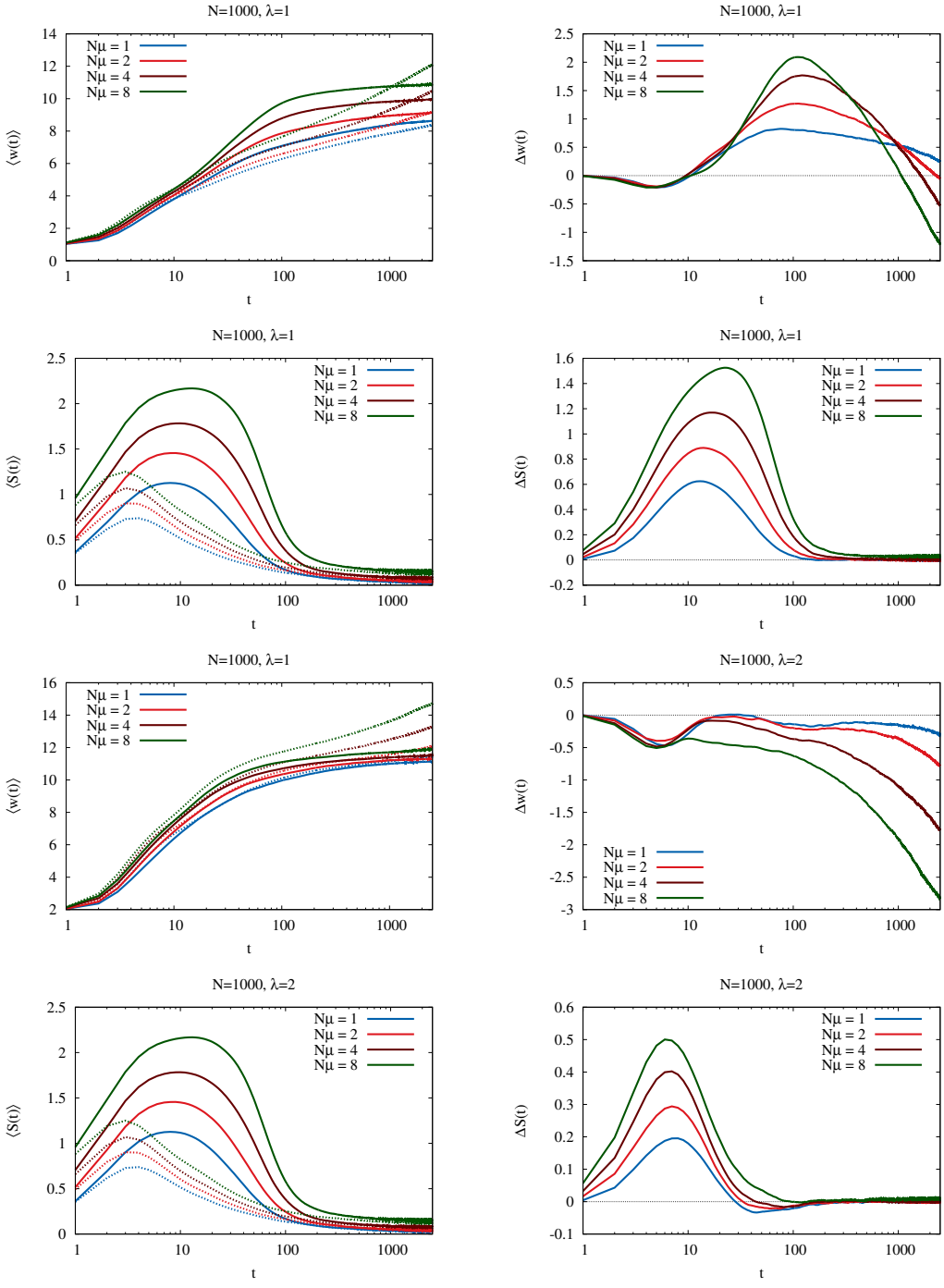


Figure 5.7: Plots of simulation results with $N = 1000$ and $c = 1$. Dashed lines correspond to non-recombining populations and solid lines to recombining ones. Parameters λ and $N\mu$ are varied.

$\Delta w(t)$ will get a negative slope at some point in time and become negative eventually. In an attempt to understand why recombining populations cannot maintain the adaptive advance it has on intermediate timescales, it is helpful to look at results in small landscapes where also analytical results are present [14, 121, 122, 123, 124, 126].

In a landscape on a two-dimensional hypercube with sign epistasis and one local and one global maximum (similar to the illustration of sign epistasis in fig. 1.2), a population which is concentrated on the local maximum would benefit if the fitness valley could be crossed and the global maximum would be reached. Therefore, a non-recombining population would have to create mutants in the valley, which survive long enough to create a double mutant on the global maximum, which gets the chance to populate the global maximum. A recombining population additionally has the possibility to create mutants in each of the valley states, which can recombine to the sequence of the global maximum. The waiting time, until valley mutants have been created and afterwards most of the population has moved to the global maximum is called *escape time* t_{esc} and is measured in generations. Despite the fact, that for larger N more mutants are created, it has been shown, that a population on a local fitness maximum has an expected escape time which increases with N. If the fitness valley around the peak has a critical depths, for mutation only populations, this time increases algebraically in N [14, 126], while for populations with recombination a critical recombination rate r_c exists and the escape time is exponentially in N, $\log(t_{\text{esc}}) \sim N(r - r_c)^{\frac{3}{2}}$ for r close to r_c so that it diverges for $r \geq r_c < 1$ for $N \to \infty$ [122]. This result follows from the fact, that the recombining population has the disadvantage, that if a mutant on the global maximum is created, it will most probably recombine with one from the local maximum, which will with probability $\frac{1}{4}$ recombine back to the local maximum and with probability $\frac{3}{4}$ away from it. This disadvantage is also present in larger fitness landscapes. Of course, the larger L provides more possible paths away from the local peak to a fitter sequence. This implies an advantage for larger L. Nevertheless, it also implies, that for a peak which is high enough, recombining populations can be *trapped*, because every possible escape path would have a corresponding critical recombination rate $r_c < 1$. If the populations with recombination are then trapped longer than those without recombination, this would explain the long time fitness deficit. From simulations, the *escape rate* $f_{\text{esc}}(r)$ was measured[2]: on a landscape with $\lambda = c = 1$ the number of escape events was measured as well as the number of trapping events for 500 generations. Then the escape rate can be calculated as $f_{\text{esc}}(r) = \frac{\#\text{escapes}}{\#\text{trappings}}$. A population

[2]The simulations on the escape rate were performed by Ivan G. Szendro

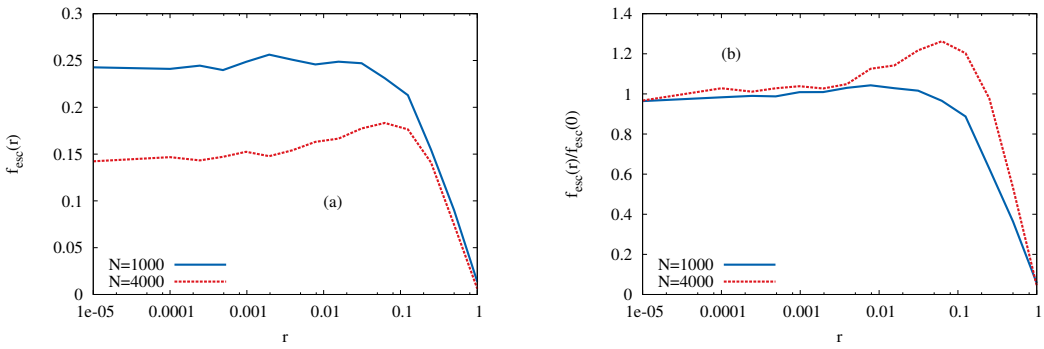


Figure 5.8: The figure shows plots of simulation data for the escape rate $f_{\text{esc}}(r)$ for two different population sizes and $N\mu = 4$. Panel (a) shows $f_{\text{esc}}(r)$ in dependence of r, while (b) shows a normalized function $f_{\text{esc}}(r)/f_{\text{esc}}(0)$. Here, a maximum larger than 1 indicates, that an optimal recombination rate exists, for which the escape time is reduced in comparison to non-recombining populations.

was assumed to be trapped on a maximum σ, if at some time t, $p_t(\sigma) \geq 0.7$. A population is considered to have escaped, if at $t' > t$: $p_{t'}(\sigma) < 0.5$. The data for $N\mu = 4$ is given in fig. 5.8. Note, that due to the fact that each escape was preceded by a trapping event $f_{\text{esc}} \to 1$ ($t \to \infty$). Nevertheless, here escapes were so rare, that this rate is still a useful measure. Obviously, the escape rate is dramatically smaller for $r = 1$, and is further decreased for larger N, as expected.

In the context of the temporal development of sex, experimental results of Becks and Agrawal [127] are at hand which can also be interpreted with the results of this chapter. There, the facultatively sexual rotifer *Brachionus calyciflorus* was used. These wheel animals can change their reproduction scheme between sexual and asexual. In the experiment, the rotifer was set under selective pressure by an environmental change. This was conducted by a change of their algal food source and the concentration of NaCl which set the population in an 'unfit' state. In the following phase of adaptation the rate of sex increased, while after a fit state was reached, it declined. Additionally, an estimate of fitness showed, that the mean fitness of the sexually reproducing rotifers starts lower than the mean fitness of the asexually reproducing directly after the environment change. But after some generations the sexually reproducing gain higher mean fitness, until at intermediate times the asexually reproducing overtake again and reach the higher mean fitness towards the end of the experiment. Thus, these experimental findings coincide nicely with the numerical results presented in this chapter and show exactly the same pattern. This demonstrates the significance of focus on temporal development and adaptation in the context

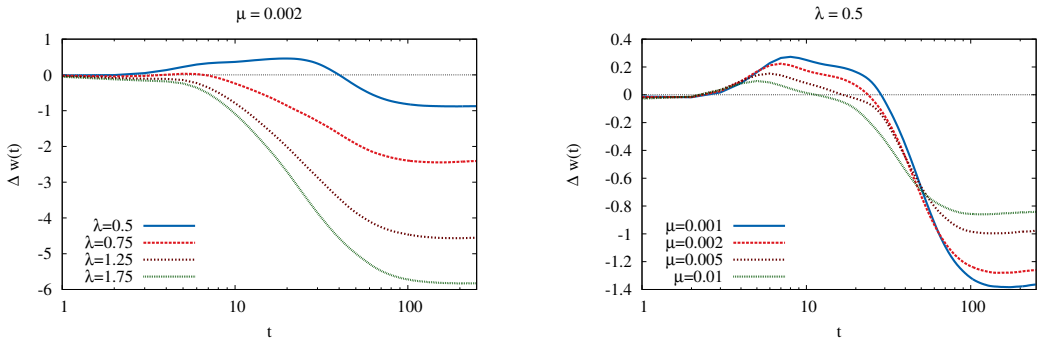


Figure 5.9: Results of simulations in the limit $N \to \infty$. Both figures show $\Delta w(t)$ for different parameters. The overall phenomenology is quite similar to the finite N results. Still there are parameters for which a transitory advantage can be found, especially for small λ and μ. For larger λ, the benefit of recombination vanishes. For increasing μ it decreases.

of evolutionary biology. Note, that the importance of transient, in contrast to asymptotically or long time effects, was also stressed by Hastings [128] in the context of ecological systems.

5.2.4 Infinite populations

Although in sec. 5.1.2 single step mutations were used in the context of infinitely large populations to make the discussed problems analytically feasible it is not very realistic. $N \to \infty$ leads to $\mu^n N \to \infty$ for every fixed $\mu < 1$ and any n. This means, that double, triple, multiple mutations in general will be produced and thus the whole sequence space will be populated after one time step, each genotype with a frequency dependent on the Hamming distance to the wild type. The single mutation matrix was introduced in def. 1.12 and will get a superscript here:

$$\mathcal{M}^{(1)} = \mathbb{I}(1-\mu) + \frac{\mu}{L}\mathcal{A}.$$

To include double mutations, the matrix becomes

$$\mathcal{M}^{(2)} = \mathbb{I}(1-\mu) + \left(\mathbb{I}(1-\mu) + \frac{\mu}{L}\mathcal{A}\right)\frac{\mu}{L}\mathcal{A}.$$

For arbitrary orders n of mutations, the matrix is expanded in the same way

$$\begin{aligned}\mathcal{M}^{(n)} &= \mathbb{I}(1-\mu) + \left(\mathbb{I}(1-\mu) + \left(\mathbb{I}(1-\mu) + \ldots\frac{\mu}{L}\mathcal{A}\right)\ldots\right)\frac{\mu}{L}\mathcal{A} \\ &= \sum_{j=0}^{n}(1-\mu)\left(\frac{\mu}{L}\mathcal{A}\right)^j. \end{aligned} \tag{5.7}$$

For matrices there exists a geometric series, which is given by $\sum_{n\geq 0} T = (\mathbb{I} - T)^{-1}$ if $\lim_{k\to\infty} T^k = 0$. Applying this yields the mutation matrix for infinite N [116]:

$$\sum_{j=0}^{n}(1-\mu)\left(\frac{\mu}{L}\mathcal{A}\right)^j \overset{n\to\infty}{\longrightarrow} (1-\mu)\left(\mathbb{I}-\frac{\mu}{L}\mathcal{A}\right)^{-1} = \mathcal{M}^{(\infty)}. \tag{5.8}$$

As discussed above, $\mathcal{M}^{\infty}$ populates every single genotype right after the first application. Hence, after the first generation, all sequences are available in the population. This means, that a faster exploration of the genotype space is not a possible benefit anymore. Nevertheless, despite the naive guess, that now recombination cannot yield an advantage, the situation is still far from trivial. Still by increasing S recombination can yield more diversity. On the other hand, as mentioned above, trapping can be more pronounced for larger populations, which indicates, that this will be an important issue especially for infinite population sizes. In fact, on empirical fitness landscapes trapping has been observed in the infinite N limit [36].

Since the process is now deterministic in the sense, that recombination, selection and mutation do not introduce stochasticity as in the finite N cases, averaging has to be performed over landscapes, only. Due to the large number of entries in the $\mathcal{S}$ matrix ($2^L \times 2^L \times 2^L$) the considered sequence length is reduced to $L = 8$ to prevent memory issues during the simulations. And since $\forall_{\mu\in\mathbb{R}} : \mu N \to \infty$ ($N \to \infty$), μ is the better parameter here. The fig. 5.9 shows, that the qualitative behavior is similar to the finite N results. Particularly the transient benefit of recombination can still be observed. This means, that recombination does in fact accelerate adaptation not only by exploring the sequence space faster, which is completed now after the first generation. Nevertheless, the long period of disadvantage is also present in the simulations. Thus, the trapping still occurs and ends the benefit of the recombining population.

5.2.5 Seascapes

Up to now, this work only considered time independent fitness landscapes. But of course, natural environments are changing. Either this might happen without the organism changing its habitat, for example due to an Ice Age, after an asteroid collision, etc., or if a population changes its location. Such a time dependent fitness landscape is also called a *fitness seascape* [129]. Time dependent fitness landscapes have already been studied, but usually the considered models had one, moving global optimum, e.g. [130, 131, 132, 38, 133]. Here, based on the RMF-model, two ways are chosen to change

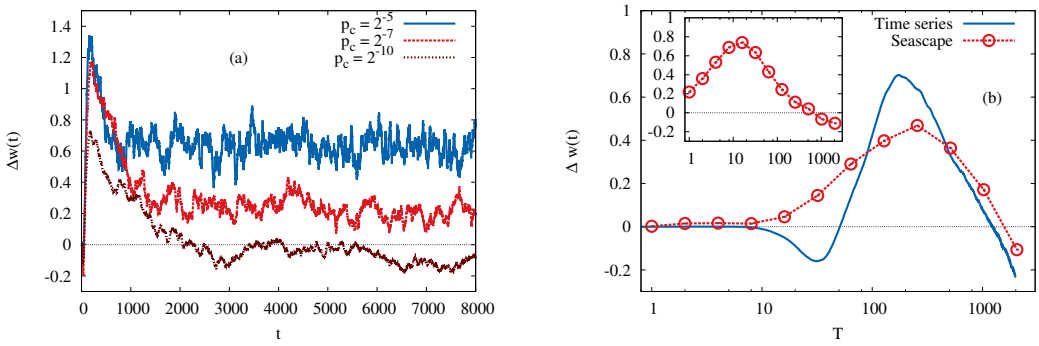


Figure 5.10: Plots of simulation results on seascapes. (a) shows time series as before for seascapes with different changing probabilities p_c. Note that in contrast to previous figures no logarithmic scales were used. The chosen changing scheme is the hard reset. The plot shows, that for some p_c a plateau occurs in $\Delta w(t) > 0$ which means, that recombination stays beneficial in these cases. In (b) the dots are *mean plateaus* of $\Delta w(t)$ in a hard reset seascape in the sense, that from data sets like the ones from (a) the mean plateau height was determined for a given p_c. For each p_c there is a typical time at which the change of the seascape happens and this is $T = 1/p_c$. The solid line is for comparison and shows a time series of $\Delta w(t)$ without reset. The inset shows the mean plateau heights for soft reseted seascapes. Parameters are $c = 0.75$, $N = 2000$, $\mu = 0.0025$.

the fitness in time. In both cases, the random numbers are redrawn and the reference sequence σ^* changes its location in each time step with the *changing probability* p_c. In the *soft reset* setup, σ^* moves to one of its neighbors, while in the *hard reset* setup σ^* moves to another sequence randomly picked from $\mathbb{H}^L$. Connected to the changing probability is the typical changing time $T_c = \frac{1}{p_c}$. In the context of fitness seascapes, recombination gives a much greater benefit if the time in which recombination leads to a faster adaptation is tuned to the timely changes in the fitness seascape. Fig. 5.10 shows timely varying fitness prolonging the benefit of recombination indefinitely. It also shows that the optimal T_c (the one for which recombination is most beneficial) coincides with the time at which in the static fitness landscape $\Delta w(t)$ has its maximum. This behavior resembles *resonance*: the long time advantage of recombination is maximized, if the changing frequency $1/T_c$ is close to the 'resonance frequency' of the intrinsic timescale of adaptation.

The Red Queen hypothesis was inspired by the following line from *Through the Looking-Glass* [134]: "Now, here, you see, it takes all the running you can do, to keep in the same place". As mentioned in the introduction, it states, that to survive in a changing environment it is not necessarily the total fitness which can be achieved by an organism but the fitness benefit

an organism can establish and maintain against a competing species. In a changing environment, a fit state can fast become unfit and then it is important to adapt fast. This coincides very nicely with the results from this and the previous chapter which show, that in static fitness landscapes the higher fitness is reached by non-recombining organisms, while in changing environments a fitness benefit can be gained and maintained by recombining populations.

Chapter 6

Conclusions

In this thesis, adaptive processes on theoretical models of fitness landscapes, as well as the fitness landscapes themselves were analyzed with help of analytical and numerical methods. Additionally two methods were presented to fit theoretical model-landscapes to experimental data.

6.1 Summary

In sec. 2.2 the fitness maxima as well as the correlations in an RMF-landscape [17, 16, 18] are discussed as characteristic features of this model. While for the number of maxima an expression was found in the case of large L on Gumbel distributed RMF-landscapes, the general behavior of M is described in terms of EVT for arbitrarily distributed RMF-models. In order to clarify the picture sec. 2.3 is devoted to the study of the number of exceedances which is identified as an important indicator for or against a fitness landscape model in the context of finding an appropriate description for a given experimental fitness dataset [66]. Analytical expressions are presented for an exponentially distributed RMF-landscape. For other distributions computer simulations deliver reliable data. With help of a parameter scan a GPD RMF-landscape is brought in line with experimental observations of Miller et al. [66].

A different approach is used in ch. 3. Here with help of a Fourier analysis [87], the family of fitness landscapes which can be represented as a superposition of LK-landscapes [19, 86] is examined and the amplitude spectra are calculated. This results in condensing the landscapes information from 2^L to L data points. Although this happens not without loss of information, it is argued, that important information about the correlations and interactions within the models is conserved. By spectral deconstruction of experimentally obtained fitness landscapes [68, 69], it is shown that by

comparison to LK-landscape superpositions parameters can be identified with which the spectrum of the model fits the experimental one. Since also the RMF-model belongs to the family of LK-superpositions, the findings do extend the previous chapter.

In ch. 4 adaptive processes in the SSWM limit, adaptive walks [30, 31], were analyzed. Based on previous results [99, 83], a more systematic calculation of the natural adaptive walk is presented on a GPD HoC-model. The extension of the dynamics to the RMF-model is elaborated and numerical as well as analytical results are shown. For the natural adaptive walk [31] the walk length is analyzed with help of computer simulations and the analytic form of the result is concluded. For the random adaptive walk [59], a phase transition in the walk length is found on exponentially distributed RMF-landscapes. The analysis of the greedy adaptive walk [97] yields analytic results for the mean walk length on RMF- landscapes, but here especially Gumbel distributed ones. It shall be emphasized, that these results have pendants in the theory of spin glasses. Especially the phase transition which is found for the random adaptive walk can also be interpreted as a zero temperature kinetic phase transition of Metropolis dynamics on a random energy model in a magnetic field. The phase transition then occurs between a phase of weak field, where the dynamics gets trapped in a metastable state and a strong field phase, in which the ground state is found.

Additionally to adaptation by mutation and selection, also recombinatorial processes are examined. Under simplifying assumptions rigorous results are obtained by the analysis of Hamming balls on the hypercube, see sec. 5.1. It can in particular be shown, that a population which proliferates with single step mutations and recombination can explore the sequence space exponentially fast in time, whereas a population without recombination explores it only linearly in time. For more complex systems of recombining populations on RMF-landscapes, results of numerical studies are presented in sec. 5.2. Here it is shown, that recombining populations yield on intermediate times more diversity, which is also known as the Weismann effect [45]. Furthermore, the simulation results show that recombining populations can have a transient fitness benefit compared to populations which adapt on the same landscape without recombination. But on the long run the non-recombining populations have a higher mean fitness. From analytical work on a two locus system with recombination [122], it is known, that recombination prolongs the escape time which is needed to move away from a local fitness maximum if a population is trapped there. By means of simulations it is shown that this effect is also present in larger systems. It is thus concluded that, if recombining population get trapped at a local fitness maximum, they cannot escape for much longer times than non-recombining ones in a similar

position. This way, the presented transient benefit of sex can be explained. These findings coincide nicely with an experimental study on the adaptation dynamics of a facultatively sexual rotifer [127], where the rate of sexual reproduction as well as the fitness reached by sexually proliferating rotifers increases in times of selective pressure and declines after the reach of fitness plateaus. Furthermore, results are shown of adaptation with recombination on a fitness landscape varying in time. Here, the fitness benefit of the recombining populations can be maintained. This is because in a fluctuating environment, the trapping is abrogated regularly while the populations have an enduring need of fast adaptation. Because in the previous chapter it was shown, that initially this is rather provided by proliferation with recombination, a permanent benefit of sex can be established. This is in accordance with the Red Queen hypothesis [48] which states, that the main reason for adaptation is the survival of the species in competition with other organisms in a permanently changing environment. This implies, that not necessarily the ultimate fitness an organism achieves is important, but more the possibility to maintain its current fitness and the fitness benefit it has against other species. Which is exactly what recombination enables according to the results of this section.

6.2 Outlook

In total, several results in evolutionary biology have been elaborated which are partially connected and some go beyond the field. Since the RMF-landscape could be fitted to different sets of experimental data it can be assumed, that it is a fairly realistic model. Nevertheless, it might be the case, that given the fact, that a landscape of size $L \leq 9$ shows only a part of the complete landscape, a full genotype to fitness mapping is in some areas RMF-like, but not measured over the entire genome length. But since the largest complete empirical fitness landscape known to the author is analyzed in this thesis, the analysis of larger fitness landscapes and the question if those are RMF-like remains to future research. This problem is also known as *the problem of scale* [64]. Still, the successful fitting increases the importance of the other results obtained for this landscape type, which are mainly properties of the landscape structure and topography. Nevertheless, the study of adaptive walks is also of relevance in this context, since the adaptive walk length is closely connected to the number of fitness maxima. The GAW length is an upper limit of the mean Hamming distance to the closest fitness maximum. The behavior of the adaptive walk length with increasing c is another indicator for the decreasing ruggedness in comparison

to the HoC landscape, as is the decreasing number of maxima. Additionally, together with the results of the study of recombination on rugged fitness landscapes, the phase transition in the RAW length might indicate not only two regimes in the context of adaptive walks, but also in the context of sexual recombination. Since in the presence of many fitness maxima, the probability, that a recombining population gets trapped is increased, recombination might especially be beneficial on landscapes in the phase of long adaptive walks. This leads to an open question: Are the fitness landscapes of recombining organisms preferably in the phase of long adaptive walks, which has a lower number of local maxima?

The study of the RMF-model and adaptive processes on it did also reveal, that in contrast to the HoC case, EVT is important to understand the phenomena, but it does not suffice. The linear drift of the landscapes increases the importance of the tail behavior of the underlying probability distribution. From this follows not only the phase transition for exponentially distributed RMF-landscapes, but it is also of importance in the discussion of the number of maxima and probably also has an impact on the number of exceedances, which could not be calculated. The fact, that the three probability classes are not enough to describe the general behavior could imply other new, yet unknown, phenomena and emphasizes the importance of analytical methods besides the standard tools, not only applied to theory but especially in the evaluation of experimental data.

Beyond the RMF-landscape, it is now possible to fit superpositions of LK-models to experimental data using the properties of the amplitude spectra of the fitness landscapes. In fact, with enough fitting parameters, every possible amplitude spectrum can be fitted by such a superposition. Nevertheless, the use of fitting such a superposition might be questioned if it is not sparse any more. But, if a sparse fit is possible, it can be done easily and the resulting parameters give directly information about the interactions in the genome and its implications to the fitness landscape. Of course, this method is not restricted only to fitness landscapes, but all kinds of functions on a hypercube, for example energy landscapes of spin glasses. This method is very promising to understand more about the nature of fitness landscapes, and it is up to future research to analyze especially larger complete data sets for a better classification of the naturally occurring landscape types.

Regarding recombination, this thesis shows, how important the time evolution of the observables of the dynamics is. Not only is the benefit of sex on a static landscape only transient, but it can be prolonged indefinitely if the fitness landscape is varying in time with an adequately chosen changing time. The nature of both effects cannot be observed, if not the full time evolution is measured. The benefit of sex in time dependent fitness landscapes indicates

the importance of the analysis of fitness seascapes. Although the time dependence in natural environments is obvious, the amount of research in this field, especially concerning epistatic and correlated landscapes has many unanswered questions and the analytical methods for large landscapes are not sufficient. For experimental research an important open question is: How can fitness landscapes of sexually reproducing organisms be characterized? Besides the phase mentioned above, the epistatic interactions, the spectrum and for time dependent fitness landscapes the investigation of the existence of the recombinatorial resonance would give great insights of the causes and nature of recombination, adaptation and fitness land- and seascapes.

Bibliography

[1] D. Adams. Speech at Digital Biota 2, 1998. URL `http://www.biota.org/people/douglasadams/index.html`. accessed: 5.5.14.

[2] R. Dawkins. *the selfish gene*. Oxford University Press, 2nd edition, 1989.

[3] Cburnett. Hamming distance cube for 4-bit binary numbers, 2006. URL `http://en.wikipedia.org/wiki/File:Hamming_distance_4_bit_binary.svg`. File: `Hamming_distance_4_bit_binary.svg`, accessed: `14.4.14`.

[4] C. R. Darwin. *The variation of animals and plants under domestication*, volume 1. John Murray, London, 1st edition, 1868.

[5] C. R. Darwin. *On the origin of species, by means of natural selection or the preservation of favoured races in the struggle of life*. John Murray, London, 5th edition, 1868.

[6] Thomas C. Leonard. Origins of the myth of social darwinism: The ambiguous legacy of Richard Hofstadter's social Darwinism in american thought. *J. Econ. Behav. Organ.*, 71:37–51, 2009.

[7] H. A. Orr. Fitness and its role in evolutionary genetics. *Nature Review Genetics*, 10:531–539, 2009.

[8] P. E. O'Maille, A. Malone, N. Dellas, B. A. Hess, Jr., L. Smentek, I. Sheehan, B. T. Greenhagen, J. Chappell, G. Manning, and J. P. Noel. Quantitative exploration of the catalytic landscape separating divergent plant sesquiterpene synthases. *Nat. Chem. Biol.*, 4:617–623, 2008.

[9] M. F. Schenk, I. G. Szendro, M. L. M. Salverda, J. Krug, and J. A. G. M. de Visser. Patterns of epistasis between beneficial mutations in an antibiotic resistance gene. *Mol. Biol. Evol.*, 30:1779–1787, 2013.

[10] J. Wynne McCoy. The origin of the "adaptive landscape" concept. *Am. Nat.*, 113(4):610–613, 1979.

[11] S. Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Int. Proceedings of the Sixth International Congress on Genetics*, 1:356–366, 1932.

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

[13] V. Rao and V. Nanjundiah. J.B.S. Haldane, Ernst Mayr and the Beanbag Genetics Dispute. *J. Hist. Biol.*, 44:233–281, 2011.

[14] L. Chao D. M. Weinreich, R. A. Watson. Perspective: Sign epistasis and genetic consraint on evolutionary trajectories. *Evolution*, 59:1165–1174, 2005.

[15] J. F. C. Kingman. A simple model for the balance between selection and mutation. *J. Appl. Prob.*, 15:1–12, 1978.

[16] T. Aita, H. Uchiyama, T. Inaoka, M. Nakajima, T. Kokubo, and Y. Husimi. Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape: application to prolyl endopeptidase and thermolysin. *Biopolymers*, 54:64–79, 2000.

[17] T. Aita and Y. Husimi. Adaptive walks by the fittest among finite random mutants on a Mt. Fuji-type fitness landscape ii. effect of small non-additivity. *J. Math. Biol.*, 41:207–231, 2000.

[18] J. Neidhart, I. G. Szendro, and J. Krug. Adaptation in tunably rugged fitness landscapes: The Rough Mount Fuji model. *Genetics*, 2014. doi: 10.1534/genetics.114.167668.

[19] S. A. Kauffman and E. D. Weinberger. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *J. Theor. Biol.*, 141:211–245, 1989.

[20] S. A. Kauffman. *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford University Press, USA, 1993.

[21] J. J. Welch and D. Waxman. The NK model and population genetics. *J. Theor. Biol.*, 234(4):329–340, 2005.

[22] T. Aita. Hierarchical distribution of ascending slopes, nearly neutral networks, highlands, and local optima at the dth order in an NK fitness landscape. *J. Theor. Biol.*, 254:252–263, 2008.

[23] B. Østman, A. Hintze, and C. Adami. Impact of epistasis and pleiotropy on evolutionary adaptation. *Proc. Roy. Soc. B*, 279:247–256, 2012.

[24] J. Franke and J. Krug. Evolutionary accessibility in tunably rugged fitness landscapes. *J. Stat. Phys.*, 148:705–722, 2012.

[25] N. G. van Kampen. *Stochastic Processes in Chemistry and Physics*. North-Holland Personal Library, 2007.

[26] J. Franke. *Statistical topography of fitness landscapes*. PhD thesis, Universität zu Köln, 2012.

[27] M. Kimura. Process leading to quasi-fixation of genes in natural populations due to random fluctuation of selection intensities. *Genetics*, 39:280–296, 1954.

[28] M. Kimura and T. Ohta. The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, 61:763–771, 1968.

[29] M. Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47:713–719, 1962.

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

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

[32] J. Maynard Smith and J. Haigh. The hitch-hiking effect of a favorable genes. *Genet. Res.*, 23:23–35, 1974.

[33] R. E. Michod and B. R. Levin. *The Evolution of Sex: An Examination of Current Ideas*. Michigan: Sinauer Associates, 1987.

[34] M. W. Feldman, S. P. Otto, and F. B. Christiansen. Population genetic perspectives on the evolution of recombination. *Annu. Rev. Genet.*, 30: 261–295, 1997.

[35] S. P. Otto and T. Lenormand. Resolving the paradox of sex and recombination. *Nature Reviews Genetics*, 3:252–261, 2002.

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

[37] S. P. Otto. The evolutionary enigma of sex. *Amer. Nat.*, 174:1–14, 2009.

[38] B. Charlesworth and D. Charlesworth. An experiment on recombination load in Drosophila melanogaster. *Genet. Res.*, 25:267–274, 1975.

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

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

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

[42] R. A. Fisher. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, 1930.

[43] H. J. Muller. Some genetic aspects of sex. *Am. Nat.*, 66:118–138, 1932.

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

[45] A. Weismann. *Essays on heredity and kindred biological subjects*. Oxford University Press, Oxford, UK, 1889.

[46] G. R. Price. Fisher's "fundamental theorem" made clear. *Ann. Hum. Genet.*, 36:129–140, 1972.

[47] W. J. Ewens. An interpretation and proof of the fundamental theorem of natural selection. *Theor. Popul. Biol.*, 36(2):167–180, 1989.

[48] L. Van Valen. A new evolutionary law. *Evol. Theor.*, 1:1–30, 1973.

[49] P. F. Stadler and G. P. Wagner. Algebraic theory of recombination spaces. *Evol. Comp.*, 5(3), 1998.

[50] R. A. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest and smallest member of a sample. *Proc. Cambridge Phil. Soc.*, 24:180–190, 1928.

[51] B. V. Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *Ann. Math.*, 44:423–453, 1943.

[52] J. Pickands. Statistical inference using extreme order statistics. *Ann. Stat.*, 3:119–131, 1975.

[53] A. Balkema and L. de Haan. Residual life time at great age. *Ann. Prob.*, 2:792–804, 1974.

[54] E. M. Furrer and R. W. Katz. Improving the simulation of extreme precipitation events by stochastic weather generators. *Water Resour. Res.*, 44:W12439, 2008.

[55] M. Mezard, G. Parisi, and M. A. Virasoro. *Spin Glass Theory and Beyond*. World Scientific Publishing Co. Pte. Ltd., Singapore, 1987.

[56] B. Derrida. Random-energy model: Limit of a family of disordered models. *Phys. Rev. Lett.*, 45:79, 1980.

[57] B. Derrida. Random-energy model: An exactly solvable model of disordered-systems. *Phys. Rev. B*, 24(5):2613–2626, 1981.

[58] D.J. Gross and M. Mezard. The simplest spin glass. *Nucl. Phys. B*, 240(4):431–452, 1984.

[59] H. Flyvbjerg and B. Lautrup. Evolution in a rugged fitness landscape. *Phys. Rev. A*, 46:6714–6722, 1992.

[60] G. Ben Arous, A. Bovier, and V. Gayrard. Glauber dynamics of the random energy model I. metastable motion on the extreme states. *Commun. Math. Phys.*, 2001. doi: 10.1007/s00220-003-0798-4.

[61] A. M. Sutton, L. D. Whitley, and A. E. Howe. Computing the moments of k-bounded pseudo-Boolean functions over Hamming spheres of arbitrary radius in polynomial time. *Theor. Comp. Sci.*, 425:58–74, 2012.

[62] D. Stauffer and A. Aharony. *Introduction to percolation theory*. Taylor & Francis Ltd, London, revised 2nd edition, 1994.

[63] S. Nowak and J . Krug. Accessibility percolation on n-trees. *EPL*, 101: 66004, 2013.

[64] J.A.G.M. de Visser and J. Krug. Empirical fitness landscapes and the predictability of evolution. *Nature Reviews Genetics*, 15:480 –490, 2014.

[65] N. Colegrave and A. Buckling. Microbial experiments on adaptive landscapes. *BioEssays*, 27:1167–1173, 2005.

[66] C. R. Miller, P. Joyce, and H. A. Wichman. Mutational effects and population dynamics during viral adaptation challenge current models. *Genetics*, 187:185–202, 2011.

[67] D. R. Rokyta, P. Joyce, S. B. Caudle, and H. A. Wichman. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. *Nat. Gen.*, 37(4):441–444, 2005.

[68] D. W. Hall, M. Agan, and S. C. Pope. Fitness epistasis among six biosynthetic loci in the budding yeast saccharomyces cerevisiae. *J. Hered.*, 101:75–84, 2010.

[69] J. Franke, A. Klözer, J. A. G. M. de Visser, and J. Krug. Evolutionary accessibility of mutational pathways. *PLoS Comput. Biol.*, 7(8), 2011. e1002134. doi:10.1371/journal.pcbi.1002134.

[70] T. Aita, M. Iwakura, and Y. Husimi. A cross-section of the fitness landscape of dihydrofolate reductase. *Protein Eng.*, 14:633–638, 2001.

[71] J. Franke, G. Wergen, and J. Krug. Records and sequences of records from random variables with a linear trend. *J. Stat. Mech.*, P10013, 2010.

[72] G. Wergen, J. Franke, and J. Krug. Correlations between record events in sequences of random variables with a linear trend. *J. Stat. Phys.*, 144:1206–1222, 2011.

[73] R. L. Graham, D. E. Knut, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley Publishing Company, 2nd edition, 1994.

[74] A. S. Perelson and C. A. Macken. Protein evolution on partially correlated landscapes. *Proc. Nat. Acad. Sci. USA*, 92(21):9657–9661, 1995. doi: 10.1073/pnas.92.21.9657.

[75] S. N. Evans and D. Steinsaltz. Estimating some features of NK fitness landscapes. *Ann. Appl. Prob.*, 12:1299–1321, 2002.

[76] R. Durrett and V. Limic. Rigorous results for the NK model. *Ann. Prob.*, 31:1713–1753, 2003.

[77] V. Limic and R. Pemantle. More rigorous results on the Kauffman-Levin model of evolution. *Ann. Prob.*, 32:2149–1287, 2004.

[78] B. Schmiegelt and J. Krug. Evolutionary accessibility of modular fitness landscapes. *J. Stat. Phys.*, 154(1-2):334–355, 2013.

[79] E. Weinberger. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biol. Cybern.*, 63:325–336, 1990.

[80] P. F. Stadler and R. Happel. Random field models for fitness landscapes. *J. Math. Biol.*, 38:435–278, 1999.

[81] E. J. Gumbel and H. von Schelling. The distribution of the number of exceedences. *Ann. Math. Stat.*, 21:247–262, 1950.

[82] D. R. Rokyta, C. J. Beisel, and P. Joyce. Properties of adaptive walks on uncorrelated landscapes under strong selection and weak mutation. *J. Theor. Biol.*, 243:114–120, 2006.

[83] J. Neidhart. Adaptive walks in random and correlated fitness landscapes. Master's thesis, Universität zu Köln, 2011.

[84] H. A. David and H. N. Nagaraja. *Order Statistics*. Wiley-Interscience, 2003.

[85] M. Kac. Can one hear the shape of a drum? *Am. Math. Mon.*, 73: 1–23, 1966.

[86] J. Neidhart, I. G. Szendro, and J. Krug. Exact results for amplitude spectra of fitness landscapes. *J. Theor. Biol.*, 332:2018–227, 2013.

[87] E. D. Weinberger. Fourier and Taylor series on fitness landscapes. *Biol. Cybern.*, 65:321–330, 1991.

[88] P. F. Stadler and R. Happel. Random field models for fitness landscapes. *J. Math. Biol.*, 38:435–478, 1996.

[89] P. F. Stadler. Landscapes and their correlation function. *J. Math. Chem.*, 20, 1996.

[90] G. Szegö. Orthogonal polynomials. *Amer. Math. Soc. Colloq. Publ.*, 23, 1975.

[91] T. Stoll. *Reconstruction problems for graphs, Krawtchouk polynomials, and Diophantine equations*. Birkhäuser Boston, 2011.

[92] R. Coleman. On Krawtchouk polynomials. *CoRR*, abs/1101.1798, 2011.

[93] P.R.A. Campos, C. Adami, and C.O. Wielke. Optimal adaptive performance and delocalization in NK fitness landscapes. *Physica A*, 304:495–506, 2002.

[94] W. Fontana, P. F. Stadler, E. G. Bornberg-Bauer, T. Griesmacher, I. L. Hofacker, M. Tacker, P. Tarazona, E. D. Weinberger, and P. Schuster. RNA folding and combinatory landscapes. *Phys. Rev. E*, 47(3):2083–2099, 1993.

[95] H. W. Gould. Additional combinatorial identities, 2010. URL `http://www.math.wvu.edu/~gould/`. edited by Jocelyn Quaintance, accessed: 27.8.2014.

[96] B. Drossel. Biological evolution and statistical physics. *Adv. Phys.*, 50: 209–295, 2001.

[97] H. A. Orr. A minimum on the mean number of steps taken in adaptive walks. *J. theor. Biol.*, 220:241–247, 2002.

[98] H. A. Orr. The population genetics of adaption: the adaption of DNA sequences. *Evolution*, 56(7):1317–1330, 2002.

[99] J. Neidhart and J. Krug. Adaptive walks and extreme value theory. *Phys. Rev. Lett.*, 107, 178102, 2011.

[100] K. Jain and S. Seetharaman. Multiple adaptive substitutions during evolution in novel environments. *Genetics*, 189(3):1029–1043, 2011.

[101] J. Krug and C. Karl. Punctuated evolution for the quasispecies model. *Physica A*, 318:137, 2003.

[102] C. Sire, S. Majumdar, and D. S. Dean. Exact solution of a model of time-dependent evolutionary dynamics in a rugged fitness landscape. *J. Stat. Mech.: Theory Exp.*, L07001, 2006.

[103] I. Bena and S. N. Majumdar. Universal extremal statistics in a freely expanding Jepsen gas. *PRE*, 75:051103, 2008.

[104] K. Jain. Number of adaptive steps to a local fitness peak. *EPL*, 96 (58006), 2011.

[105] P. Joyce, D. Rokyta, C. Beisel, and H. A. Orr. A general extreme value theory model for the adaptation of DNA sequences under strong selection and weak mutation. *Genetics*, 180:1627–1643, 2008.

[106] P. Hegarty and A. Martinsson. On the existence of accessible paths in various models of fitness landscapes. *Ann. Appl. Probab.*, 24(4): 1375–1395, 2014.

[107] S.-C. Park, J. Neidhart, and J. Krug. Greedy adaptive walks on a correlated fitness landscape. *in preparation*, 2014.

[108] T. Ernst. A method for q-calculus. *J. Nonlinear Math. Phy.*, 10(4): 487–525, 2003.

[109] S.-C. Park, I. G. Szendro, J. Neidhart, and J. Krug. Phase transition in random adaptive walks on correlated fitness landscapes. *PRL (submitted)*, 2014.

[110] F. A. Kondrashov and A. S. Kondrashov AS. Multidimensional epistasis and the disadvantage of sex. *Proc. Natl. Acad. Sci. USA*, 98:12089–12092, 2001.

[111] R. A. Watson, D. M. Weinreich, and J. Wakeley. Genome structure and the benefit of sex. *Evolution*, 65:523–536, 2011.

[112] R. A. Watson and J. Wakeley. Multidimensional epistasis and the advantage of sex. *IEEE C. Evol. Computat.*, 3:2792–2799, 2005.

[113] J. A. G. M. de Visser, S. C. Park, and J. Krug. Exploring the effect of sex on empirical fitness landscapes. *Amer. Nat.*, 174:S15–S30, 2009.

[114] D. Misevic, R. D. Kouyos, and S. Bonhoeffer. Predicting the evolution of sex on complex fitness landscapes. *PLoS Comput. Biol.*, 5(9): e1000510, 2009.

[115] D. Moradigaravand and J. Engelstädter. The effect of bacterial recombination on adaptation on fitness landscapes with limited peak accessibility. *PLoS Comput. Biol.*, 8(10):e1002735., 2012.

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

[117] T. Nagylaki. The evolution of multilocus systems under weak selection. *Genetics*, 134:627–647, 1993.

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

[119] N. Nei. Modification of linkage intensity by natural selection. *Genetics*, 57:625–641, 1967.

[120] M. W. Feldman and U. Liberman. An evolutionary reduction principle for genetic modifiers. *Proc. Nat. Acad. Sci. USA*, 83:4824–4827, 1986.

[121] D. B. Weissman, M. W. Feldman, and D. S. Fisher. The rate of fitness-valley crossing in sexual populations. *Genetics*, 186:1389–1410, 2010.

[122] A. Altland, A. Fischer, J. Krug, and I. G. Szendro. Rare events in population genetics: stochastic tunneling in a two-locus model with recombination. *Phys. Rev. Lett.*, 106:088101, 2011.

[123] K. Jain. Time to fixation in the presence of recombination. *Theor. Popul. Biol.*, 77:23–31, 2010.

[124] S. C. Park and J. Krug. Bistability in two-locus models with selection, mutation, and recombination. *J. Math. Biol.*, 62:763–788, 2011.

[125] R. A. Neher, M. Vucelja, M. Mezard, and B. I. Shraiman. Emergence of clones in sexual populations. *J. Stat. Mech.*, P01008, 2013.

[126] D. B. Weissman, M. M. Desai, D. S. Fisher, and M. W. Feldman. The rate at which asexual populations cross fitness valleys. *Theor. Popul. Biol.*, 75:286–300, 2009.

[127] L. Becks and A. F. Agrawal. The evolution of sex is favoured during adaptation to new environments. *PLoS Biol.*, 10(5):e1001317, 2012.

[128] A. Hastings. Transients: the key to long-term ecological understanding? *Trends Ecol. Evol.*, 19:39–45, 2004.

[129] V. Mustonen and M. Lässig. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends Genet.*, 25:111–119, 2009.

[130] J. Maynard Smith. Selection for recombination in a polygenic model - the mechanism. *Genet. Res.*, 51:59–63, 1988.

[131] B. Charlesworth. Directional selection and the evolution of sex and recombination. *Genet. Res., Cambridge*, 61:205–224, 1993.

[132] M. Turelli and N. H. Barton. Dynamics of polygenic characters under selection. *Theor. Popul. Biol.*, 38:1–57, 1990.

[133] R. Bürger. *The Mathematical Theory of Selection, Recombination, and Mutation*. John Wiley & Sons, Ltd, 2000.

[134] L. Carroll. *Through the looking glass and what Alice found there*. Harper & Brothers Publishers, New York and London, 1902.

[135] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, and F. Rossi. *GNU Scientific Library Reference Manual*. Network Theory Ltd., 3rd edition, 2009. URL `https://www.gnu.org/software/gsl/`.

[136] M. Matsumoto and T. Nishimura. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM T. Model. Comp. S.*, 8(1):3–30, 1998.

Appendices

Appendix A

Various definitions and remarks

For convenience it might be helpful to reintroduce some standard notations to avoid confusion. The harmonic numbers are defined by the sum

$$\begin{aligned} H_n &= \sum_{k=1}^{n} \frac{1}{k} \\ H_0 &= 0 \end{aligned} \tag{A.1}$$

and they are connected to the logarithm in the limit of large n by the Euler-Mascheroni constant $\gamma \approx 0.577$

$$\begin{aligned} H_n - \log(n) \to \gamma \ (n \to \infty) \\ H_n = \log(n) + \gamma + \mathcal{O}\left(\frac{1}{n}\right). \end{aligned}$$

The Gamma function is a generalization of the factorial $n! = 1 \cdot 2 \cdot \cdots \cdot n$ to the non-natural numbers and is defined by the properties

$$\begin{aligned} \forall_{n\in\mathbb{N}} : \Gamma(n+1) = n! \\ \text{and } \forall_{x\in\mathbb{R}} : \Gamma(x+1) = x\Gamma(x). \end{aligned} \tag{A.2}$$

The Beta-function is defined by

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}\mathrm{d}t\ . \tag{A.3}$$

Since at some points the binomial coefficients might be used ambiguously, it is useful to reintroduce them:

$$\binom{L}{k} = \begin{cases} \frac{L!}{k!(L-k)!}, & L \geq k \text{ and } L, k \geq 0, \\ 0, & \text{else.} \end{cases} \tag{A.4}$$

Probability distribution functions will usually be denoted by P and the corresponding density by p. The exponential distribution with mean λ and variance λ^2 has a density

$$p(x) = \begin{cases} \frac{1}{\lambda} e^{-\frac{1}{\lambda}x} & x \geq 0, \\ 0 & x < 0 \end{cases} \tag{A.5}$$

and distribution function

$$P(x) = \begin{cases} 1 - e^{-\frac{1}{\lambda}x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

The Weibull distribution with mean $\Gamma(1 + 1/\beta)$ and variance $\Gamma\left(1 + \frac{2}{\beta}\right) - \left(\Gamma\left(1 + \frac{1}{\beta}\right)\right)^2$ has a density

$$p(x) = \begin{cases} \beta x^{\beta-1} e^{-x^\beta} & x \geq 0, \\ 0 & x < 0, \end{cases} \tag{A.6}$$

and a distribution function

$$P(x) = 1 - e^{-x^\beta}.$$

The Kumaraswamy distribution with mean $\frac{\Gamma(1+\frac{1}{\alpha})}{\Gamma(2+\frac{1}{\alpha})}$ and support $[0, 1]$ has a density

$$p(x) = \alpha x^{\alpha-1}. \tag{A.7}$$

and distribution function

$$P(x) = 1 - (1 - x^\alpha).$$

The Pareto distribution with mean $\frac{\alpha}{\alpha-1}$ for $\alpha > 1$, variance $\frac{\alpha}{(\alpha-1)^2(\alpha-2)}$ for $\alpha > 2$ and support $[1, \infty)$ has a density

$$p(x) = \alpha x^{-(\alpha+1)} \tag{A.8}$$

and a distribution function

$$P(x) = 1 - x^{-\alpha}.$$

Appendix B

On the algorithms used in simulations

In this Appendix, the algorithms used for the simulations are presented but for the ones in sec. 5.2 which are given directly there. All programs are written in C++ and use the GNU Scientific Library [135], especially its implementation of the Mersenne Twister [136] pseudorandom number generator. In the following the main idea of the algorithms is given in arranged prose to avoid confusion of ambiguous definitions from pseudocode or distract by unnecessary syntax from the original program code. Although not quite accurate, for convenience pseudorandom numbers will be referred to as random numbers. In most simulations, the code creates the neighborhoods on-the-fly which leads to an error of $\mathcal{O}(\frac{1}{L})$ but enables the simulation of larger landscape dimensions. Usually only the main part of the simulation is described and has to be repeated for statistics.

B.1 Single adaptive steps on RMF-landscapes

The simulations for sec. 4.3.1 used the following algorithm.

- Draw $L + 1$ GPD random numbers
- Create an RMF neighborhood with given d by adding c to d random numbers, substract c from $L - d$ and leave one unaltered, call it F_0.
- If F_0 is a fitness maximum repeat the previous step.
- Choose a fitness value $F' > F_0$ at random and make a step with probability $(F' - F_0)/(F_{max} - F_0)$ where F_{max} is the maximal fitness value from the neighborhood.

- If the step is taken proceed, else repeat the previous step.
- Create a new RMF neighborhood with d altered according to the direction F' is with respect to F_0.
- Measure the rank and return the NoE.

B.2 Fitting an RMF-model with the NoE

In sec. 2.3, to fit an RMF-model to experimental data with help of the NoE, a neighborhood change was simulated. Then a parameter scan was performed and for $\kappa \in [a, b]$ a c was found, such that the landscape has a typical NoE which equals the one measured in experiments. The simulation starts with $\kappa = a$ and $c = 0$ while d has to be set by hand. Then the following algorithm returns the parameters of interest.

1. Repeat the following for statistics
   - Draw $L + 1$ GPD random numbers
   - Create an RMF neighborhood with given d by adding c to d random numbers, substract c from $L - d$ and leave one unaltered.
   - Sort the RMF fitness values and pick the fitness with the desired rank. If this fitness value is the one which is neither uphill nor downhill, repeat step 1 because then no step is taken. The chosen fitness value is the one, the step will go to.
   - Redraw the L remaining fitness values of the new RMF neighborhood and alter d to $d + 1$ if the step is taken downhill and to $d - 1$ if the step was taken uphill.
   - Measure the new rank.
2. Compare the mean rank after step with the experimental result, if the simulated NoE is lower, increase c and repeat step 1. If it is higher, save the parameters and go to step 3.
3. If $\kappa < b$ increase κ and repeat step 1.

B.3 NAW on an RMF-landscape

Simulations in sec. 4.3.2 used the following algorithm.

- Draw $L + 1$ GPD random numbers.

- Create an RMF neighborhood with given d by adding c to d random numbers, substract c from $L-d$ and leave one unaltered, this will be the momentary fitness F_0. Make sure that F_0 has the desired starting rank.

  1. Choose a fitness value $F' > F_0$ at random and make a step with probability $(F'-F_0)/(F_{max}-F_0)$ where F_{max} is the maximal fitness value from the neighborhood.
  2. If the step is taken proceed, else repeat step 1.
  3. Create a new RMF neighborhood with d altered according to the direction F' is with respect to F_0.
  4. Set $F_0 = F'$. If F_0 is a fitness maximum return the number of taken steps, else go to step 1.

B.4 GAW in high dimensions

Simulations in sec. 4.3.3 required a large sequence length, therefore the algorithm for adaptive walks had to be altered. To produce random numbers according to the distribution of the largest of n random numbers which are drawn from a distribution with distribution function $P(x)$ one can do the following. Given a random number generator ξ which produces uniformly distributed random numbers in $(0,1)$, the Ansatz $P(x)^n = \xi$ can be solved for $x(\xi)$ which will have the desired distribution, generated from ξ. This way in each step only comparably few random numbers have to be generated and large sequence lengths can be simulated. In the dRMF case, each step requires only one new random number and the following algorithm was used.

- Set L and start with the number of steps $l = 0$. Draw the starting fitness F_0 from the desired distribution.

- 1. Draw a random number from the distribution of the largest of $L-l$ random numbers and call it h.
  2. If $h + c > F_0$ take the step by setting $F_0 = h$, increase l by one and go to step 1. Else return the number of steps taken.

If the dRMF-model was not sufficient, e.g. in the study of small α, two new random numbers have to be generated each step. The following algorithm was used.

- Set L and d as desired and start with the number of steps $l = 0$. Draw the starting fitness F_0 from the desired distribution.

- 1. Draw a random number from the distribution of the largest of $L - d$ random numbers, call it h, for the downhill and from the distribution of the largest of d random numbers, call it g, for uphill neighborhood.
  2. If $\max(g+c, h-c) > F_0$ take the step uphill (downhill) by setting $F_0 = g$ ($F_0 = h$) if $\max(g+c, h-c) = g+c$ ($\max(g+c, h-c) = g+c$), alter d accordingly and increase l, afterwards go to step 1. Else return the number of steps taken.

B.5 RAW in high dimensions

For study of the phase transition in sec. 4.3.4, long RAWs had to be simulated. The following algorithm was used for walks on the dRMF.

- Set L and start with the number of steps $l = 0$. Draw the starting fitness F_0 from the desired distribution.

- 1. Create a counting variable $i = 1$, which counts the number of neighbors which have been seen.
  2. Check if a local maximum is reached by generating a random number $u \in (0, 1)$ and compare it to the probability that F_0 is a fitness maximum $P(F_0 - c)^{L-l}$. If u is smaller, a maximum is found, return the number of steps, else proceed.
  3. Draw a random number from the distribution of the RMF-model and call it h.
  4. If $h + c > F_0$ take the step by setting $F_0 = h$, increasing l by one and go to step 1. Else proceed to step 4.
  5. If $i \leq L$ increase i by one and go to step 3. Else all neighbors have been seen and have lower fitness, return the number of steps taken.

For the walks with back-steps the following algorithm was used.

- Set L and d. Draw the starting fitness F_0 from the desired distribution.

- Create a variable $s \in \{+1, -1\}$.

- 1. Set $d' := d$ and $d'' := L - d$ which are the neighbors up (d'') and down (d') which have not been seen yet.

2. Draw a random number in the range $[1, d' + d'']$ to cast a neighbor from up- or downhill. If it is smaller or equal to d', a step will be made in direction of σ^* and $s := -1$. Otherwise $s := +1$. Decrease d' (d'') by one if $s = -1$ ($s = +1$).
3. Draw a random number from the distribution of the RMF-model and call it h.
4. If $h + s \cdot c > F_0$ take the step by setting $F_0 = h$. Set $d := d + s$ and go to step 1. Else proceed to step 5.
5. If $d' > 0$ or $d'' > 0$ go to step 2. Else return the number of steps taken because all neighbors have been seen and no higher fitness value was found.

Anhang C

Teilpublikationen & Erklärung

- Johannes Neidhart, Ivan G. Szendro and Joachim Krug, Exact Results for Amplitude Spectra of Fitness Landscapes. *J. Theor. Biol.*, **32**:2018–227, (2013).
- Johannes Neidhart, Ivan G. Szendro and Joachim Krug, Adaptation in tunably rugged fitness landscapes: The Rough Mount Fuji Model, *Genetics*, doi:10.1534/genetics.114.167668, (2014).
- S. Nowak, J. Neidhart, I. G. Szendro and J. Krug, Multidimensional epistasis and the transitory advantage of sex, *PLoS Comp. Biol.*, **10**(9):e1003836, (2014).
- Su-Chan Park, Ivan G. Szendro, Johannes Neidhart and Joachim Krug, Phase transition in random adaptive walks on correlated fitness landscapes. *In review*, (2014).
- Su-Chan Park, Johannes Neidhart and Joachim Krug, Greedy adaptive walks on the Rough Mount-Fuji Fitness Landscapes. *In preparation*, (2014).

Ich versichere, dass ich die von mir vorgelegte Dissertation selbständig angefertigt, die benutzten Quellen und Hilfsmittel vollständig angegeben und die Stellen der Arbeit einschließlich Tabellen, Karten und Abbildungen, die anderen Werken im Wortlaut oder dem Sinn nach entnommen sind, in jedem Einzelfall als Entlehnung kenntlich gemacht habe; dass diese Dissertation noch keiner anderen Fakultät oder Universität zur Prüfung vorgelegen hat; dass sie abgesehen von oben angegebenen Teilpublikationen noch nicht veröffentlicht worden ist, sowie, dass ich eine solche Veröffentlichung vor Abschluss des Promotionsverfahrens nicht vornehmen werde. Die Bestimmungen der Promotionsordnung sind mir bekannt. Die von mir vorgelegte Dissertation ist von Prof. Dr. Joachim Krug betreut worden.

Köln, den 12. Dezember 2014

J. Neidhart