Statistical topography of fitness landscapes

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

Vorgelegt von

Jasper Franke

aus Hamburg

Berichterstatter: Prof. Dr. Joachim Krug, Universität zu Köln
Prof. Dr. Heiko Rieger, Universität des Saarlandes

Datum der mündlichen Prüfung: 30.1.2012

Zusammenfassung

Fitnesslandschaften sind verallgemeinerte Energielandschaften, die eine wichtige konzeptionelle Rolle in der Evolutionsbiologie spielen. Diese Landschaften stellen eine Beziehung her zwischen der genetischen Konfiguration eines Organismus und seinen adaptiven Eigenschaften.
In dieser Arbeit werden globale topographische Eigenschaften dieser Fitnesslandschaften an theoretischen Modellen untersucht und die daraus resultierenden Ergebnisse mit empirischen Landschaften verglichen. Dabei wird gezeigt, daß diese Landschaften zumindest mit Hinblick auf die betrachteten Eigenschaften grob in zwei Klassen eingeteilt werden können, deren Verhalten sich im biologisch Relevanten Grenzfall stark unterscheidet. Obwohl für empirische Landschaften nicht allgemein gezeigt werden kann, in welche dieser Klassen sie fallen, werden einige der bekannten empirischen Datensätze im Rahmen dieser Arbeit von Fall zu Fall klassifiziert.
Theoretische Modellstudien führen zur Betrachtung von Sequenzen von Rekordereignissen unabhängiger, nicht identisch verteilter Zufallsvariablen und der Korrelationen dieser Rekordereignisse untereinander. Diese Untersuchungen stehen zunächst zum Großteil außerhalb der Hauptrichtung dieser Arbeit, führen jedoch zu einer rekordbasierten Methode der Datenauswertung, die auf einen der hier betrachteten Datensätze angewandt wird.

Abstract

Fitness landscapes are generalized energy landscapes that play an important conceptual role in evolutionary biology. These landscapes provide a relation between the genetic configuration of an organism and that organism's adaptive properties.
In this work, global topographical features of these fitness landscapes are investigated using theoretical models. The resulting predictions are compared to empirical landscapes. It is shown that these landscapes allow, at least with respect to the properties considered, for a rough classification into two types. In the biologically relevant limit, these two types of landscapes show very different behavior. While empirical landscapes cannot be classified on purely theoretical grounds, some of the known landscapes are classified here on a case-by-case basis.
The study of theoretical models leads to consider sequences of record events of independent but non-identically distributed random variables and correlations between these record events. While these considerations appear at first to be outside the main direction of this thesis, they lead to a record-base tool of data analysis that is applied to one of the data sets considered here.

Contents

1. Evolutionary Biology and Population Genetics

In this chapter, the basics ideas and some key results of evolutionary biology and population genetics necessary for the remainder of this thesis are introduced. Some basic elements of the biochemical mechanisms of heredity are recalled and finally the basic questions to be adressed within this work are stated.

1.1. Evolutionary Biology

In his famous book 'On the Origin of Species by Means of Natural Selection' [28], Charles Darwin in 1859 proposed an explanation for the diversity and the spacial distribution of species he had observed during his voyage on the 'Beagle' from 1831 through 1836. These explanations, which laid the foundations of the Theory of Evolution, rely on three separate principles:

Heredity: Each individual in a given generation carries hereditary material, the individuals 'genotype', which is passed down to its offspring in the next generation (In the case of sexual reproduction, the individual's genotype is 'mixed' with that of its mate, thus only part of the hereditary material of each parent gets passed onto the offspring).

Mutation: In the reproductive process, 'copying errors' called 'mutations' occur at a very low rate. Thus the parental genotype is not passed down in a perfectly conserved manner. This way, genetic diversity occurs.

Natural Selection: The individuals carrying these mutant genotypes may have different properties (reproduction rate, fertility, metabolism, resistance to environmental influences such as heat, cold or radiation, etc.) than their parents or other individuals present in the same generation. If these properties are better suited to the environmental conditions, the organism has an advantage. Thus it is expected to have more offspring than an organism without this advantage, leading to a higher percentage of organisms with this advantage in the next generation. This selection mechanism is termed 'Natural Selection'[1]. In allusion to Herbert Spencer calling this principle of natural selection 'survival of the fittest' [138], it is said that such a better adapted mutant is 'fitter' than the others, with 'fitness' generically referring to the degree of adaptation to the environment.

If these three principles act repeatedly on a population for many generations, the resulting population is expected to be better adapted to the surroundings. Darwin proposed that over

[1] As opposed to 'Artificial Selection' in e.g. animal or plant breeding, where humans select those individuals of the population that most strongly possess a desired property as the only one to have offspring.

long (comparable to geological) time scales and with sub-populations of the same species occupying different habitats, this mechanism shapes the form and diversity of life observed.
Darwin's idea was quite controversial at the time for several reasons. Most of the scientific criticism founded to a large extend on the work of Gregor Mendel [98], who found that hereditary information is transmitted through the generations in 'blocks'. This seemed to contradict the Darwinian picture, within which the population adapted to the environmental surroundings gradually and in small steps. An account of this controversy can for example be found in the book by Ewens [45].
This controversy was only resolved in the 1930's by the works of Fisher, Haldane and Wright who described what is today known as the 'Modern Synthesis' [72], thereby founding the field of Theoretical Population Genetics.

1.2. Theoretical Population Genetics

In his seminal work of 1866 [98], Mendel systematically studied the result of mating between plants of the same species, but different appearance (or different 'phenotypic traits'), e.g. color or shape of the seed. Denoting one variant or 'allele' of a given trait by A and another by a (if only two alleles are present), he observed that besides the two variants A and a, the next generation contained also a certain proportion of intermediate phenotypes that Mendel denoted Aa. The fact that such mixed states Aa exist but no higher degree of mixing such as Aaa or even a continuum of states between A and a implies that each individual carries *two* copies of the 'block of genetic information'[2] considered, one from the mother and one from the father. Such organisms are called 'diploid' as opposed to 'haploid' organisms which only carry one copy. Both types of organisms exist in nature. An assumption consistent[3] with the findings of Mendel is that each parent in sexual reproduction contributes one of its copies to the offspring. The 'pure' individuals carry two identical copies of the genetic information A and a respectively (denoted by AA or aa) and are called 'homozygotes' , while the intermediate type Aa is called 'heterozygote'.
If X individuals in the population are in the AA-state, $2Y$ individuals in the Aa state (where the Aa and aA are counted as equivalent and denoted as Aa) and Z in the state aa, one can define the 'frequencies' of each state by dividing each of these numbers by the total population size N,

$$x \equiv \frac{X}{N}, \qquad 2y \equiv \frac{2Y}{N} \qquad \text{and} \qquad z \equiv \frac{Z}{N}.$$

Evidently, if one organism is randomly picked without replacement from the population, the probability that it is in, say, the AA state is equal to that state's frequency. If in a given population, the allele frequency for a given trait is not unity, i.e. there are two or more alleles coexisting, the population is said to be genetically diverse.
Every population observed in nature shows some degree of genetic diversity [45,58]. One of the central purposes of the field of population genetics is to quantitatively describe these variations. Population genetics can be formulated independently of the idea of evolution in the Darwinian sense, yet it is intimately connected to the study of evolution. This can be seen by considering

[2] later identified as chromosomes
[3] and later directly verified

one of the early landmark results of population genetics, the Hardy-Weinberg law ([66, 142], see also [45]). This law describes how a population of sexually reproducing diploid organisms can maintain a steady level of genetic diversity simply by the combinatorics of randomly mating two of its individuals. Furthermore, by the same combinatorics, the Hardy-Weinberg law makes predictions about these steady-state frequencies which are observed in large classes of organisms ranging from model systems like *Drosophila melanogaster* to humans, see [45, 58] and references therein.
The fact that diversity is maintained in a population is important for that population's evolution. It is only on this diversity that selection in the Darwinian sense can act, showing that the Mendelian mechanism of inheritance and the gradual picture of Darwinian evolution are not mutually exclusive but strongly related to the point where they are complementary [45, 108], thus resolving this apparent contradiction.
However, the Hardy-Weinberg law only applies to sexually reproducing haploid populations while for the remainder of this thesis, focus will be on asexually reproducing haploid organisms.

1.2.1. A Master equation for frequencies

In stating the Hardy-Weinberg law, the notion of allele frequencies was introduced along with their interpretations as the probability that a randomly selected individual has this allele. This law was formulated solely on the basis of heredity, the first of the three principles on which the Darwinian theory of evolution relies. The others can be included using the probabilistic interpretation of allele frequencies.
To include selection, assume that with a small probability, the offspring carries genetic information different from that of its parent. Now if the copying step in reproduction goes wrong and the genetic information A is not perfectly transmitted to the offspring and instead, the offspring receives information a, say[4], the frequency of a will be $1/N$, assuming that it is the first time that this mutation a has come about. In principle, if the a mutant is fitter than the original variety A - also referred to as the 'wild type' (WT) since it is the variety without any mutation 'as found in the wild' - the a frequency should increase and approach unity with each generation. However, if an accident happens, e.g. the new mutant is eaten by a predator or similar events which cannot possibly be taken into account theoretically, the possibly very advantageous mutation a is lost to the population.
Both the process by which mutations arise and the process by which single individuals die are very complicated and even if understanding in one single instance were possible, this understanding would in general not carry over to another instance. Thus a major theoretical step towards understanding the change in frequencies in a *typical* population is to introduce a stochastic equation for the time development of trait frequencies. This endeavor, pioneered by Fisher in in 1922 [47] and later extended by him [48], Haldane [64] and Wright [153] (see also [80] for a review) led to the formulation of what is now called 'the modern synthesis' of evolutionary biology. One of its core messages is that for one single individual, its ultimate fate cannot be determined with certainty: Only *probabilities* for a given outcome can be computed. In what follows, a short overview over the mathematical formalism necessary for such

[4]Note that there are many of these errors possible such that the next copying error will most likely *not* yield an offspring a but yet another with information a'

an approach will be given.

A recursive integral equation Assume a population homogeneously carrying information A within which a mutation a has just arisen. Of course, it now has frequency $x_a = 1/N$ and, since for the moment the possibility that another mutation arises is ignored, one can drop the index on the frequency and only deal with the frequency of the mutant x and the frequency of the wild type $1 - x$. One assumption that will turn out to make computations considerably more easy is that the population size N is so large that the minimal change in frequency, $\delta x = 1/N \ll 1$. This will allow to treat *difference* equations like *differential* equations but does not yet constitute the assumption of infinite population size. Let $\phi(x, t|x_0, t_0 = 0)$ denote the probability density that the mutation has reached frequency x at time t after it first arose with frequency x_0 at time $t_0 = 0$. Another reasonable assumption is that changes in the frequency in the time step $t \to t + \delta t$ only depend on the frequency at time t. The probability density for taking a step of size δx in time δt is expressed by $g(\delta x, \delta t|x - \delta x, t)$. This property, known as Markov property [140], allows immediately to set up a recursion equation for ϕ in t,

$$\phi(x, t + \delta t|x_0, t_0) = \int \phi(x - \delta x, t|x_0, t_0) g(\delta x, \delta t|x - \delta x, t) d(\delta x). \tag{1.1}$$

This equation, known as Chapman-Kolmogorov equation - in simple forms used as early as 1900 by Bachelier [8] and later by Einstein in 1905 [41] but only developed to full theory by Kolmogorov and Chapman, see e.g. [140] and references therein - can be paraphrased as follows: The probability $\phi(x, t + \delta t|x_0, t_0 = 0)$ of going from frequency x_0 to frequency x in time $t + \delta t$ can be expressed as the probability $\phi(x - \delta x, t|x_0, t_0 = 0)$ of going from x_0 to some state x in time t and subsequently taking the remaining step δx in time δt and then summing (integrating) over all possible intermediate states x.

Deriving the corresponding differential equation Now under very general mathematical conditions, the function $\phi(x, t|x_0, t_0) g(\delta x, \delta t|x - \delta x, t)$ can be expressed as a Taylor series around x in δx using the short notation $y \equiv x - \delta x$ to emphasize that the derivative is with respect to the spacial argument

$$\phi(y, t|x_0, t_0) g(\delta x, \delta t|y, t) = \sum_{n=0}^{\infty} \frac{(-\delta x)^n}{n!} \frac{d^n}{dy^n} [\phi(y, t|x_0, t_0) g(\delta x, \delta|y, t)] \Big|_{y=x} \tag{1.2}$$

Now subtracting the zeroth order term on both sides, dividing by the time step δt and letting the time step δt go to zero, the left hand side of eq (1.1) becomes the time derivative of ϕ. On the right hand side, more care is needed because dividing by something that subsequently goes to zero might pose some problems. However if it can be assumed that the frequency x is a continuous stochastic process in time, the frequency change δx in time δt also tends to zero with the time step. This allows not only to divide by δt but also to ignore terms in $(\delta x)^3$ and smaller on the right hand side. Thus, if integration, summation and differentiation can be interchanged and with the definition

$$M(x,t) \equiv \lim_{\delta t \to 0} \frac{1}{\delta t} \int \delta x g(\delta x, \delta t | x, t) d(\delta x) \tag{1.3}$$

$$V(x,t) \equiv \lim_{\delta t \to 0} \frac{1}{\delta t} \int (\delta x)^2 g(\delta x, \delta t | x, t) d(\delta x), \tag{1.4}$$

$\phi(x, t|x_0, t_0)$ obeys the differential equation

$$\frac{\partial}{\partial t}\phi(x,t|x_0,t_0) = -\frac{\partial}{\partial x}\left\{M(x,t)\phi(x,t|x_0,t_0)\right\} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\left\{V(x,t)\phi(x,t|x_0,t_0)\right\}. \tag{1.5}$$

In the physics literature, eq (1.5) is known as Fokker-Planck equation and the method by which it was obtained from the Chapman-Kolmogorov equation is also called Kramers-Moyal expansion [140].

To find the first and second moment, consider the microscopical behavior of growth of a population of $N_a(t)$ individuals,

$$\frac{\mathrm{d}}{\mathrm{d}t}N_a(t) = F_a N_a(t) + \mu_{A\to a}(N - N_a(t)) - \mu_{a\to A}N_A(t) + \eta_a(t), \tag{1.6}$$

where F_a is the growth rate of mutant a, $\mu_{a\to A}$ and $\mu_{A\to a}$ the mutation rates from a to A and A to a respectively and $\eta_a(t)$ is a symmetric random variable encoding any influences that cannot be accounted for. From eq (1.6), a Langevin equation for the change of *frequency* $x(t)$ of the new mutant a can be derived which in turn gives the first and second moments from eq (1.3) to finally yield the Kimura equation ([78–80, 82] and [103] for this presentation)

$$\begin{aligned}\frac{\partial}{\partial t}\phi(x,t|x_0,t_0) =& \frac{1}{2N}\frac{\partial^2}{\partial x^2}\left[x(1-x)\phi(x,t|x_0,t_0)\right] - s\frac{\partial}{\partial x}\left[x(1-x)\phi(x,t|x_0,t_0)\right] \\ &+ \mu_{A\to a}\frac{\partial}{\partial x}\left[(1-x)\phi(x,t|x_0,t_0)\right] - \mu_{a\to A}\frac{\partial}{\partial x}\left[x\phi(x,t|x_0,t_0)\right]\end{aligned} \tag{1.7}$$

where $s = F_a - F_A$ is the difference in growth speed between the WT A and the mutant a. This difference is also called 'fitness'. Note that it is only defined with respect to the WT (or in general to some reference type).

1.2.2. Fixation probability

One would clearly expect that a mutation that has newly arisen will fix in the population if it is fitter than the WT, i.e. if $s > 0$. This can be computed from the solution $\phi(x, t|x_0 = 1/N, t_0)$ by sending $x \to 1$, as this represents the scenario of fixation. This fixation probability $u(N, s)$ was found by Kimura as [79]

$$u(N,s) = \frac{1 - e^{-2s}}{1 - e^{-4Ns}} \approx \frac{2s}{1 - e^{-4Ns}} \tag{1.8}$$

where the approximation is only valid for small $|s|$. The associated time, called 'fixation time' depends on the population size.

In the Kimura equation (1.7), the mutation rates appear as constants $\mu_{a\to A}$ and $\mu_{A\to a}$. However, in order to understand their meaning, basic notions of the biochemical form in which the hereditary information is stored are needed.

1.3. Elements of molecular biology

At the time when Darwin formulated his ideas on evolution and Mendel discovered the laws of genetics, the form in which the hereditary material is passed on from parent to offspring was unknown. When these two idea were reconciled in the modern synthesis in the 1920s and 1930s, this was still the case. Nonetheless it was possible to formulate the ideas present this far only by using abstract notions of genetic information[5]. However in connection with the stochastic master equation (1.7) for the fate of a single mutation, the Kimura equation, it was shown that in order to take mutations into account properly, further knowledge of the mutation mechanism is in order. The mutations are essentially errors in copying the genetic information, and to understand them better, some basic notions of the structure of the hereditary material are needed and will be provided in this section. The presentation here follows the standard textbook [5] and is only intended as a very basic summary of the principal mechanism.

1.3.1. The double helix

The genetic information of almost any living organism is stored in the same form, as a double-stranded molecule called Desoxyribonucleic Acid (DNA). The molecule is made up of two parallel backbones consisting of sugar-phosphate connected by hydrogen bonded base pairs. There are four of these bases, adenine (A), guanine (G), cytosine (C) and thymine (T). T preferably binds to A and and C to G. In thermal equilibrium, the double stranded DNA molecule is twisted, thus forming the famous double helix structure.
Each cell contains the complete genetic information for the whole organism, sometimes distributed across units called 'chromosomes'. The chromosomes contain many 'genes', which are stretches of DNA that code for one protein. Primitive organisms such as bacteria or Archeae have about 1000 genes, while higher organisms like mammals have many more [5].
When a cell reproduces, the double-stranded DNA splits up to form two single-stranded molecules which act as templates for the synthesis of *two* new double-stranded DNA molecules. Thus the information is copied during cell reproduction. In order to synthesize the complementary strand, it is not sufficient for the single stranded DNA to be in a medium with an excess of G, A, T and C. The hydrogen bonds between the bases are not strong enough for spontaneous synthesis, thus a helper enzyme is needed, the polymerase. Working its ways along the sugar-phosphate backbone, it 'captures' the complementary base for each site and catalyzes the binding reaction [5], see fig 1.1. However, the polymerase does not spontaneously attach to the single-stranded DNA molecule. Rather, it needs part of the double-stranded DNA

[5]It is quite a remarkable success of the stochastic approach that it allows to make such important quantitative predictions as the Hardy-Weinberg laws purely based on very general assumptions.

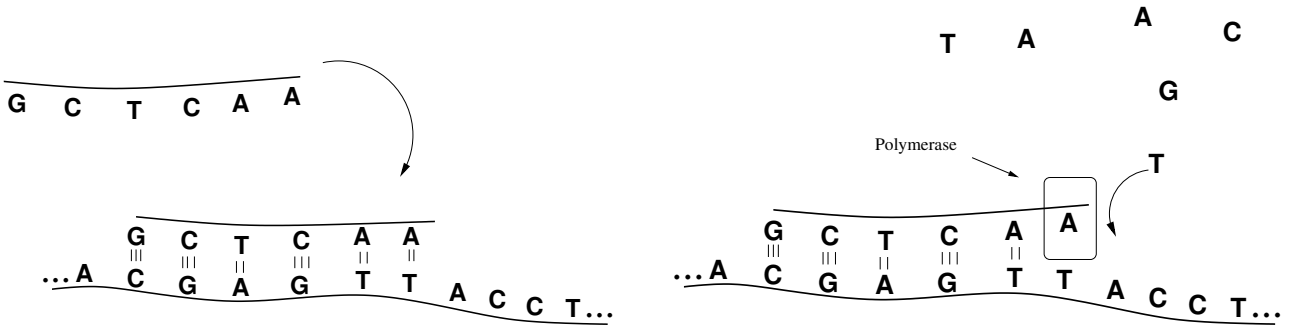


Figure 1.1.: A polymerase catalyzing the DNA synthesis, leaving a double-stranded DNA molecule. The left panel shows a primer spontaneously attaching to single stranded DNA. For purposes of illustration, the primer depicted here is only five bases long, while in reality it is usually of the length of several dozen bases. Then the polymerase attaches at the primer and starts catalyzing the synthesis of the complementary strand, using the excess G, A, T, C present in the medium as shown in the right panel.

to already be in place. This starting point is provided by 'primers', i.e. short (on the order of several dozens of bases in length) pieces of DNA that freely float around in the medium[6]. Unlike single nucleotides, these primers *can* spontaneously bind to the DNA when they find a segment that matches the entire primer, thus forming a starting point for the polymerase. After synthesis of the complementary DNA strand is completed, the polymerase detaches from the double-stranded DNA molecule.

In order to use the information stored in form of DNA, this information must be read out. This is done by an enzyme called RNA-polymerase (of the same class as the enzyme which catalyzed the synthesis of the complementary DNA strand in the reproduction process) which attaches under certain conditions to the double stranded DNA and locally opens the hydrogen bonds holding the two strand together. Then the RNA-polymerase synthesizes a molecule called ribonucleic acid (RNA) complementary to the part of the DNA that is being read. The RNA has essentially the same structure as a single strand DNA molecule but instead of thymine (T), uracil (U) is used. Then the information stored in form of DNA and now present as RNA can be transported out of the nucleus[7]. This process is called 'transcription'.
The transcript RNA is then 'translated' into a protein, i. e. a string of amino acids, by an organel called ribosome. The ribosome considers three bases of the RNA at the same time and synthesizes the amino acid corresponding to that triplet. This way the information stored and multiplied in the DNA is ultimately turned into proteins which then can be used by the cell in a large variety of ways, see fig 1.2 for a very simplified sketch of this process.
The protein thus produced can then fulfill its function. The universality of mechanisms ends on this level, as the type and amount of protein produced vary from organism to organism and

[6]The discussion here actually only applies to the situation *in vitro*: *In vivo*, primers are pieces of RNA synthesized by a type of RNA-polymerase called 'primase' to be complementary to the single-stranded DNA [91]. *In vitro*, however, the DNA primer is added 'by hand'.
[7]This only applies *in vivo*.

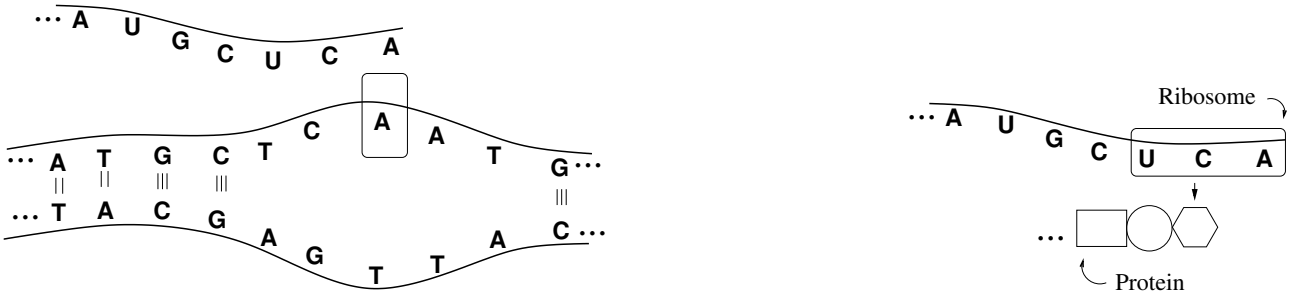


Figure 1.2.: A double stranded DNA is at a certain position opened and transcribed to RNA (left panel) by a polymerase. The dark lines represent the sugar backbone of the DNA polymer, the light lines between bases represent the double (A-T) or triple (G-C) hydrogen bonds. Note that in the RNA, uracil (U) has taken the place of thymine (T). The RNA is then translated into amino acids by a ribosome. Here, the hexagon stands for the amino acid serine, which is coded for by any triple of bases beginning with UC (right panel).

even from cell to cell within one multi-cellular organism. One of the many functions of some proteins is to regulate transcription of genes.
Since there are a total of $4^3 = 64$ different triplets but only 21 amino acids, there is a certain amount of redundancy, i.e. different triplets of DNA sequence corresponding to the same amino acid. Thus a mutation on the DNA level does not necessarily mean a change on the amino acid level. Such mutations are called 'silent' or 'synonymous' mutations and are in fitness assays sometimes used as control since such mutations are not expected to have any fitness effect. However these mutations, as well as those parts of the DNA that is never transcribed, can have a strong effect on the process of transcription and translation by changing the shape or the (thermal) stability of the DNA and RNA molecule [20, 133] and thus affect fitness [71, 117].

1.3.2. DNA synthesis and the Polymerase Chain Reaction

The two strands of DNA are complementary to each other. Thus when the double stranded form breaks apart into two single strands of DNA, each strand can serve as 'template' for synthesis of the other strand as long as there is enough G, A, T and C present. This mechanism can be utilized to amplify the amount of DNA present in a sample (see e.g. [67]): If double-stranded DNA in a sample is heated to the point where it takes the single-strand form, then is allowed to cool in a medium containing polymerase, primers and an excess amount of G, A, T, C, one will very likely end up with twice the amount of double-stranded DNA. If this procedure is repeated several (usually on the order of 10) times, one has amplified the amount of DNA by an exponentially large factor. This process is called polymerase chain reaction (PCR) and a central tool in biology with applications randing from fundamental research (an example will be presented in the next chapter) to criminology [54].

Except for chapter 2 where the data sets are described, this level of detail is not needed in this thesis. Important is the key idea that the complete information about the whole organism is stored as a sequence

$$\vec{\sigma} = \{\sigma_1, \ldots, \sigma_L\} \tag{1.9}$$

of length L. In the case of DNA, each σ_i is drawn from the alphabet $\{G, A, T, C\}$. However for theoretical purposes, the essential behavior can be captured by considering the more restricted alphabet $\{0, 1\}$.

1.4. Fitness landscapes

Since an organism is uniquely defined by its genome represented by the DNA sequence $\vec{\sigma} = \{\sigma_1, \ldots, \sigma_L\}$, i.e. by a sequence of length L, the fitness that was introduced in the context of the Kimura-equation (1.7) is ultimately a function of this sequence. Many empirical data sets, however, do not record the base at a given position, but only record whether or not a given mutation was present or absent. Thus it suffices to consider binary sequences, i.e. $\sigma_i \in \{0, 1\}$ where 0 denotes absence and 1 presence of a mutation at site i. Thus the configuration space is $\mathcal{C}_L = \{0, 1\}^L$, the L dimensional Boolean hypercube. This configuration space consists of 2^L different sequences. Since labels can be set without changing the system, the wild type introduced in sec 1.2.1 will for the remainder of this thesis be labeled as $\vec{\sigma}_0 = \{0, \ldots, 0\}$.
Fitness is a mapping from this configuration space $\mathcal{C}_L$ into the real numbers

$$W : \mathcal{C}_L \to \mathbb{R}, \tag{1.10}$$

meaning that each possible sequence $\vec{\sigma}$ has a fitness $W(\vec{\sigma})$. Then the collection of all 2^L fitness values gives the complete Fitness Landscape (FL). FLs are a central paradigm of evolutionary biology since their introduction (more as a metaphor than an actual concept) by Sewall Wright in the 1930's [152]. Evolution can be thought of as a hill climbing process happening in such a landscape, where each population 'tries to get as high a possible'.
While FLs have theoretically been studied every since the modern synthesis, empirical FLs have only become available in recent years, with the earlier works in the late 1990s [30] and an increasing number of FLs having been published since, see e.g. [21, 77, 146] and [120, 127] and references in these reviews. Thus comparing the theoretical work to empirical FLs has only recently become possible and the work done for this thesis is part of this effort.

1.4.1. Epistasis

One typical feature of empirical fitness landscapes seems to be a certain degree of ruggedness, (see e.g. [30, 146] for examples with a large amount of ruggedness, [21, 77] for examples with moderate amounts of ruggedness). The presence of this ruggedness is a sign of 'epistasis'. This term is used to indicate that the effect two mutations, if jointly present in one organism, differs from the individual effects of these mutations, see [33, 147] and references therein. The mutations then interact in some non-trivial way. Five of the possible motifs of interaction are shown in fig 1.3.
If epistasis is present, the effect of any given mutation depends crucially on the genetic background in which it occurs: If it has a beneficial effect in one background, it might have no or even a deleterious effect in another. While epistasis does not necessarily lead to local maxima (or optima), i.e. configurations that are surrounded[8] by states of lower fitness on a FL, observing such optima is a sign that epistasis plays an important role.

[8] in a suitably defined sense, see next subsection.

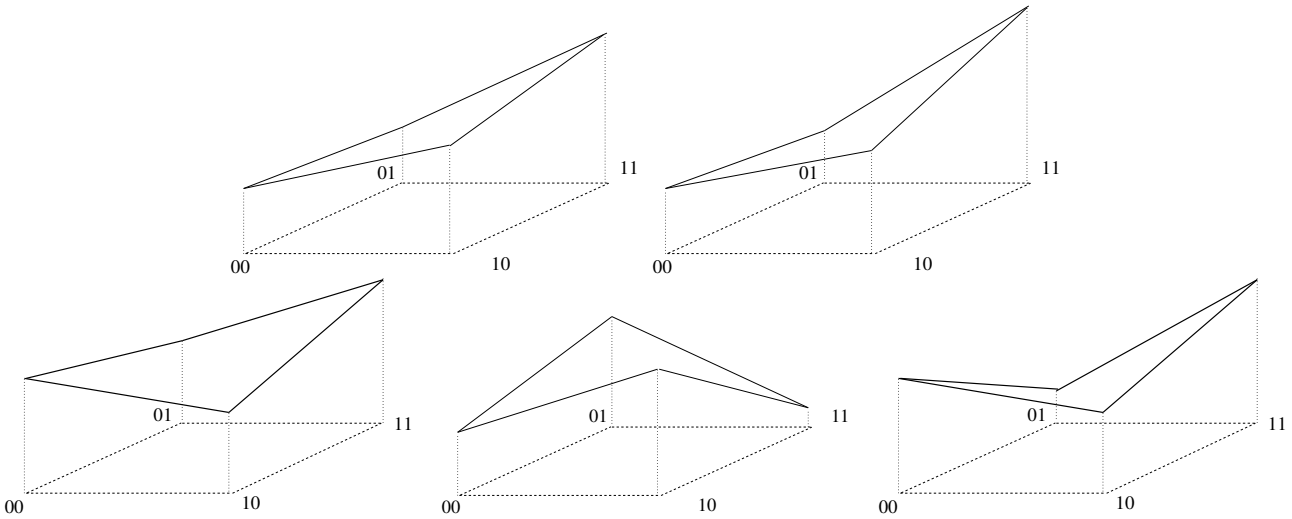


Figure 1.3.: An illustration of some of the possible consequences of epistasis on fitness landscapes on a landscape with $L = 2$. The top row shows a non-epistatic landscape (left), where the combined effect is just the sum of the two individual effects and a landscape with magnitude epistasis (right), on which the combined effect is more beneficial than the sum of the two individual effects. The bottom row shows sign epistasis (left) inverse sign epistasis (middle) and reciprocal sign epistasis (right). All fitness values are shown in arbitrary units and on arbitrary scales. Schematic plots and naming of epistasis according to [120].

1.4.2. Regimes of evolutionary adaptation

With the selection coefficients s, selection strength has been put in a quantitative form. Similarly, the mutation rate μ controls the rate at which a single member of the population mutates. Thus the rate at which new alleles enter the population is given by $N\mu$, since every single individual can mutate. These are two of the three forces of evolution. The third, called drift[9], is associated with the second derivative in the Kimura equation (1.7) . This 'drift' term encodes the random loss of alleles regardless of their fitness by frequency fluctuations. Note, however, that the term is proportional to N^{-1}. Thus for small populations, this is more important [116] than for large ones and can even give a certain advantage to small populations since they essentially can explore more options [74] (in a sense that will be specified later).
The mutation rate clearly plays a crucial role for the adaptation of the population. One effect that needs to be mentioned in this context is that of 'clonal interference'. If $N\mu$ is so low that the fixation time of each beneficial mutation is short compared to the expected time for a new mutation to arise, the process of adaptation will consist of mutation arising and, if they are beneficial and survive drift, subsequently fixating. If, however, $N\mu$ is large, then it is possible that two beneficial mutations are present at the same time and compete for fixation [114].
Within this thesis, two different regimes of adaptation will be discussed.

[9]Within the standard nomenclature of stochastic processes, the term 'drift' is used for ballistic motion and associated with the first derivative in a Fokker-Planck equation. In the biological context, 'drift' refers to the diffusive term, see also the discussion of this point in [80].

Strong selection/weak mutation

If selection is strong and the rate at which mutations are supplied $N\mu$ is small, the population is genetically homogeneous, since any deleterious mutation will be strongly selected against while a beneficial mutation will strongly be favored. Since the $N\mu$ is small, a beneficial mutation will have time to fixate in the population before a new mutation arises. This means that the population can only move by point-mutations, i.e. only adapt in steps of Hamming distance 1, where the Hamming distance between two binary sequences $\vec{\sigma}_1$, $\vec{\sigma}_2$ of length L is defined by

$$d(\vec{\sigma}_1, \vec{\sigma}_2) = \sum_{l=1}^{L} \left(1 - \delta_{\sigma_{1,l},\sigma_{2,l}}\right) \tag{1.11}$$

where $\delta_{a,b}$ is the Kronecker δ. This adaptive regime is the 'strong selection/weak mutation' (SS/WM) regime [58].

Greedy regime

If $N\mu \gg 1$ but $N\mu^2 \ll 1$ and selection is strong, the population sees all neighbors at Hamming distance one at each time, but never any neighbor further away. Thus in each step, that mutation which offers the greatest increase in fitness fixates in a population. Adaptation by this process can be likened to an optimization by steepest descent[10] and is called the 'greedy' regime of adaptation.

1.4.3. Adaptation as record driven process

Under strong selection, only mutations with higher fitness than that of the current state fixate in the population. Thus each fitness that fixates is the highest encountered so far by the population, in other words it is a fitness *record* in the sense of the famous 'Guinness book of records' [1]. This makes evolutionary adaptation in the SS/WM regime an example of a record driven process. There are many other examples of such processes from areas as diverse as dirty superconductors [109,136], sand pile models [36], stock prices [149] and of course from many areas of biology such as species extinction [9, 135] and related aspects of evolutionary biology [87, 88, 104] and many others. Thus it can be expected that record statistics will also play a central role in the study of properties of fitness landscapes.

1.4.4. Fitness landscapes as spin glasses

Spin glasses (see e.g. [100] and references therein) are, together with percolation models [139], one the most important models in the study of disordered systems [17, 61] and were originally introduced as models for disordered magnets. The configuration space of spin glasses is usually the same configuration space $\mathcal{C}_L$ as for fitness landscapes defined above (or straightforward generalizations, see e.g. the Potts model [61]). In spin glasses, each point of $\mathcal{C}_L$ is given a real-valued 'energy' similar to the fitness. The state with the lowest energy is then called 'ground state'. In the FL-picture, the stable state is the one with the *highest* fitness, but this difference is trivial and can be compensated by a change in sign. Another similarity between

[10] with the trivial difference that in this context, the population searches to *increase* fitness.

fitness landscapes and spin glasses is that they both show epistasis (or 'frustration' in the spin glass context [100]).
These similarities in configuration space and empirical observation of ruggedness have led to similar parallel developments of models, most prominently Derrida's famous 'Random Energy Model' (REM) of spin glass [34, 35] and the 'House of Cards' (HoC) model by Kingman [83], see next chapter.
The list of similarities also includes certain schemes for the dynamics on spin glasses and FLs [68] and there is quite a bit of knowledge transfer between the communities studying the two subjects, having e.g. given rise to 'genetic algorithms' that mimic a population adapting on a FL to solve certain optimization problems that can be formulated in terms of a spin glass [68]. While spin glasses are a natural frame of mind when thinking about such problems, the properties of interest in the study of spin glasses are different from those considered here, thus none of the methods used in the study of spin glasses will be applied here.

1.5. The central questions of this thesis

As was mentioned above, aim of this thesis is to study properties of FLs both theoretically and on empirical data. In this section, these properties will be introduced.

1.5.1. Accessible paths

A concept proposed to be central to adaptation on FLs in the context of development of antibiotic resistance in bacteria by Weinreich *et al.* [146] are accessible paths, that is sequences of point mutations joining the global maximum (GM) and its antipodal sequence (AS), i.e. the sequence with maximum Hamming Distance from the GM. This property is believed to be most important for the adaptation of a population in the SS/WM regime. Such a path crosses the entire state space by going from one state to another, always taking steps covering Hamming distance 1 (see fig 1.4 for two example landscapes on which such steps are indicated) and is called accessible, if and only if fitness increases in each step.

On each given fitness landscape, there is a certain number n of such paths, which can be determined by counting them (for large sequence lengths L using a computer). Counting this number of accessible paths on many FLs, one obtains a distribution of accessible paths where $\pi_L(n)$ stands for the probability of having n accessible paths. Since $\pi_L(n)$ is typically a nontrivial object to compute, one can consider the expected number of accessible paths

$$\langle n_L \rangle \equiv \sum_{l=0}^{L!} l \pi_L(l). \tag{1.12}$$

Here the upper limit of the sum ends at $L!$ which is the total number of paths when only those without loops are considered, called 'direct paths'. This is a justifiable approximation since in the SS/WM regime, mutations occur only at rate $\mu N \ll 1$ and thus even the smallest detour (taking two extra steps) is suppressed by a factor of $(\mu N)^2$ compared to paths without loops. Only direct (or shortest) paths are considered here.
Another feature to study on the distribution of accessible paths, introduced by Carneiro and Hartl [19], is the probability of having no accessible path at all on a given FL, $\pi_L(0)$. These

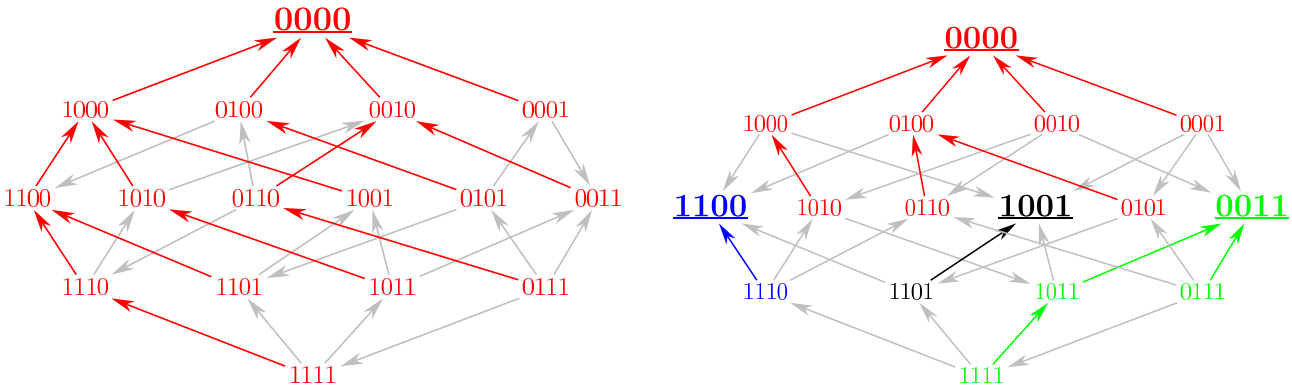


Figure 1.4.: Two excerpts of the empirical *Aspergillus niger* fitness landscape (see next chapter for details on this landscape) to illustrate the notion of accessibility. The arrows indicate direction of increasing fitness. Left: The subgraph containing the four mutations arg, pyr, leu ald oli with nine accessible paths as defined below. Right: Subgraph consisting of the mutations fwn, pyr, phe and oli with only one accessible path. Under greedy dynamics, a state of a given color leads to the maximum of the same color. Those edges which will be taken in a greedy adaptation are also given the corresponding color, the others are left gray.

two objects, $\langle n_L \rangle$ and $\pi_L(0)$, will be studied here.

1.5.2. Basin of attraction

In the regime of greedy adaptation, the population chooses in each step the neighboring state which offers the largest increase in fitness, until a local maximum is reached. Thus it performs a 'greedy adaptive walk'
Under such a dynamics, the state space $\mathcal{C}_L$ falls apart into disjoint[11] subsets, each leading to a different local fitness maximum, see fig 1.4. The statistics of these 'basins of attraction' of the local maxima and specially of the global maximum will be discussed in this thesis.

1.5.3. Sub-graphs

In natural populations, mutations can occur anywhere on the genome. Fitness landscapes in empirical studies, however, are subject to the constraint that only very few (on the order of a dozen at most) mutations can be considered, as even for $L = 10$ different mutations, $2^{10} = 1024$ different mutant strains must be created and measured to obtain a complete landscape (and even numerical simulations cannot reach realistic sequence lengths of $L \sim \mathcal{O}(10^6)$ [5] and greater). Thus theoretical predictions on a given property of the FL can only be compared to empirical data obtained for a subset of all possible mutations.
Furthermore theoretical results obtained in this thesis will make statements about the behavior of the objects considered as a function of sequence length L while empirical FLs of a given number L of mutations only correspond to one data point for that value of L without error

[11] provided that no two fitness values are equal, which will be the case with probability one in the models considered here, see next chapter.

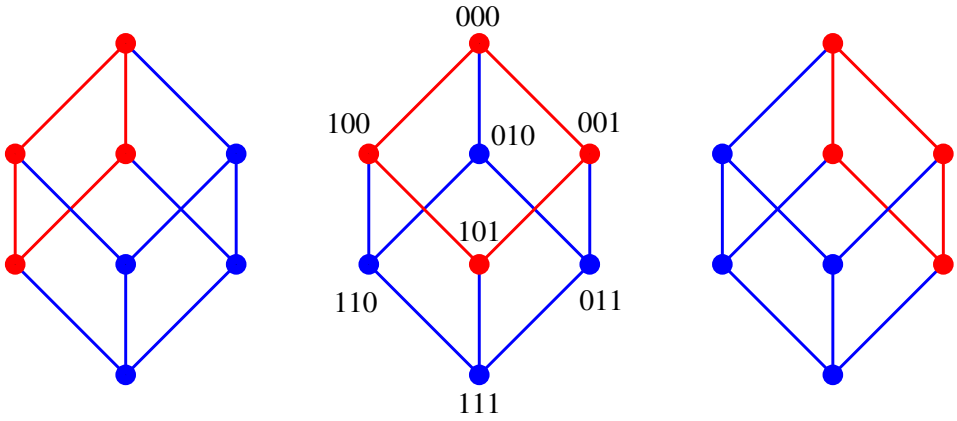


Figure 1.5.: The three possible $m = 2$ sub-graphs (squares) of the $L = 3$ state space (cube). Note that the choice of m sites to mutate on only fixes the sub-graph with the additional condition that the WT should be part of the SG. Otherwise, there is an ambiguity on the state, in which the remaining $L - m$ state should be in, leaving 2^{L-m} choices. In the example here, this corresponds to the two opposite faces of the cube.

bars. Thus comparison to predictions is not straightforwardly possible.
One way to circumvent these problems is to systematically pick a subset of $m < L$ mutations and consider the FL consisting of the 2^m fitness values of the possible states with mutations *only* at those m sites. If one chooses these sub-graphs (SG) of the full state space such that the wild type is part of the SG, there are $\binom{L}{m}$ such SGs, see fig 1.5. Thus under the condition that FLs on two different SGs are reasonably uncorrelated, one obtains one data point for each value of m and, due to the $\binom{L}{m}$ sub-graphs, also an estimate for the error bars, allowing a comparison between theoretical predictions as function of L and the m-SGs of an empirical data set and ultimately to extrapolate to large sequence lengths.
SGs have been analyzed before (see e.g. [30, 31, 115] where they were referred to as 'complete subgraphs') but in order to use them systematically, it must be established that the condition of 'reasonable uncorrelatedness' mentioned above is actually met. This is shown in appendix C.

2. Models for fitness landscapes and the data sets used

Of the many models proposed for fitness landscapes, those considered in this thesis are introduced. The data sets used for comparison of the models to experiment are presented.

2.1. Theoretical models for fitness landscapes

Fitness landscapes are a central paradigm of theoretical studies of evolution, see e.g. [56,76,81], thus a lot of theoretical work has been devoted to them. In particular, different (and often contradictory) intuitions about the biological effects determining fitness were used to develop model landscapes capturing these effects. Since all of the underlying intuitions of the models introduced [3, 56, 76, 81] are believed to capture some aspect of the biochemical interactions underlying fitness landscapes, none can be rejected or accepted on theoretical grounds alone.
One goal of this thesis is to investigate how strongly the properties mentioned in section 1.5 depend on the choice of model. Recently, it has been recognized that universality might be as pervasive a feature of evolutionary biology and genomics as it is in statistical physics [85]. In the light of this fact, such an approach probing fitness landscapes across different models for universality seems promising.

2.1.1. The House Of Cards model

The biochemical machinery responsible for reading and transcribing DNA as presented in section 1.3 consists of a large number of reactions happening in a precise timing. Thus one intuition, first introduced by J. F. C. Kingman [83], is that any random change such as a mutation destroys *the biochemical 'house of cards' built up by evolution*, as Kingman puts it. Thus the House of Cards (HoC) models asserts that the fitness of every mutant has to be 'rebuilt from scratch' and is therefore independent of the wild type.
A formal definition of the HoC model assigns to each possible state $\vec{\sigma}$ of the genome a fitness $x_{\vec{\sigma}}$ where the family of fitness values $\{x_{\vec{\sigma}}\}$ for all $\vec{\sigma} \in \{0,1\}^L$ is a family of iid RVs. Using the mapping between the configuration space $\{0,1\}^L$ and the integer numbers provided by interpreting each sequence $\vec{\sigma}$ as binary representation of an integer, this can be put as

$$W(\vec{\sigma}_n) = x_n, \quad x_n \sim f(x). \tag{2.1}$$

The only choice left to be made is the probability density for the fitness values $f(\cdot)$.

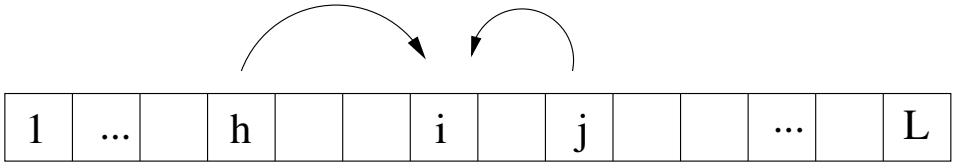


Figure 2.1.: A neighborhood consisting of sites h and j assigned to site i. Here, the $K = 2$ states are scattered all over the genome, emphasizing the fact that the expression 'neighborhood' does not necessarily imply local adjacency.

2.1.2. Kauffman's LK model

As was stated in section 1.4, mutations interact in a non-trivial way. To model these epistatic interactions, the LK model (originally called NK model) was introduced by S. A. Kauffman [76]. It starts out from the same intuitions about the finely tuned machinery of biochemical interactions, but attempts to model them in more detail: The fitness of a sequence $\vec{\sigma}$ is the sum over the contributions $w_l(\vec{\sigma})$ of each site l,

$$W(\vec{\sigma}) = \sum_{l=1}^{L} w_l(\vec{\sigma}) \tag{2.2}$$

which in turn depend on the sequence $\vec{\sigma}$.
To determine the site contributions $w_l(\cdot)$, to each site l a set of indices $\nu_l = \{l, l_1, \ldots, l_K\}$ called 'neighborhood' is assigned, see fig 2.1. Note that the neighborhood contains the site l as well, thus it contains a total of $K + 1$ indices. For each possible state of the sub-sequence $\vec{\nu}_l = \{\sigma_l, \sigma_{l_1}, \ldots, \sigma_{l_K}\}$, the fitness contribution is independently drawn from a probability density $f(\cdot)$, thus $w_l(\cdot)$ is actually just a (random) function of this sub-sequence $\vec{\nu}_l$ of length $K+1$ instead of the whole sequence $\vec{\sigma}$. One can say that each site l 'interacts' with K other sites in determining that sites fitness contribution. Since no index can appear more than once in a neighborhood, the value of K is constrained to the range of $0 < K < L - 1$.

The LK model, much like the HoC model, is defined by a family of iid RVs. However the numbers of RVs needed to fully characterize the model differ: One set of 2^{K+1} iid RVs for each neighborhood, making a total of $L2^{K+1}$ iid RVs as opposed 2^L RVs in the HoC case. In numerical simulations performed for this thesis, the 2^{K+1} RVs for each state of each neighborhood are stored in look-up tables.

The parameter K which governs the number of interaction partners for each site is the most important tuning parameter in the model. It enables the model to cover a wide range of fitness landscapes: For $K = 0$, each site l has only two fitness contributions corresponding to the two states $\sigma_l = 0$ and $\sigma_l = 1$ independently drawn from the probability density $f(\cdot)$. Since $f(\cdot)$ is assumed to be continuous, one of these two fitness values will be greater than the other with probability one. Thus each site can optimize its fitness independently of the others and there is one unique global maximum with all sites in their 'good' position and one unique global minimum with all sites in their 'bad' position. In between, the landscape is smooth in the

sense that one can always flip one spin from the bad to the good position, corresponding to a mutation that allows further adaptation on the landscape. This type of landscape where all pathways are accessible and, due to a lack of maxima other than the global one, all states lead to the GM under greedy adaptive dynamics, is called the 'Mount Fuji' type landscape, since it resembles the famous Mt Fuji of Japan [76].
The other extreme is the case of $K = L - 1$, where each site interacts with all others (provided that no site can appear twice in a given neighborhood). Here, any mutation will cause all L single-site contributions to be replaced with another set of iid RVs, thus as far as order statistics considerations are concerned, this case is equivalent to the HoC model. Note however that the iid RVs will not be drawn from the original distribution $f(\cdot)$ but from its L-fold convolution [46] $f^{*L}(\cdot)$, since the fitness of each state is still the sum over L single site contributions.
For values of $0 < K < L - 1$, the FL will display an intermediate degree of ruggedness. Since the value of L is usually set by the system under consideration, the value of K is the most important parameter for fitting the LK model to empirical data.

The neighborhood and specially the way in which it is assigned constitutes another important degree of freedom when defining the model. For a given site l, does one simply choose the next K sites as neighborhood, $\nu_l = \{l, l+1, \ldots, l+K\}$ with periodic boundary conditions if necessary? Does one divide the K sites evenly and distribute them to both sides of l, or is the neighborhood simply assigned by uniformly picking K sites at random without replacement from the L sites? It appears that the way in which the neighborhood is chosen has implications for the computational complexity of the problem [145], but the properties considered here did not seem to be influenced by the choice of neighborhood [84]. Thus throughout this thesis, for numerical simulations the LK model was considered for randomly chosen neighborhoods.

2.1.3. Rough Mt Fuji model

In many empirical fitness landscapes, see e.g. [30,31,51], the states at small Hamming distance from the global maximum tend to be higher in fitness than those further away, thus there is some sort of effective trend in fitness toward the GM. To systematically study the effect of this overall trend, the 'Rough Mt Fuji' (RMF) model introduced by Aita and coworkers [3] was used. In this model, the fitness values of each state are assigned independently, but not all fitness values have the same distribution anymore. Rather, the distribution keeps its shape but has its mean value moving linearly with distance from the GM. If the GM is at state $\vec{\sigma}_0$, i.e. the WT, then a state $\vec{\sigma}_n$ has fitness

$$W(\vec{\sigma}_n) = x_n - cd(\vec{\sigma}_0, \vec{\sigma}_n) \tag{2.3}$$

where the $\{x_n\}$ are iid RVS and $d(\vec{\sigma}_1, \vec{\sigma}_2)$ is the Hamming distance between the two sequences in the argument as defined in eq (1.11). Thus states in the shell of Hamming distance 1 from the GM are typically c units of fitness lower than the GM but c units of fitness higher than those states in the Hamming distance 2 shell. For the numerical simulations, only those realizations of a RMF landscape were considered where the state $\vec{\sigma}_0$ towards which the FL is oriented actually was the global optimum.
For $c \equiv 0$, this model is again the HoC case, while in the limit $c \to \infty$, all paths are accessible

again and, due to a lack of maxima of other than the GM, all states are in the BoA of the GM. It is thus equivalent to the Mt Fuji model. Much like the LK model defined above, this model has a tunable degree of ruggedness controlled by the parameter c in this case, by which it can be fitted to empirical data. From a theoretical perspective, however, this model is a prototype for landscapes with some underlying trend that is asymptotically amenable to analytic computations. It will mostly be used to perturbatively examine the influence of a very small average slope c on the HoC model.

2.1.4. The Neutral model

Yet a different intuition about fitness starts from the assumption that the fitness *value* is not the important feature but only whether or not a given mutant is viable at all because if a mutation hits a non-coding part of the DNA or produces a silent mutation, this will have no effect[1] while a non-synonymous mutation on a coding part of DNA will be lethal. Based on this idea, Kimura and Ohta [81] developed the Neutral Theory of Evolution, here referred to as 'Neutral Model'. They imagined fitness landscapes as vast stretches of a plateau of validity with valleys of lethality.
In the simplest case, each state $\vec{\sigma}$ is randomly and independently of the others either viable (fitness $W(\vec{\sigma}) = 1$) with probability p or lethal (thus $W(\vec{\sigma}) = 0$) with probability $1 - p$. This model of FLs is equivalent to the problem of site percolation [139] defined on the L-dimensional hypercube [57]. Within this model, there is a natural reformulation of accessibility: A mutational pathway is called accessible if none of the intermediate states is lethal. Thus asking whether two given states are joined by an accessible path becomes in some sense equivalent to asking whether both states are part of the 'percolating cluster' [139], an object that has been studied in the percolation context from the very beginning [49].
A more complicated version of this model where lethality is not uniformly distributed across sequence space but rather depends on the identity of mutations present will be discussed in the context of one of the empirical landscape used during this thesis.

2.2. Empirical fitness landscapes

Empirical fitness landscapes were the driving force behind this work as they inspired many of the questions asked. While theoretical work, ongoing since the works by Wright [152], has given rise to the models introduced above (among others), empirical fitness landscapes have only become available in the last decade, starting with the *Aspergillus niger* landscape by J. A. G. M. de Visser and coworkers in 1997 [30]. These empirical data sets have proven to be very provocative, leading for example to questions about the distribution of accessible paths, first asked in the context of the *Escherichia coli* landscape by D. Weinreich *et al.* [146] and treated theoretically in [19, 51]. While the questions asked of these fitness landscapes (and the methods used here to answer them) are in principle applicable to any of these landscapes ([21, 69, 77, 93, 94], see also [120, 127] and references therein), in this section, only those data

[1] This is actually not correct since synonymous mutations are known to influence the expression levels of certain genes as well as the folding stability of the resulting RNA, see the references given in the previous chapter on that point.

name	abbrev.	effect
fwnA1	fwn	color mutation (fawn-colored conidiospores)
argH1	arg	arginine deficiency
pyrA5	pyr	pyrimidine deficiency
leuA1	leu	leucine deficiency
pheA1	phe	phenyl-alanine deficiency
lysD25	lys	lysine deficiency
oliC2	oli	oligomycin resistance
crnB12	crn	chlorate resistance

Table 2.1.: The eight mutations present in the strain N890 in order of increasing chromosome number and their effects as stated in [30]. The abbreviations given will be employed throughout this thesis.

sets primarily considered in this thesis will be presented, namely the *A. niger* data set [30] and a very recent *E. coli* data set by M. Schenk and coworkers [134].

2.2.1. The Aspergillus niger landscape

The data set most important for this thesis was obtained from the filamentous fungus *Aspergillus niger* and first described in [30].

Creating the data set

Aspergillus niger usually has a haploid life cycle but nuclei can spontaneously merge to form diploid organisms. This process happens at a very low rate and the resulting diploid organism is unstable: The diploid nucleus can spontaneously separate again to form haploid nuclei [30]. As this happens, the chromosomes are randomly redistributed among the two emerging nuclei, see fig 2.2 for a caricature sketch of this process also called 'para-sexual cycle'. This chromosome redistribution was used to create the mutant strains that make up the data set. Initially, only two strains were present: One, called N411, carried only a color mutation conferring olivine-color spores, while the other, called N890, contained - in addition to the color mutation also present in N411 - exactly one mutation of each of eight chromosomes. These mutations, induced by low doses of UV light (see [30] and references therein for details), consisted of two resistance mutations, five deficiencies and one marker mutation that changed spore color, see table 2.1. The two strains had a sufficiently recent common ancestor to assure that they only differed by these mutations.

Now if these two strains N411 and N890 are left to merge and the combined diploid is then haplodized again under random chromosome redistribution, the resulting nuclei will each contain a random number of mutations, see fig 2.2. Thus a possible number of $2^8 = 256$ different mutants could be produced, corresponding to the 2^8 different states of a binary sequence of length $L = 8$. In the total of about $S = 2500$ segregants, however, only 186 of these 256 possible mutant were found. On those that were present, fitness was measured in terms of the mycelial growth rate, which turned out to be highly correlated to spore count and thus fecundity. The full data set of all mutations found and their respective fitness values is given in

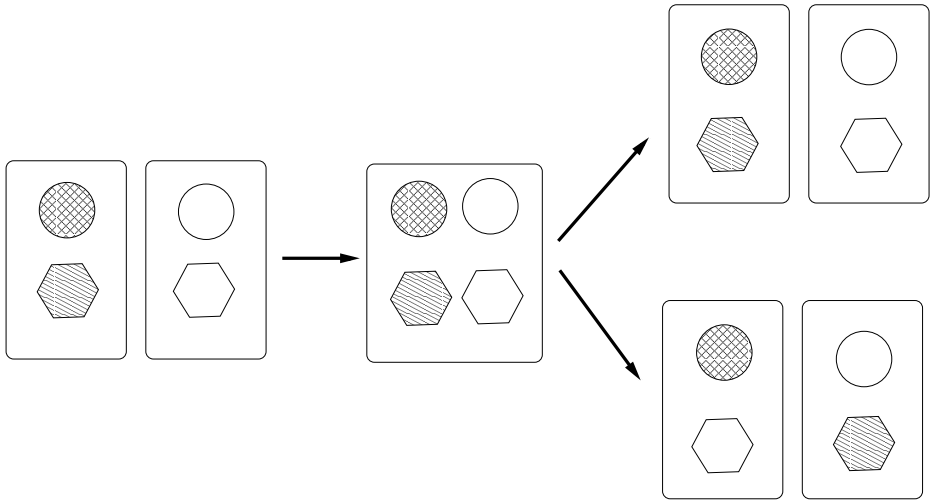


Figure 2.2.: A sketch of the process of spontaneous diploidization and subsequent re-haploidization under random distribution of chromosomes for a hypothetical organism with two chromosomes (circle and hexagon). Chromosomes carrying a mutation are shaded while those without mutation are left open. In this illustration with two chromosomes, only two outcomes are possible: Either the two mutations are in one organism after re-haploidization or they are split up between the two segregants. With three chromosomes, already $2^3/2 = 4$ outcomes are possible, corresponding to $2^3 = 8$ different mutants.

appendix A. Note, however, that the eight-mutant N890 is not on the list as it was not found among the segregants. Even though this strains is available, its fitness was not measured as, for consistency, the data set was only made up of segregants found after haploidization [32].

Discussion of the data set

As mentioned above, not all 256 possible states were found. Assuming that all mutants were equally likely to be found[2], the chance of any single one mutant having been missed by chance is $p = (1-1/256)^S \approx 5.6\times10^{-5}$. The probability p_n for at most n genotypes to have been missed is then given by a Poisson distribution with mean $256 \times p \approx 0.014$ [51]. Thus the probability of having missed none or at most one of the mutants can be estimated as $p_0 = 1-256\times p \approx 0.986$ and $p_1 \approx 1-(256\times p)^2 \approx 0.9998$ respectively. A more conservative estimate can be obtained by taking into account that finding a mutant can depend on its fitness. Assuming that each mutant has a different likelihood to be found, which is uniformly distributed on the interval $[r, 1]$ with $r = 0.274$ corresponding to the lowest fitness measured among the 186 mutants found (see appendix A), numerical simulations yielded an estimate of $p_0 \approx 0.74$ for the probability of finding each mutant, and for the probability of missing at most one $p_1 \approx 0.956$. Under both assumptions it is very unlikely to have missed more than, say, 3 states, let alone all 70 missing mutants. Thus they were assigned a fitness value of 0. It was assumed that those combinations of mutations were lethal for the organism.

[2]A nontrivial assumption, as even though each combination of mutations is equally likely to arise by segregation, detecting them depends on growth of that mutant.

These missing genotypes were, however, not uniformly distributed across all mutations [51]: 62 out of 70 lethal strains contained the lysine deficiency mutation *lys*. This could be explained as outcome of a model where each mutation can independently induce lethality if present in a given mutant. In the simplest version of this model, all mutations have the same probability q of inducing lethality, thus a mutant with n mutations is viable with probability $\text{Prob}[viable] = (1-q)^n$. Then, taking into account that $\binom{L}{n}$ different combinations of n mutations with sequence length L are possible, the average number of viable states is given by

$$N_{viable} = \sum_{n=0}^{L} \binom{L}{n} (1-q)^n = (1+(1-q))^L = (2-q)^L \tag{2.4}$$

where the binomial formula was used. The almost two-fold over representation of one mutation among lethal genotypes, however, does not seem to be compatible with this single-parameter model.
A similar picture is obtained by analyzing the distribution of lethal states across sub-graphs of this data set as defined in section 1.5.3, where one observes that many SGs contain lethals. One approach to circumvent this problem, used for example in [31], was to only analyze sub-graphs without lethal states, called 'viable sub-graphs' (VSGs). In [51] it was found that none of the VSGs of the *A. niger* data set contained the *lys* mutation, while the other mutations were approximately evenly distributed across VSGs, which also points to a more complicated situation. The next simplest thing is to assume one lethality probability q_{lys} for the *lys* mutation and another q_0 for all others. Under this model, the average number of viable states is given by

$$N_{viable} = (2-q_{lys})(2-q_0)^7 \tag{2.5}$$

for all mutations and

$$\tilde{N}_{viable} = (2-q_0)^7 \tag{2.6}$$

for mutants that did not contain the *lys* mutation. Equating the expected value N_{viable} with the number of states found, $N_{viable} = 186$, and $\tilde{N}_{viable}$ with the number of these that did not contain the *lys* mutation, $\tilde{N}_{viable} = 120$, one obtains the estimates $q_{lys} \approx 0.45$ and $q_0 \approx 0.018$. Computing the expected number of VSGs for each SG-size m for this model with these two parameters and comparing to the numbers of VSGs actually observed in the data set supports the model of two lethality probabilities with these two values, see table 2.2. Thus it was concluded that the missing states can be assumed to be lethal and that lethality of a given state is strongly influenced by the number and identity of mutations present in that state. It would be very interesting to understand the strong influence of the lysine deficiency mutation *lysD*25 on the lethality of a state on a biochemical or biological basis.

2.2.2. TEM 1 β-lactamase resistance data set

β-lactam antibiotics are one of the most important classes of antibiotics, containing for example the penicillins and cephalosporins [4]. The TEM 1 β-lactamase enzyme is one of the most prominent antibiotic resistance enzymes and much effort has been devoted to understanding

m	No. of SG	No of VSG	Expected from model
2	28	20	19.5
3	56	29	28.1
4	70	19	19.5
5	56	4	4.9
6	28	0	0.2

Table 2.2.: From [51]. The first column shows the sub-graph size m, the second column the possible number $\binom{L}{m}$ of sub-graphs of each value of m. The third column shows the number of VSGs observed in the empirical *A. niger* data set and the fourth column the expected number of VSGs from the model with two different lethality probabilities above with $q_{lys} \approx 0.45$ and $q_0 \approx 0.018$. Note the good agreement with model predictions and empirical fitness data.

the mechanisms underlying this enzymes' function, see e.g. [131] and references therein. Study of this resistance enzyme has also given rise to the *Escherichia coli* fitness landscapes by Weinreich *et al.*, [146], which however will not be discussed here. One approach to investigating this enzyme was to turn the gene that codes for it into a plasmid, i.e. a closed loop of double-stranded DNA that can be brought ('transformed' [5]) into the bacterium, where it is then expressed and copied just like the 'native' DNA of the bacterium. Comparing 'regular' *E. coli* and those that have the plasmid then allows to probe the effect of the TEM 1 gene on antibiotic resistance.

Creating the mutants

The advantage of having the TEM 1 gene on a plasmid is that one can induce mutations on the plasmid, bring it into the bacterium and then observe the effect of that mutation on antibiotic resistance [65]. Unlike in the *A. niger* data set, now it is possible to control the precise position at which the mutation hits. In the *A. niger* data set, mutations were induced by UV radiation and the only way to be sure that *only* those mutations detected were present was to use very low doses of light. For the TEM 1-plasmid, mutants can be controlled on a base-pair level by a process called 'site directed mutagenesis' (SDM) which underlies the *E. coli* data set by Schenk *et al* [134] that will be presented here.
SDM by the QuickChange^{TM}protocol uses the principle of Polymerase Chain Reaction (PCR), not only to amplify the amount of DNA present but also to induce an error at one precise location. As was stated in section 1.3, the individual hydrogen bonds between bases are quite weak. As a consequence, single-stranded DNA does not spontaneously synthesize its complementary strand even when in an excess of G, A, T, C but needs a polymerase to do so. First, however, a primer needs to attach to the single stranded DNA, thus giving the polymerase someplace to start, as explained in section 1.3. However, because the individual hydrogen bonds are so weak, single errors ('mismatches') in the primer are allowed. If the primer contains one mismatch[3], it can still spontaneously bind and thus the mismatched primer is incorporated into the complementary strand, see fig 2.3.
If the resulting double stranded DNA is then submitted to further cycles of PCR, by far the

[3]provided it is sufficiently far away from the ends of the primer

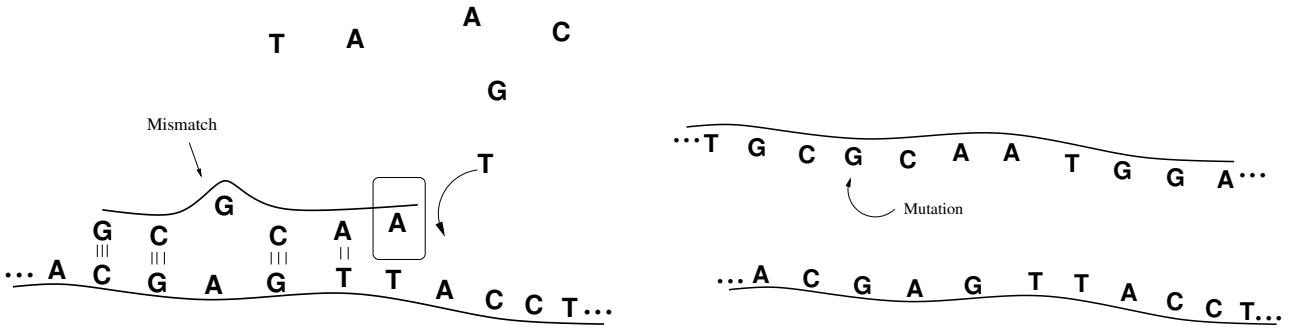


Figure 2.3.: Left: A mismatched primer can still spontaneously attach and serve as starting point for the polymerase. If the resulting double-stranded DNA is then heated and thus turned into two single DNA strand for another cycle of PCR, the DNA with the mutation will be amplified.

majority of DNA will have the mutation imposed by the mismatch primer. Then any residual plasmid without mutation is removed and the plasmid containing the mutant is, by transformation, put into *E. coli* organisms which are left to multiply in a nutrient-rich environment for about one day. Since the plasmid is copied together with the 'native' DNA of the bacterium, the amount of plasmid is strongly amplified and subsequently harvested by destroying the bacteria and extracting the plasmid, which is then prepared for long time storage.
Now primers can be synthesized industrially at reasonable cost, thus introducing a single mutation at a given site of a known gene is (at least in principle) quite straightforward. Once a mutation has been introduced into the plasmid, one can proceed in creating double mutants and so forth. The process described above, however, is subject to the constraint that only one mutation at a time can be introduced into the plasmid, thus creating a large set of mutations if rather laborious and costly.

Measuring antibiotic resistance of the mutants

In the data set obtained by Martijn Schenk [134], three landscapes of $L = 4$ mutations each were constructed using the scheme presented above. One landscape consisted of mutations known to have a large effect on the antibiotic resistance, one of mutations which were known to have a small effect and one landscape consisted of synonymous mutations. For each of these 3×2^4 mutants, its resistance to Cefotaxime was measured. In order to do so, the mutant plasmid created as described above were, by transformation, put into a reference strain of *E. coli*. Then the minimal inhibitory concentration (MIC) of the antibiotic Cefotaxime needed to kill a precisely defined fraction of the bacteria was measured. Even though this antibiotic is a cephalosporin [4], the TEM1 β-lactamase does not infer a high resistance to it [131, 132]. However only very few mutations on the TEM1 gene are necessary to increase Cefotaxime resistance by about five orders of magnitude [146]. The aim [134] was to investigate the fitness increase induced by each combination of mutations on TEM 1 in *E. coli*.
Furthermore, one data set was obtained which only consisted of a number of individual mutations induced on the WT. The purpose of this data set was to obtain information about the distribution of the resulting fitness effects.

3. Record statistics of non-identically distributed random variables

Analytic results on record statistics beyond the standard setting of independent, identically distributed random variables are presented. Some of these results have direct implications in the context of fitness landscapes, but the material presented here is also of interest in the general context of record statistcs and will be applied to study the effects of mutations.

To the extent that evolution can be seen as a process of perpetual improvement of a population, it consists of a large sequence of record values (in the sense of e.g. 'fastest growth') of the trait under selection. Thus part of the work on the present thesis was concerned with the statistics of records. The precise way in which records are related to fitness landscapes is omitted here and the results obtained will be presented without reference to fitness. The implications for fitness landscapes will become apparent in the next chapter.
Only section 3.2 below will be needed in the context of fitness landscapes. The results presented in sections 3.3 and 3.4, even though they are new in the sense that they were derived in the context of this thesis [53, 150], do not have immediate applications in the context of fitness landscapes. However, based on these results, a new tool of data analysis was derived, which, as proof of principle, will be used to extract information from the data set of antibiotic resistance of the bacterium *Escherichia coli* introduced in section 3.4. Thus the quite sizable digression into the realm of record statistics presented in this chapter will tie back into the main theme of this work in a surprising yet useful way.

3.1. Record and Extreme Value statistics

Consider a family $\{x_n\} \equiv \{x_i\}_{i=1,2,\ldots,n}$ of n real-valued random variables. Then an entry x_n is called a *record* whenever $x_n > \hat{x}_{n-1}$, with $\hat{x}_{n-1} \equiv \max_{i\leq n-1}\{x_i\}$, i.e. whenever it is greater than all previous entries in the sequence. Because only *previous* entries in the family of RVs are considered, $\{x_n\}$ can be interpreted and will sometimes be referred to as a time series. Now record statistics is concerned with the study of record events by, for example, considering the probability p_n that the n^{th} entry is a record,

$$p_n \equiv \text{Prob}[x_n \text{ is record}] = \text{Prob}\left[x_n > \max_{i\leq n-1}\{x_i\}\right]. \tag{3.1}$$

Since x_n must be *greater* than any previous entry in the time series, p_n is the probability for an *upper* record. The probability for a lower record, i.e.

$$\text{Prob}[x_n < \min_{i\leq n}\{x_i\}] \tag{3.2}$$

can be considered along the same lines as upper records. In the following, only upper records will be considered.

For many practical applications, it is only of secondary importance, *when* a record will occur, and of much greater interest *how large* the record value $\hat{x}_n$ will be. Perhaps the oldest of these applications is the problem of building a dike large enough to resist the highest flood to be expected in a given number of years to come [29,63] while on the other hand obeying economic constraints on the size. If flood heights can be modeled as RVs, one needs a precise estimate of the distribution of the RV $\hat{x}_n$, i.e. the *maximal* value of the time series $\{x_n\}$ of flood heights. This estimate is provided by Extreme Value Theory (EVT) [29, 55, 63].

In the simplest case, the RVs $\{x_n\}$ are independent and identically distributed with a common density $x \sim f(x)$ such that $\text{Prob}[x \leq y] = \int^y dx f(x) = F(y)$ where $F(y)$ is called the cumulative distribution function[1] called 'parental distribution'. By independence, the distributions of the individual entries of the time series factorize and thus

$$F_{max}(y) = \text{Prob}\left[\max_{i\leq n}\{x_i\} \leq y\right] = \prod_{i=1}^{n} F_i(y) = F^n(y), \tag{3.3}$$

where in the last step the fact was used that all RVs are iid, which allowed to drop the index i. Above, the convention was employed that an omitted lower (upper) limit of integration stands for the lower (upper) boundary of the support of $f(\cdot)$. This convention will be used throughout this thesis.

For some choices of $f(\cdot)$, eq (3.3) can explicitly be evaluated. For example setting $f(x) = \exp[-x]$ for $x > 0$, one has $F(y) = 1 - \exp[-y]$ and thus for $n \to \infty$ and $y \to \infty$

$$F^n(y) = \left(1 - e^{-y}\right)^n \approx \exp(-ne^{-y}) = \exp\left(-e^{-(y-\ln(n))}\right). \tag{3.4}$$

One observes that as $n \to \infty$, the probability that all RVs are below any given finite y tends to zero, reflecting the fact that the maximum of a time series tends to the upper bound of the support of $f(\cdot)$, which is in this case infinite. Under the rescaling $\tilde{y} = y - \ln(n)$, the distribution function for the maximum of the time series takes the form

$$F_{max}(\tilde{y}) = \exp\left(-e^{-\tilde{y}}\right). \tag{3.5}$$

This function does not depend on n anymore and thus one has found the *limit distribution*, which in this case is known as Gumbel distribution. A more detailed study (see for example the review [63], the textbooks [29,55] or for a more recent derivation [14]) shows that any parental distribution $f(\cdot)$ falls with respect to the distribution of the maximal value into one of three universality classes, each of which associated with a limit distribution for the maximum value $\hat{x}$. For each distribution $f(\cdot)$ there is a sequence of parameters a_n, b_n such that $\hat{x}$ has distribution

$$\text{Prob}\left(\frac{\hat{x}_n - a_n}{b_n} \leq y\right) = \exp\left\{-(1+\xi y)^{-\xi^{-1}}\right\} \text{ as } n \to \infty \tag{3.6}$$

for all y such that $1 + \xi y > 0$. The shape parameter ξ is determined by the tail behavior of the parental distribution. There are three different regimes of this shape parameter ξ, each with a

[1] For the remainder of this thesis, it will be assumed that f has no contributions in the form of a Dirac-δ and that thus F is a continuous function.

different form of limiting distribution associated with it. These regimes are

Gumbel class ($\xi = 0$): If $f(x)$ decays *faster than any power law* as $x \to \infty$, eq (3.6) takes the form

$$\lim_{n\to\infty} \text{Prob}\left(\frac{\hat{x}_n - a_n}{b_n} \leq y\right) = \exp\left(-e^{-y}\right) \tag{3.7}$$

Fréchet class ($\xi > 0$): If $f(x)$ decays like a power law as $x \to \infty$, eq (3.6) takes the form

$$\lim_{n\to\infty} \text{Prob}\left(\frac{\hat{x}_n - a_n}{b_n} \leq y\right) = \begin{cases} 0 & y \leq 0 \\ e^{(-y)^{-\xi^{-1}}} & y > 0 \end{cases} \tag{3.8}$$

Weibull class ($\xi < 0$): If the upper limit of the support of $f(x)$ is finite, eq (3.6) takes the form

$$\lim_{n\to\infty} \text{Prob}\left(\frac{\hat{x}_n - a_n}{b_n} \leq y\right) = \begin{cases} e^{-(-y)^{\xi^{-1}}} & y < 0 \\ 1 & y \geq 1 \end{cases} \tag{3.9}$$

The result that any parental distribution $f(\cdot)$ falls in one of these three universality classes and the connection between these universality classes and the form of $f(\cdot)$ is called Fisher-Tippet-Gnedenko theorem [29]. It is quite a remarkable results, since the parental distributions can have very diverse structures. It is similar to the famous Central Limit Theorem explaining the 'omnipresence' of the Gaussian distribution, see e.g. [16] and references therein. The above statement of universality in EVT was not only given here for completeness but also because in the remainder of this chapter, traces of these universality classes will be recovered.

3.2. Records and sequences of records from random variables with a linear trend

As one-dimensional version of the RMF-model presented in section 2.1.3, consider a family of RVs

$$x_n = y_n + cn, \tag{3.10}$$

where the $\{y_n\}$ are a family of iid RVs with density $f(\cdot)$ and c is a real positive constant. This model was introduced in [10, 11]. Here, it will be referred to as the 'Linear Drift Model' (LDM). Since this model's initial formulation, record events from the LDM have not received much attention. Only in a recent surge of interest in record and extreme value statistics beyond the iid case (see e.g. [96, 149] for random walk correlated RVs with or without drift or [52, 89, 150, 151] for records from the LDM), which was partially triggered by problems from biological contexts [87, 115], this question has been treated again.
The $\{x_n\}$ are a family of independent, but not identically distributed RVs. Rather, the entries of the time series have a distribution that keeps its overall shape but is shifted (or 'drifted'[2], hence the name of the model) by a factor cn with respect to the zeroth entry of the time series, see fig 3.1.

[2]Here, the word 'drift' is used in the sense of standard physics terminology, meaning ballistic motion, as opposed to the use of this word in the biological context.

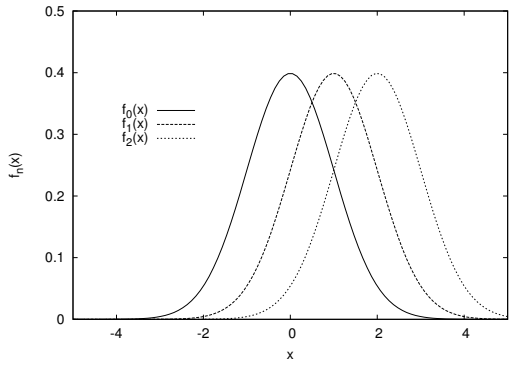


Figure 3.1.: Parental probability density $f_0(x)$ and the probability densities $f_1(x)$ and $f_2(x)$ of the first two entries x_1 and x_3 of the time series $\{x_n\}$ for $f_0(x)$ a standard Gaussian density and $c = 1$. The overall shape remains the same, but f is shifted by a factor cn. Thus the probability density of the n^{th} RV can be mapped onto that of the zeroth entry by the substitution $x \to x - cn$.

In other words, if the shape of the probability density centered around zero, say, is given by the function $f(\cdot)$, then the probability density of the n^{th} entry $f_n(\cdot)$ is related to $f(\cdot)$ via

$$f_n(x) = f(x - cn) \iff f_n(x + cn) = f(x), \tag{3.11}$$

see also fig 3.1.

3.2.1. Records from the LDM

The random variables are still independent, thus their joint probability factorizes (cf. eq (3.3)) and the probability that the n^{th} RV is a record is given by

$$p_n(c) \equiv \text{Prob}\left[x_n = \max_{i \le n}\{x_n\}\right] = \int \mathrm{d}x f_n(x) \prod_{i=1}^{n-1} F_i(x). \tag{3.12}$$

For $c \equiv 0$, one recovers the iid case, where the indices can be removed. In that case $p_n(c = 0)$ can explicitly be computed as

$$p_n(c = 0) = \int \mathrm{d}x f(x) F^{n-1}(x) = \int_0^1 \mathrm{d}u u^{n-1} = \frac{1}{n}, \tag{3.13}$$

where the substitution $u \equiv F(x)$ was used. This result can directly be verified, since in the iid case, each of the n RVs has equal probability of being the largest.
In the case of $c > 0$, eq (3.11) can be used to remove the indices in eq (3.12) and express p_n in terms of the parental probability density $f(\cdot)$,

$$p_n(c) = \int \mathrm{d}x f(x - cn) \prod_{i=1}^{n-1} F(x - (n - i)c) = \int \mathrm{d}x f(x) \prod_{i=1}^{n-1} F(x + ic). \tag{3.14}$$

This probability was the central object of study in [10], where it was noticed that it is in

general very difficult to compute. However for the Gumbel distribution with probability density $f(x) = \exp\left[-e^{-x} - x\right]$ and cumulative distribution function $F(x) = \exp\left[-e^{-x}\right]$, Ballerini and Resnick showed that p_n can exactly be computed: Using the fact that

$$F(x+c) = \exp(-e^{-x-c}) = \exp(-e^{-x}e^{-c}) = \exp(-e^{-x})^{e^{-c}} = F(x)^{\alpha} \tag{3.15}$$

with $\alpha \equiv e^{-c}$ and substituting this into eq (3.14), one obtains

$$\begin{aligned} p_n(c) &= \int \mathrm{d}x f(x) \prod_{i=1}^{n-1} F(x)^{\alpha^i} = \int \mathrm{d}x f(x) F(x)^{\sum_{i=1}^{n-1}\alpha^i} \\ &= \left(\sum_{i=0}^{n-1} \alpha^i\right)^{-1} = \frac{1-e^{-c}}{1-e^{-nc}} \end{aligned} \tag{3.16}$$

by use of the same substitution as above and the incomplete geometric series [60].

3.2.2. Sequences of Records from the LDM

So far, as in [52], the effect of a small drift c on the record rate $p_n(c)$ was studied. Of immediate importance to the study of fitness landscapes however is the probability P_n that all n RVs are ordered. This probability can be expressed as

$$\begin{aligned} P_n(c) &= \int \mathrm{d}x_n f(x_n) \int \mathrm{d}x_{n-1} f(x_{n-1}) \ldots \int \mathrm{d}x_1 f(x_1) \mathbb{1}_{x_1<x_2<\cdots<x_n} \\ &= \int \mathrm{d}x_n f(x_n) \int^{x_n} \mathrm{d}x_{n-1} f(x_{n-1}) \int^{x_{n-1}} \ldots \int^{x_2} \mathrm{d}x_1 f(x_1) \\ &= \int \mathrm{d}x_n f(x_n) \int^{x_n+c} \mathrm{d}x_{n-1} f(x_{n-1}) \int^{x_{n-1}+c} \ldots \int^{x_2+c} \mathrm{d}x_1 f(x_1), \end{aligned} \tag{3.17}$$

where the indicator function $\mathbb{1}_{x_1<x_2<\ldots x_n}$ was absorbed in the integral boundaries and eq (3.11) was used in the last step.

Explicitly solvable examples

This chain of integrals is, just like the integral expression for $p_n(c)$ in eq (3.14) above, usually hard to evaluate, but in the iid case $c \equiv 0$ and for the Gumbel distribution, it can be computed explicitly. In the iid case, this yields

$$\begin{aligned} P_n(c=0) &= \int \mathrm{d}x_n f(x_n) \int^{x_n} \mathrm{d}x_{n-1} \ldots \int^{x_2} \mathrm{d}x_1 f(x_1) \\ &= \int \mathrm{d}x_n f(x_n) \int^{x_n} \mathrm{d}x_{n-1} \ldots \int^{x_3} \mathrm{d}x_2 f(x_2) F(x_2) \\ &= \int \mathrm{d}x_n f(x_n) \int^{x_n} \mathrm{d}x_{n-1} \ldots \int^{F(x_3)} \mathrm{d}u u \\ &= \frac{1}{2} \int \mathrm{d}x_n f(x_n) \int^{x_n} \mathrm{d}x_{n-1} \ldots \int^{F(x_4)} \mathrm{d}u u^2 = \cdots = \frac{1}{n!}. \end{aligned} \tag{3.18}$$

As in the case of record rates $p_n(c=0)$, this has an intuitive interpretation since all $n!$ orderings or n iid RVs are equally likely and the one that increases in magnitude occurs in $1/n!$ of the times. The two quantities $p_n(c=0)$ and $P_n(c=0)$ are related via

$$P_n(c=0) = \frac{1}{n!} = \prod_{i=1}^{n} \frac{1}{i} = \prod_{i=1}^{n} p_i(c=0). \tag{3.19}$$

For the Gumbel case $F(x) = \exp(-e^{-x})$, eq (3.17) can also be evaluated explicitly using the Gumbel formula eq (3.15):

$$\begin{aligned} P_n(c) &= \int \mathrm{d}x_n f(x_n) \int^{x_n+c} \mathrm{d}x_{n-1} f(x_{n-1}) \int x_{n-1}+c \ldots \int^{x_3+c} \mathrm{d}x_2 f(x_2) F(x_2+c) \\ &= \int \mathrm{d}x_n f(x_n) \int^{x_n+c} \mathrm{d}x_{n-1} f(x_{n-1}) \int^{x_{n-1}+c} \ldots \int^{F(x_3+c)} \mathrm{d}u u^{\alpha} \\ &= \frac{1}{\alpha+1} \int \mathrm{d}x_n f(x_n) \int^{x_n+c} \mathrm{d}x_{n-1} f(x_{n-1}) \int^{x_{n-1}+c} \ldots \int^{F(x_4+c)} \mathrm{d}u u^{\alpha(\alpha+1)} \\ &= \cdots = \prod_{l=1}^{n-1} \frac{1}{\sum_{i=0}^{l} \alpha^i}. \end{aligned} \tag{3.20}$$

With the expression for incomplete geometric series used above, one obtains

$$P_n(c) = (1-e^{-c})^n \frac{1}{\prod_{i=1}^{n}(1-e^{-ic})} = (1-e^{-c})^n \mathcal{Z}_n, \tag{3.21}$$

where $\mathcal{Z}_n$ is the grand canonical partition sum of bosonic particles with energy levels $i = 1, \ldots, n$ at inverse temperate c, which also occurs as one limit in the integer partition problem (cf. [26,27] and references therein).
The product $\prod_{i=1}^{n}(1-\exp(-ic))$ in the denominator is also known as the $q-$Pochhammer symbol $(q:q)_n$ with $q = e^{-c}$, which in the limit $N \to \infty$ for fixed c has the asymptotics [148]

$$\lim_{n\to\infty} \prod_{i=1}^{n} \left(1-e^{-ic}\right) \equiv \left(e^{-c}\right)_{\infty} \approx \sqrt{\frac{2\pi}{c}} \exp\left(-\frac{\pi}{6c} + \frac{c}{24}\right). \tag{3.22}$$

Inserting this expression into eq (3.21) gives the approximation

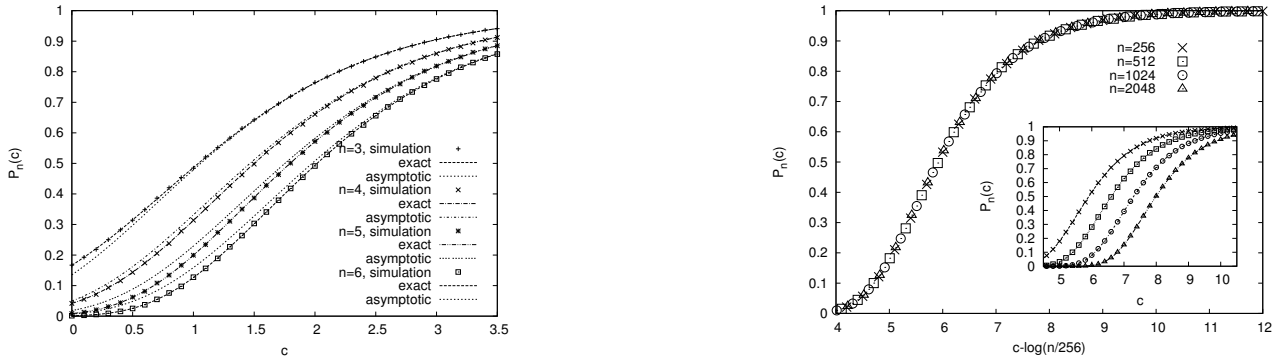


Figure 3.2.: The left panel shows a comparison of the asymptotic expression (3.24) and the exact expression (3.21) to numerical simulations of the problem. Note that the exact expression matches the data, while the asymptotic expression is slightly off. The right panel shows that for larger n, the asymptotic expression is indistinguishable from the data (inset). The main plot on the right confirms the scaling behavior $P_n(c)$ found above.

$$P_n(c) \approx \sqrt{\frac{c}{2\pi}} \exp\left[n \ln\left(1 - e^{-c}\right) + \frac{\pi}{6c} - \frac{c}{24}\right] \text{ for } n \gg 1. \tag{3.23}$$

The asymptotic for large c and fixed n is easier to obtain. Since then $\alpha = e^{-c} \ll 1$, each of the sums in the denominator can be expanded to first order in α to yield $\left(\sum_{i=1}^{l} \alpha^i\right)^{-1} \approx 1 - \alpha + \mathcal{O}(\alpha^2)$ and thus

$$P_n(c) \approx \exp\left[-(n-1)\alpha\right] = \exp\left[-(n-1)e^{-c}\right]. \tag{3.24}$$

This asymptotic expression is distinguishable from the exact expression (3.21) only in the region $c \sim \mathcal{O}(1)$ as can be seen in fig 3.2, where also the scaling behavior $P_n(c) = P(ne^{-c})$ is shown.

Comparing eqs (3.16) and (3.21), one observes the same connection between $p_n(c)$ and $P_n(c)$ as for the iid case in eq (3.19). This is a consequence of the fact that in both cases, record events are stochastically independent [105]. This is not generally true for the LDM, as will become evident in section 3.3. First however, the effect of small[3], positive drift on the ordering probability $P_n(c)$ for arbitrary probability density $f(\cdot)$ will be studied.

General expansion for small drift

The effect of positive a drift on $P_n(c)$ for arbitrary f can be studied by a first order series expansion of eq (3.17) in c around $c = 0$,

$$P_n(c) = P_n(0) + c \left.\frac{\mathrm{d}}{\mathrm{d}c} P_n(c)\right|_{c=0} + \mathcal{O}(c^2). \tag{3.25}$$

[3]Clearly c has to be measured in units of the standard deviation σ of the parental distribution $f(\cdot)$ or, if σ does not exist, in units of some other scale set by $f(\cdot)$, otherwise 'small c' has no meaning. Throughout this thesis, c will be measured in units of the scale set by $f(\cdot)$ and any statement such as $c \ll 1$ or $c \gg 1$ will be subject to an interpretation according to this.

The zeroth order term is simply the iid result $P_n(0) = (n!)^{-1}$. The coefficient of the first order term can be computed by noting that the differentiation with respect to c interchanges with the first integral over x_n as that integral is independent of c, cf. eq (3.17). Defining

$$\begin{aligned} P(n-1, x_n, c) &\equiv \int^{x_n+c} \mathrm{d}x_{n-1} f_{n-1}(x_{n-1}) \int^{x_{n-1}+c} \mathrm{d}x_{n-2} \ldots \int^{x_2+c} \mathrm{d}x_1 f(x_1) \\ &= \int \mathrm{d}x_{n-1} f(x_{n-1}) P(n-2, x_{n-1}, c), \end{aligned} \tag{3.26}$$

and, omitting terms in c^2 and smaller, the expansion eq (3.25) becomes

$$P_n(c) \approx \frac{1}{n!} + c \int \mathrm{d}x_n f(x_n) \left.\frac{\mathrm{d}}{\mathrm{d}c} P(n-1, x_n, c)\right|_{c=0}. \tag{3.27}$$

Now the derivative of $P(n-1, x_n, c)$ clearly obeys the recursion relation

$$\begin{aligned} \left.\frac{\mathrm{d}}{\mathrm{d}c} P(n-1, x_n, c)\right|_{c=0} =& f(x_n) P(n-2, x_n, 0) \\ &+ \int^{x_n} \mathrm{d}x_{n-1} f(x_{n-1}) \left.\frac{\mathrm{d}}{\mathrm{d}c} P(n-2, x_{n-1}, c)\right|_{c=0}. \end{aligned} \tag{3.28}$$

$P(n-2, x_n, c=0)$ can be evaluated by the same means as the ordering probability for the iid case in eq (3.18) to yield

$$P(n-1, x_n, c=0) = \frac{1}{(n-2)!} F^{n-2}(x_n) \tag{3.29}$$

and therefore

$$\begin{aligned} \left.\frac{\mathrm{d}}{\mathrm{d}c} P_n(n-1, x_n, c)\right|_{c=0} =& \frac{f(x_n)}{(n-2)!} F^{n-2}(x_n) \\ &+ \int^{x_n+c} \mathrm{d}x_{n-1} f(x_{n-1}) \left.\frac{\mathrm{d}}{\mathrm{d}c} P(n-2, x_{n-1}, c)\right|_{c=0}. \end{aligned} \tag{3.30}$$

Applying eq (3.28) $n-1$ times, the first order coefficient can be expanded into a sum of $n-1$ terms, each of which is a chain of integrals of lengths varying between 1 and $n-1$ and, noting that $P(1, x_2, c) = \int^{x_2+c} \mathrm{d}x_1 f(x_1) = F(x_2 + c)$ and thus $\left.\frac{\mathrm{d}}{\mathrm{d}c} P(1, x_2, c)\right|_{c=0} = f(x_2)$, this coefficient reads

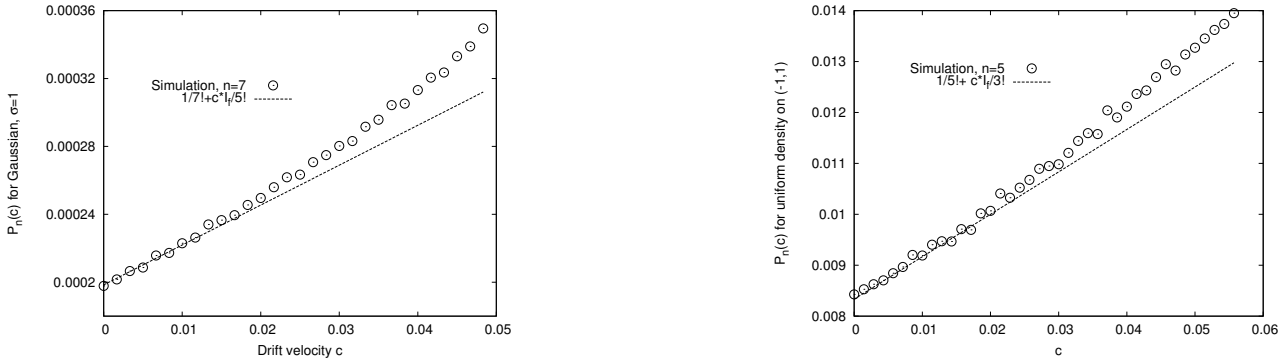


Figure 3.3.: Simulations comparing the ordering probability $P_n(c)$ to the asymptotic expansion eq(3.32) with $I_f = \int \mathrm{d}x f^2(x)$: For $f(\cdot)$ a Gaussian density with $\sigma^2 = 1$ (left) and for a uniform density on $(0, 1)$ (right).

$$\begin{aligned}\left.\frac{\mathrm{d}}{\mathrm{d}c}P_n(c)\right|_{c=0} =& \frac{1}{(n-2)!}\int \mathrm{d}f^2(x_n)F^{n-2}(x_n) \\ &+ \frac{1}{(n-3)!}\int \mathrm{d}x_n f(x_n)\int^{x_n} \mathrm{d}x_{n-1}f^2(x_{n-1})F^{n-3}(x_{n-1}) + \dots \\ &+ \frac{1}{0!}\int \mathrm{d}x_n f(x_n)\int^{x_n} \mathrm{d}x_{n-1}f(x_{n-1})\dots\int^{x_2} \mathrm{d}x_1 f^2(x_1). \end{aligned} \tag{3.31}$$

Interestingly, this sum can be performed in closed form to give the simple result (see [52] or Appendix B for details) $\left.\frac{\mathrm{d}}{\mathrm{d}c}P_n(c)\right|_{c=0} = \frac{1}{(n-2)!}\int \mathrm{d}x f^2(x)$.

With this expression, one finally obtains for the expansion eq (3.25)

$$P_n(c) \approx \frac{1}{n!} + \frac{c}{(n-2)!}\int \mathrm{d}x f^2(x). \tag{3.32}$$

As will be shown later, this result has important consequences for the study of fitness landscapes. Note that the probability density of the iid part of the RVs only enters as a non-universal constant[4] $\int \mathrm{d}x f^2(x) > 0$, but the scaling of $P_n(c)$ with c remains universal. Thus the complete universality observed in the iid case eq (3.18) remains up to a multiplicative constant and is thus only broken in a very weak sense.

As a check, this expression is compared to numerical simulations in fig 3.3.

Ordering probability for large drift

In the case of infinitely large drift, trivially $P_n(c) \to 1$ with probability one. For large but finite drift, however, an asymptotic expression for $P_n(c)$ can be derived. Ballerini and Resnick showed that the record rate $p_n(c)$ approaches a ($f(\cdot)$-dependent) constant $p(c) > 0$ as $n \to \infty$, [10, 11]. Numerical studies [52] showed that this limit is attained faster for larger c, thus in the case

[4]For some choices of f, this integral might not exist. In that case the expansion is of course not valid, see e.g. the discussion in [150] or [130].

$c \gg 1$, one can assume that it is to good accuracy attained from the very beginning. Since $c \gg 1$, the probability densities of any two RVs x_{n-k} and x_n with $k \geq 2$ have next to no overlap. Thus only the RV x_{n-1} has an appreciable chance of being greater than its successor x_n and the probability that x_n is *not* a record can be approximated by the probability that $x_{n-1} > x_n$,

$$\text{Prob}[x_n \text{ not record}] \approx \text{Prob}(x_{n-1} > x_n) = \int_c^\infty \mathrm{d}x f^{*2}(x) \equiv \epsilon(c) \tag{3.33}$$

where $f^{*2}(x)$ denotes the twofold convolution of the probability density $f(x)$ of the iid part of the RVs. Note that in the last two steps of eq (3.33), the absolute position n within the time series did not appear any more, reflecting the fact that the large-n regime of $p_n(c)$ is attained to good accuracy.
To obtain an expression for the ordering probability $P_n(c)$, one can make the assumption that in the limit considered here, record events are stochastically independent (a detailed study of the correlations between record events presented in the next section justifies this assumption, but for the moment it is just an ansatz). Then the ordering probability factorizes into the individual record probabilities,

$$P_n(c) \approx p(c)^n = (1 - \epsilon(c))^n \approx \exp[-n\epsilon(c)], \tag{3.34}$$

where $\epsilon(c)$ is the probability that x_n is *not* a record defined in eq (3.33). Now the form of $P_n(c)$ depends on $\epsilon(c)$. To quote a few examples,

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \Rightarrow \epsilon(c) = \frac{1}{2}\text{erfc}(c/2) \approx \frac{1}{c\sqrt{\pi}} e^{-c^2/4} \tag{3.35}$$

$$f(x) = \frac{1}{2} e^{-|x|} \Rightarrow \epsilon(c) = \left(\frac{1}{2} + \frac{c}{4}\right) e^{-c} \tag{3.36}$$

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \mathrm{d}k e^{-ikx-|k|^\mu} \Rightarrow \epsilon(c) = \frac{1}{2\pi} \int_c^\infty \mathrm{d}x \int_{-\infty}^{\infty} \mathrm{d}k e^{-ikx-2|k|^\mu} \approx \gamma_\mu c^{-\mu} \tag{3.37}$$

with

$$\gamma_\mu \equiv \frac{2\Gamma(1+\mu)\sin(\pi\mu/2)}{\pi\mu}, \tag{3.38}$$

where $\Gamma(\cdot)$ is the standard Gamma function. The expressions quoted above are either readily available ([2] for eq (3.35)) or straightforward to compute, see appendix B. These expressions compare well with numerical simulations as can be seen in figs 3.4 and 3.5.

The first two examples are from the Gumbel class of extreme value statistics while the third example, the Lévy stable distribution, is from the Fréchet class. To some extent these universality classes, which were defined with respect to the extreme *values* of iid RVs seem to apply here as well: Putting the corresponding expressions for $\epsilon(c)$ back into eq (3.34), one observes that for the two representatives of the Gumbel class, $P_n(c) \sim \exp\left(-ne^{-c^2/4}/(c\sqrt{\pi})\right)$ or $P_n(c) \sim \exp\left(-nce^{-c}/4\right)$ while for the example from the Fréchet class, $P_n(c) \sim \exp\left(-nc^\mu\right)$ similar to eqs (3.7) and (3.8).

The fact that the parameter c plays a role vaguely similar to the rescaled extreme value in

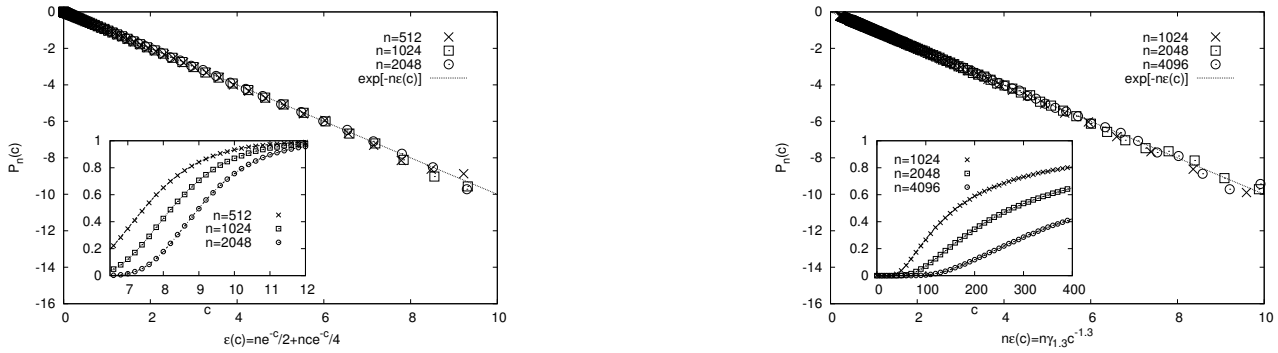


Figure 3.4.: The main plot in both figures shows a scaling collapses of the numerically obtained data (points) to the asymptotic expressions of $P_n(c)$ as function of $n\epsilon(c)$ for two-sided exponential (also called Laplace) density on the left and Lévy density with parameter $\mu = 1.3$ on the right. The expression $\epsilon(c)$ used are the ones obtained in eqs (3.36) and (3.37) respectively. The fact that the simulations collapse for different values of n shows that the approximations were justified and the asymptotic expressions for $P_n(c)$ check to good accuracy. The inset shows $P_n(c)$ as function of c to show how the different scaling in c leads to different shapes of $P_n(c)$.

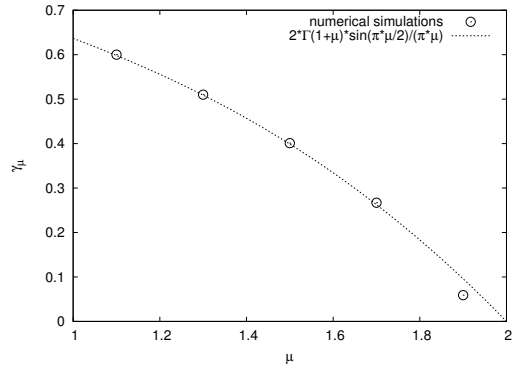


Figure 3.5.: To check the asymptotic expression derived for the Lévy distribution to more accuracy, the expression for γ_μ is compared to numerical simulations. For $n = 1024$, the numerical simulations were fitted to the form $\exp(-n\gamma c^{-\mu})$ via the value of γ, which was recorded and is represented by the circles. It matches the predicted expression (dashed lines).

the EVT context is consistent with the fact that c sets a scale for the values that are to become records. This also allows for a heuristic explanation why the complete universality observed in the iid case, where $P_n(c=0) = 1/n!$ as shown in eq (3.18), breaks down: In the iid case, there is no scale to compare to, while in the LDM case considered here, a scale is set by the value of c.
Nonetheless, this is just the interpretation of particular examples and more rigorous studies are needed to understand this breakdown of universality observed throughout this section.

3.3. Correlations between record events in the linear drift model

In deriving the asymptotic expressions for the ordering probability $P_n(c)$ in eq (3.34), it was assumed that record events are independent in the limit considered there. To explore the validity of this assumption, correlations between record events were studied. These correlations, however, are also of general interest and can be motivated with the same application that motivated the study of record values as stated in section 3.1, namely the problem of containing floods: Given that a record flood just occurred, one would intuitively believe that it will be a while before the next record flood. But is this intuition true? This question can be answered by studying correlations between record events.
As was stated in the last section, in the iid case $c = 0$ and for the Gumbel-case and arbitrary c, record events are known to be statistically independent [105]. In general, however, no such result is known and, except for one special case studied by Ballerini and Resnick [10], the problem had not received systematic attention before the work to be presented here. In the different, but related problem of records from random-walk correlated RVs [96], it is obvious that record events are not independent: For a symmetric random walk, the probability of a record in the n^{th} step, *given* that a record occurred in the $(n-1)^{th}$ step is simply $1/2$, since the random walk either goes beyond the current value, or does not, with equal probability. On the other hand, the unconditioned probability of a record event in the n^{th} step was found to be $p_n \approx 1/\sqrt{n\pi}$ [96] for large n.
To investigate the correlations between record events from the LDM defined in eq (3.10), consider the joint probability of two record events at a given distance k

$$p_{n,n-k}(c) \equiv \text{Prob}[x_n \text{ and } x_{n-k} \text{ records}]. \tag{3.39}$$

Then to compare this probability to the unconditioned record rate $p_n(c)$ as defined in eq (3.12), $p_{n,n-k}(c)$ is normalized by $p_n(c)p_{n-k}(c)$. Hence the object used to quantify these correlations is

$$l_{n,n-k}(c) \equiv \frac{p_{n,n-k}(c)}{p_n(c)p_{n-k}(c)}. \tag{3.40}$$

By the definition of the conditional probability (see e.g. [128])

$$l_{n,n-k}(c) = \frac{\text{Prob}[x_n \text{ record } | x_{n-k} \text{ record}]}{\text{Prob}[x_n \text{ record }]}. \tag{3.41}$$

Thus if record events are stochastically independent, $l_{n,n-k}(c) = 1$. If on the other hand

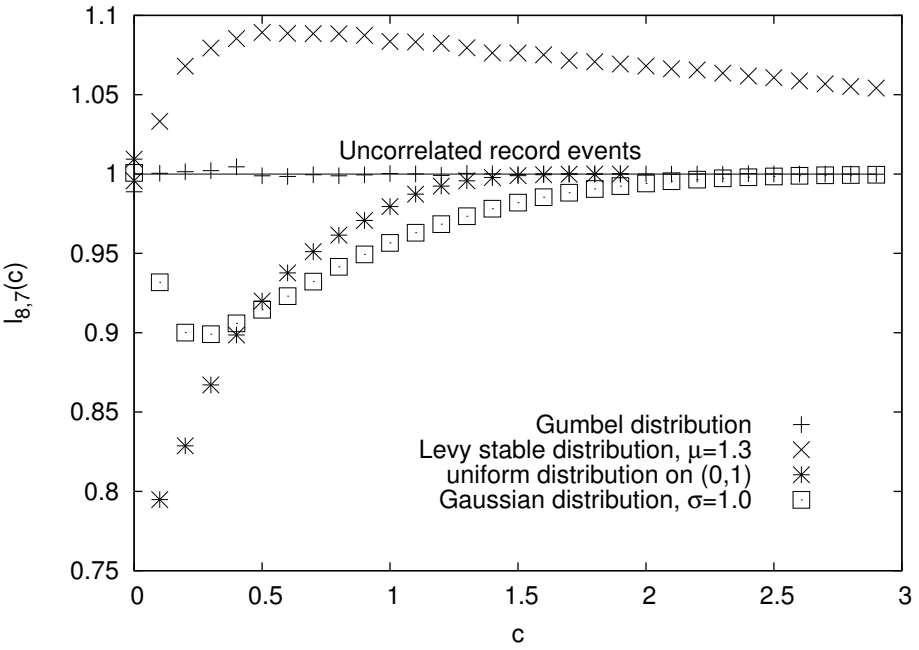


Figure 3.6.: Numerical simulations of $l_{n,n-1}(c)$ for several distributions. Note that the curve for the Gumbel distribution remains unity, recovering the known result that in this case, record events are independent. For the symmetric Lévy-stable distribution, record events attract.

$l_{n,n-k}(c) \neq 1$, then record events cannot be independent. In particular, if $l_{n,n-k}(c) < 1$, it is *less* likely to have a record in the n^{th} position if a record occurred k positions earlier than without knowledge of previous records, while if $l_{n,n-k}(c) > 1$, it is *more* likely to have a record in the n^{th} position if a record occurred in position $n-k$. In the former case, record events are said to *repel* each other, in the latter, they *attract*. To illustrate this, consider the numerical simulations presented in fig 3.6, where for $k = 1$, $l_{n,n-k}(c)$ was observed for different distributions of the iid part. There are several notable features: First of all, the curves all meet at unity for $c = 0$, which corresponds to the fact that in the iid case, record events are independent. Furthermore, as $c \to \infty$, the curves asymptotically approach unity again[5], which was expected because in this limit, $p_{n,n-k}(c)$, $p_n(c)$ and $p_{n-k}(c)$ trivially approach unity. However between these two limiting values of c, the curves of $l_{n,n-k}(c)$ show quite different features: While for the uniform distribution, record events seem to repel each other quite strongly, for the Lévy-distribution, they attract. The fact that the curves deviate significantly from unity indicates that record events in the LDM are in general not independent.
The differences between the curves and in particular the attraction between record events observed in the Lévy case seem surprising. Explaining how this difference comes about and characterizing it in terms of the universality classes of EVT introduced in section 3.1 is the main goal of this section.

[5]This explains why the assumption about stochastically independent record events yielded such precise estimates for the ordering probability $P_n(c)$ in the last section.

3.3.1. General theory

Similar to the expressions for $p_n(c)$ and $P_n(c)$ derived in eqs (3.14) and (3.17) respectively, an integral equation for $p_{n,n-k}(c)$ can formally be set up. For $k = 1$, it reads

$$p_{n,n-1}(c) = \int \mathrm{d}x_n f(x_n) \int^{x_n+c} \mathrm{d}x_{n-1} f(x_{n-1}) \prod_{l=1}^{n-2} F(x + cl). \tag{3.42}$$

This equation, together with eq (3.14), gives an expression for $l_{n,n-1}(c)$ which can be examined perturbatively for small c to determine the direction with which $l_{n,n-1}(c)$ departs from unity. Before doing so, however, stochastic independence for the iid and the Gumbel case will explicitly be checked.

Stochastic independence for two cases

For the iid case, one can again use the substitution $u = F(x)$ with $\mathrm{d}u = f(x)\mathrm{d}x$ to solve eq (3.42) to

$$p_{n,n-1}(c = 0) = \int_0^1 \mathrm{d}u \int_0^u \mathrm{d}u' u'^{n-2} = \frac{1}{n-1}\frac{1}{n} = p_n(0)p_{n-1}(0). \tag{3.43}$$

Clearly this factorization also holds under insertion of spacings between the record events, thus $p_{n,n-k}(c = 0) = p_n(c = 0)p_{n-k}(c = 0)$. It is also possible to consider more than two record events. This factorization property implies the known [7, 105] result on the independence of record events in the iid case.
For the Gumbel case, one can again use the Gumbel formula eq (3.15) to compute $p_{n,n-1}(c)$. With $\alpha = e^{-c}$ and the same substitution as above, it reads

$$\begin{aligned} p_{n,n-1}(c) &= \int \mathrm{d}x_n f(x_n) \int^{x_n+c} \mathrm{d}x_{n-1} f(x_{n-1}) \prod_{l=1}^{n-2} F(x)^{\alpha^l} = \int_0^1 \mathrm{d}u \int_0^{u^\alpha} du' u'^{\sum_{l=1}^{n-2}\alpha^l} \\ &= \frac{1}{\sum_{l=0}^{n-2}\alpha^l} \int_0^1 \mathrm{d}u u^{\sum_{l=1}^{n-1}\alpha^l} = \frac{1}{\sum_{l=0}^{n-2}\alpha^l}\frac{1}{\sum_{l=0}^{n-1}\alpha^l} \\ &= p_{n-1}(c)p_n(c). \end{aligned} \tag{3.44}$$

This factorization again holds under insertion of an arbitrary number of arbitrarily spaced record events, thus one recovers the known stochastic independence of record events. Of course, this was expected in the light of eq (3.20).

Expansion for small drift

For the iid case, $l_{n,n-1}(c = 0) = 1$, thus allowing a perturbative study for small c around $c = 0$ by the same approach used in the previous section for $P_n(c)$. Assuming that $p_n(c)$ is analytic in c, a first-order Taylor expansion can be set up as in eq (3.25).
A straightforward computation, ignoring terms in c^2 or higher then yields (see appendix B for

this and all other computations omitted in the section)

$$p_n(c) \approx \frac{1}{n} + c\frac{n(n-1)}{2} I(n-2) \tag{3.45}$$

with the definition[6]

$$I(n) \equiv \int \mathrm{d}x f^2(x) F^n(x). \tag{3.46}$$

For $c = 0$, the expression given in eq (3.13) is recovered from eq (3.45). An analogous expansion of $p_{n,n-1}(c)$ around $c = 0$ yields

$$p_{n,n-1}(c) \approx \frac{1}{n(n-1)} - c\frac{(n-1)(n-2)-2}{2} I(n-2) + c\frac{(n-1)(n-2)}{2} I(n-3) \tag{3.47}$$

with the expression for the term independent of c taken from eq (3.43). One sees that there are terms of the same order in c and n but with different signs. Which one dominates will now depend on $I(n)$ and hence on the underlying distribution $f(\cdot)$ as numerically observed in fig 3.6. For $l_{n,n-1}(c)$, the expansion reads

$$l_{n,n-1}(c) \approx 1 + cJ(n) \tag{3.48}$$

with

$$\begin{aligned} J(n) =& n(n-1)I(n-2) + \frac{n(n-1)^2(n-2)}{2}\left[I(n-3) - I(n-2)\right] \\ &- \frac{n^2(n-1)}{2} I(n-2) - \frac{(n-1)^2(n-2)}{2} I(n-3). \end{aligned} \tag{3.49}$$

Again there are terms of both positive and negative sign.
Because of the complicated structure of these expansions and in particular of the integral $I(n)$, an analysis of these expansions is most convenient in the limit of large n, which will be considered for the rest of this section. For large n and given $c \ll 1$, one has to compare the leading order terms in n of the coefficient eq (3.49) to predict the sign with which $l_{n,n-1}(c)$ departs from unity. One notes that the term in n^4 has a positive sign but depends on the difference $I(n-3) - I(n-2)$, while the terms in n^3 have negative sign and coefficients $I(n)$. It is still not obvious how the attractive correlations and the strong dependence on the underlying distribution $f(\cdot)$ arise. Rather, one needs a detailed study of the properties of $I(n)$ and its 'derivative' $I(n-3) - I(n-2)$ for large n as functional of $f(\cdot)$, which will be provided in the next subsection. Before entering in such a discussion, however, heuristic derivations of known results on the asymptotics of $l_{n,n-k}(c)$ for large n and large k will be provided.

[6]This integral also appeared in a seemingly unrelated context [130]. For certain choices of $f(\cdot)$, this integral does not exist. The consequences is this context are clear since then the leading order term is not linear in c.

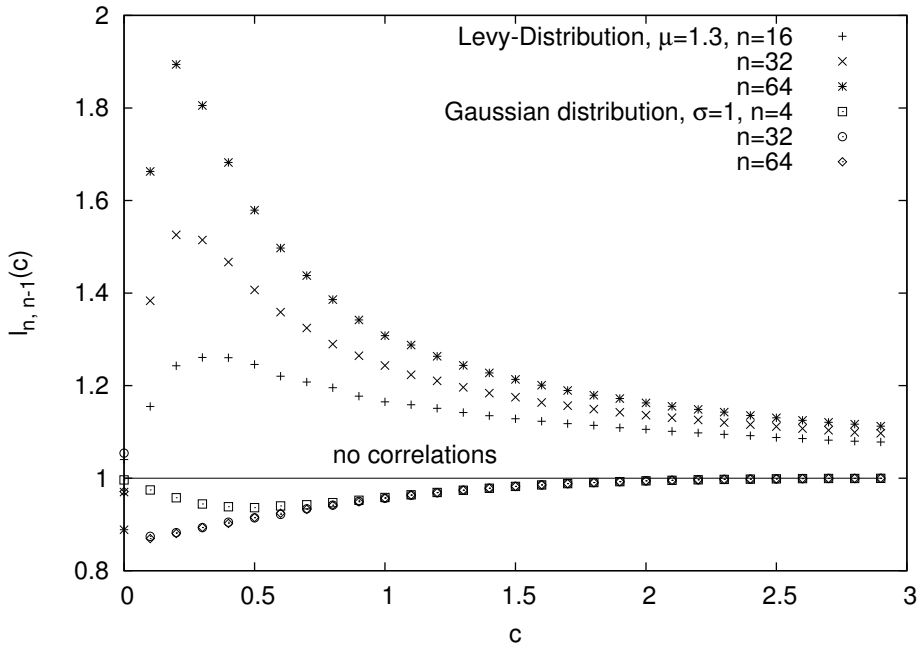


Figure 3.7.: $l_{n,n-1}(c)$ for Gaussian (standard deviation $\sigma = 1$) and symmetric Lévy stable (tail parameter $\mu = 1.3$) distributions of the iid part. As n increases, the attractive correlations in the Lévy case become more pronounced. In the Gaussian case, on the other hand, the correlations eventually saturate to a non-trivial limiting curve which so far has eluded a detailed description.

Asymptotics for large n and k

Fig 3.7 shows $l_{n,n-1}(c)$ for increasing values of n and two different distribution. These numerical simulations, similar to those shown in fig 3.6 show that the deviations from unity of the correlations are not a finite size effect. For the Gaussian distribution, the correlations saturate to a non-trivial limit curve for intermediate c, while for the Lévy-distribution, they increase in the regime numerically explored so far.

However it can be shown that for any $f(\cdot)$ with finite first moment, a limiting function exists. The existence of this limiting function $l(c) \equiv \lim_{n\to\infty} l_{n,n-1}(c)$ greater than zero is a direct consequence of the proof for the existence of $0 < p(c) \equiv \lim_{n\to\infty} p_n(c)$ as given by Ballerini and Resnick [10]: By inspection of eqs (3.14) and (3.42) it is clear that both $p(c)$ and $\lim_{n\to\infty} p_{n,n-1}(c)$ exist and are greater than zero if

$$G_c(x) \equiv \lim_{n\to\infty} \prod_{j=1}^{n} F(x + cj) \tag{3.50}$$

exists and is greater than zero, which was proven in [10]. For completeness, heuristic arguments that $G_c(x)$ and thus $\lim_{n\to\infty} l_{n,n-1}(c)$ are positive will be provided here [150]. Taking the logarithm of eq (3.50), the question $G_c(x) > 0$ becomes equivalent to the convergence of the sum

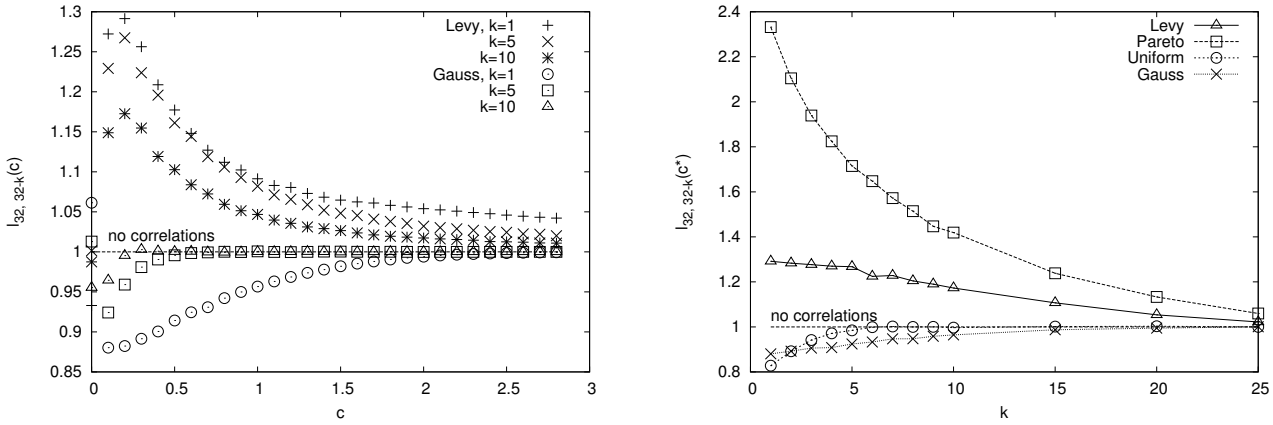


Figure 3.8.: Illustration of the behavior of the correlations as given by $l_{32,32-k}(c)$ as k increases. The left figure shows, analogous to fig 3.7, standard Gaussian ($\sigma = 1$) and symmetric Lévy stable distributions ($\mu = 1.3$) for increasing k. The right figure explores the behavior of the correlations at the value of the drift velocity c^* chosen such that deviation from unity is maximal. The values are $c^* = 0.1$ for the Gaussian case and the uniform distribution on $(0, 1)$, $c^* = 0.2$ for the Lévy-stable distribution $\mu = 1.3$ and $c^* = 0.4$ for the Pareto distribution with tail parameter $\mu = 2.0$. All curves approach unity as k grows.

$$\ln\left[G_c(x)\right] = \sum_{j=0}^{\infty} \ln\left[F(x + cj)\right]. \tag{3.51}$$

Since $F(\cdot)$ is the cumulative probability distribution of the RV X,

$$F(x + cj) = \text{Prob}\left[X \leq x + cj\right]) = 1 - \text{Prob}\left[X > x + cj\right] \tag{3.52}$$

and thus for any finite x and any $c > 0$ there is a $\tilde{j}$ large enough such that $\text{Prob}\left[X > x + cj\right] \ll 1$. With the common expansion of the logarithm $\ln(1 - y) \approx -y$ for sufficiently small $|y|$,

$$\ln\left[G_c(x)\right] \approx \sum_{j=1}^{\tilde{j}-1} \ln\left[F(x + cj)\right] - \sum_{j=\tilde{j}}^{\infty} \text{Prob}\left[X > x + cj\right]. \tag{3.53}$$

The first sum is a finite sum over finite summands and thus converges. The second, infinite sum converges whenever $\text{Prob}[X > x]$ decays faster than $1/x$, which is also the condition for the first moment of the underlying probability density $f(\cdot)$ to exists. This confirms the criterion for existence of a limiting function $p(c) > 0$ as stated by Ballerini and Resnick, namely that the distribution of the iid part of the RVs from the LDM must have finite first moment. Note, however, that such a limit function only exists for c strictly greater than zero.

Finally consider the effect of large distance k between record events. So far, only the case $k = 1$ was considered. Numerical simulations presented in fig 3.8 show that correlations decrease in magnitude with k, thus in focusing on $k = 1$, the case of most pronounced correlations was

considered. This also matches with a rigorous result [10] on the joint probability $p_{n,n-k}(c)$ of the two records in the event n and $n-k$ which states that

$$\lim_{k\to\infty}\lim_{n\to\infty} p_{n,n-k}(c) = \left[\lim_{n\to\infty} p_n(c)\right]^2 = p(c)^2. \tag{3.54}$$

3.3.2. Explicit examples

While the properties of $l_{n,n-1}(c)$ for a wide range of distributions has been discussed (see [52,150] and also implicitly [130], since the integral $I(n)$ considered in eqs (3.45), (3.47) and (3.48) also appears in the seemingly unrelated context considered there), here only a small selection of these results will be presented. The aim is to motivate a unified picture given at the end of this subsection that will allow a rough (and non-rigorous) classification of the sign of correlations according to the three universality classes of extreme value theory discussed in section 3.1.

The Weibull class

As a first example, consider the case of $f(x) = \frac{1}{2a}\mathbb{1}_{(-a,a)}(x)$. A straightforward computation yields $I(n) = (2a(n+1))^{-1}$ and thus

$$l_{n,n-1}(c) \approx 1 + \frac{c}{2a}\left(n + \frac{n(n-1)}{2} - \frac{n^2}{2} - \frac{(n-1)^2}{2}\right) \approx 1 - \frac{c}{4a}n^2 \tag{3.55}$$

for $n \gg 1$.
As a more general representative of the Weibull class, consider the Kumaraswamy distribution with probability density

$$f_\zeta = \zeta(1-x)^{\zeta-1} \tag{3.56}$$

for $\zeta > 0$ and $x \in [0,1]$ with cumulative distribution function $F_\zeta(x) = 1-(1-x)^\zeta$, which contains the case of the uniform distribution discussed above for $\zeta = 1$. With the substitution $(1-x) = z^{1/\zeta}$, for $\zeta > 1/2$ the integral $I_\zeta(n)$ is given by

$$I_\zeta(n) = \zeta B\left(2 - \frac{1}{\zeta}, n+1\right) = \zeta\frac{\Gamma\left(2-\frac{1}{\zeta}\right)\Gamma(n+1)}{\Gamma\left(n+3-\frac{1}{\zeta}\right)} \tag{3.57}$$

with $B(\cdot,\cdot)$ and $\Gamma(\cdot)$ the standard Beta and Gamma functions, respectively [2]. Note that the condition $\zeta > 1/2$ is necessary because it ensures that the first argument of the Beta function is larger than zero. Otherwise the integral $I(n)$ does not exist.
Then, since for any $z \in \mathbb{C}$, $\Gamma(z+1) = z\Gamma(z)$, the second term on the right hand side of eq (3.49) is proportional to

$$
\begin{aligned}
I_\zeta(n-3) - I_\zeta(n-2) &= \zeta\Gamma\left(2-\frac{1}{\zeta}\right)\left[\frac{\Gamma(n-2)}{\Gamma\left(n-\frac{1}{\zeta}\right)} - \frac{\Gamma(n-1)}{\Gamma\left(n+1-\frac{1}{\zeta}\right)}\right] \\
&= \zeta\Gamma\left(2-\frac{1}{\zeta}\right)\left[\frac{\Gamma(n-2)}{\Gamma\left(n-\frac{1}{\zeta}\right)} - \frac{(n-2)\Gamma(n-2)}{\left(n-\frac{1}{\zeta}\right)\Gamma\left(n-\frac{1}{\zeta}\right)}\right] \\
&= I_\zeta(n-2)\frac{2-\zeta^{-1}}{n-\zeta^{-1}}. \qquad (3.58)
\end{aligned}
$$

Thus for n large enough that $n \approx n-1 \approx n-2$ and $B(\cdot, n+1) \approx B(\cdot, n) \approx B(\cdot, n-1)$ and ζ such that $\zeta^{-1} \sim \mathcal{O}(1)$ or smaller (which is guaranteed by the requirement $\zeta > 1/2$), the first order coefficient of eq (3.48) becomes

$$
\begin{aligned}
J_\zeta(n) &\approx n^2 I_\zeta(n) + \left(2-\frac{1}{\zeta}\right)\frac{n^3}{2}I_\zeta(n) - 2\frac{n^3}{2}I_\zeta(n) \approx -\frac{n^3}{2}B\left(2-\frac{1}{\zeta}, n\right) \\
&\approx -\frac{\Gamma\left(2-\frac{1}{\zeta}\right)}{2}n^{1+\zeta^{-1}} \qquad (3.59)
\end{aligned}
$$

to leading order in n, where the expression for $I_\zeta(n)$ in eq (3.57) and the asymptotic expression $B(z,n) \sim \Gamma(z)n^{-z}$ for fixed z and large n were used [2].

3.3.3. The Gumbel class

In this context the simplest representative of the Gumbel class is the one-sided exponential distribution with density $f(x) = a^{-1}\exp\left(-\frac{x}{a}\right)$ and cumulative distribution function $F(x) = 1-\exp(-x/a)$ for $x \geq 0$. With the substitution $t = e^{-x}$, the integral from eq (3.46) is given by

$$
I_a(n) = \frac{1}{a}B(2, n+1) = \frac{1}{a(n+1)(n+2)} \qquad (3.60)
$$

and the leading order coefficient from eq (3.49) is straightforward to compute as $J(n) = \frac{1}{2a}$ and thus eq (3.48) reads

$$
l_{n.n-1}(c) \approx 1 + \frac{c}{2a} \qquad (3.61)
$$

for all n. Note that the sign of the leading order term is positive, indicating attractive record events.

For the Gaussian distribution (with second moment σ), another important member of the Gumbel class, computations are a bit more complicated and were performed in [52,150]. There, asymptotic expressions for the integral $I(n)$ for large n were derived by means of a saddle point approximation. This way, the integral (3.46) can be computed as

$$I(n) \approx \frac{1}{n^2\sigma}\frac{4\sqrt{\pi}}{e^2}\sqrt{\ln\left(\frac{n^2}{8\pi}\right)} \tag{3.62}$$

giving a leading order expansion of $l_{n,n-1}(c)$ of the form

$$(l_{n,n-1}(c)-1) \propto -n\frac{c}{\sigma}\sqrt{\ln\left(\frac{n^2}{8\pi}\right)} \tag{3.63}$$

valid for $n\sigma \ll c$. In contrast to the exponential case discussed above, here the sign is negative. For a more general example of the Gumbel class, consider the generalized Gaussian distribution

$$f(x) = A_\beta \exp[-|x|^\beta] \tag{3.64}$$

with normalization constant A_β and $\beta > 0$. In [52] it was shown that

$$I(n) \propto n^{-2}\ln(n)^{1-\beta^{-1}} \tag{3.65}$$

and [150]

$$J(n) \propto -d_1\left(1-\frac{1}{\beta}\right)n\ln(n)^{1-\beta^{-1}} + d_2\ln(n)^{1-\beta^{-1}} \tag{3.66}$$

with d_1, d_2 real positive constants independent of n. From eq (3.66) it is clear that for $\beta \neq 1$, the negative term dominates for $n \gg 1$ while for $\beta = 1$, corresponding to the exponential distribution, the positive term dominates since the negative term vanishes. Also, since in this case $\ln(n)$ is raised to the power 0, the correlations become in this case to leading order independent of n. For $\beta \leq 1$, i.e. for a probability density decaying like a stretched exponential, $\beta^{-1} > 1$ and the correlations are positive.
Thus in the Gumbel class of distributions, both signs were observed, with the stretched exponential distribution showing attractive correlations.

3.3.4. The Fréchet class

Perhaps the easiest member of the Fréchet class to consider is the Pareto distribution with density $f_\mu(x) = \mu x^{-\mu-1}$ for $\mu > 1$ and $x \geq 1$ and distribution function $F_\mu(x) = 1 - x^{-\mu}$. Then the integral $I_\mu(n)$ from eq (3.46) can be computed as

$$I_\mu = \mu B\left(2+\frac{1}{\mu}, n+1\right) = \mu\frac{\Gamma\left(2+\frac{1}{\mu}\right)\Gamma(n+1)}{\Gamma\left(3+\frac{1}{\mu}+n\right)}. \tag{3.67}$$

Analogous to the computation in eq (3.58), the difference term is then given by

$$I_\mu(n-3) - I_\mu(n-2) = \mu\Gamma\left(2+\frac{1}{\mu}\right)\frac{\Gamma(n-2)}{\Gamma\left(1+\frac{1}{\mu}+n\right)}\left[1-\frac{n-2}{n+\mu^{-1}}\right]$$
$$= I_\mu(n-2)\frac{2+\mu^{-1}}{n+\mu^{-1}}. \tag{3.68}$$

Then for sufficiently large n, the leading order coefficient $J_\mu(n)$ can be approximated by

$$J_\mu(n) \approx n^2 I_\mu(n) + \frac{n^3}{2\mu} I_\mu(n) \approx \frac{\Gamma\left(2-\frac{1}{\mu}\right)}{2} n^{1-\mu^{-1}}, \tag{3.69}$$

again using the asymptotics for the Beta function quoted above. Note however that in contrast to eq (3.59), the equation above has positive sign.

3.3.5. Unified picture

The results for specific examples for the three universality classes of EVT can be summarized in a rather simple scaling picture [150] by expressing these universality classes in terms of the generalized Pareto distribution [118] with probability density

$$f(x) = (1+\kappa x)^{-\frac{\kappa+1}{\kappa}}. \tag{3.70}$$

For $\kappa > 0$, this distribution is of the Pareto type with $\mu = 1/\kappa$, for $\kappa < 0$ it is a Weibull-class distribution similar to the Kumaraswamy distribution given in eq (3.56), with $\zeta = -1/\kappa$. The Gumbel class is represented in the limit $\kappa \to 0$, where the exponential distribution is recovered. Using the expressions for $I(n)$ derived for the cases considered above in eqs (3.65), (3.67) and (3.57), one observes that they can be expressed as

$$I(n) \sim n^{-(2+\kappa)}. \tag{3.71}$$

This can be related to the behavior of correlations by noting that in the limit of large n, the difference term $I(n-3) - I(n-2)$ in eq (3.49) can be approximated by a derivative. Then

$$J(n) \approx -\frac{1}{2} n^4 \frac{\mathrm{d}}{\mathrm{d}n} I(n) - n^3 I(n) + \mathcal{O}(n^2 I(n)). \tag{3.72}$$

The derivative of $I(n)$ is negative, since the integral is obviously monotonically decreasing in n, and thus, with eq (3.71), the sign of the correlations is asymptotically given by the sign of κ,

$$J(n) \sim \frac{\kappa}{2} n^3 I(n). \tag{3.73}$$

Thus for large n, record events are expected to attract each other in the Fréchet class and repel each other in the Weibull class. For the Gumbel class, a refined analysis like the one given in eq (3.66) is necessary. There, correlations are expected to be negative for distributions decaying faster than exponential and positive for stretched exponentials, with the pure exponential as

a marginal case with positive correlations. This classification, albeit just based on specific examples and not shown in general, is consistent with numerical simulations.
Furthermore, for the heavy-tailed distributions considered, the correlations grow with n.

3.4. A test for heavy-tailed distributions

Since one can roughly classify the type of correlations (attractive or repulsive) by the type of probability densities $f(\cdot)$ from which the iid part of the RVs is drawn (Weibull class or heavier-than-exponential tails), some information about the parental distribution is still contained in the statistics of record events from the LDM. In this section, it will be shown that by artificially imposing a linear drift on any set of *iid* RVs, this information can partially be recovered.

3.4.1. Record based tests

Imagine a data set of N entries. Now there exist various tests to check that these data points can be treated as iid RVs. One possibility [50, 59] is to take sub-sequences of $n < N$ entries, keep these entries in the same order they have in the original data set and note if the last entry is a record. If the data set is composed of iid RVs, then the fraction of times that the last entry was a record should approach $\hat{p}_n \sim p_n = 1/n$. If this is not the case, the data set cannot be treated as iid[7]. This constitutes perhaps the simplest record based test.
Note that there was no need to formulate a hypothesis about the distribution of the RVs (to assume say, a Gaussian shape). Thus this method is said to be 'non-parametric' or 'distribution free', which is a general feature of record-based test (see [50, 59] or [62] and references therein). Another advantage is that even for a data set of moderate size N on the order of a few dozen entries, the number $\mathcal{N}$ of ordered subsets of size $n < N$ is given by

$$\mathcal{N} = \binom{N}{n} = \frac{N!}{n!(N-n)!}, \tag{3.74}$$

which can be quite large for $n \sim N/2$, thus one can use this combinatorial proliferation to obtain a reasonable statistics. However, record based tests usually do not allow to estimate the shape of the distribution. Here, the results obtained throughout this chapter will be used to devise a record based test that allows to tell whether or not a given data set is drawn from a distribution with heavy tails.

3.4.2. The test for heavy tails

Assume that a given data set $\{\eta_i\}_{i\in\{1,\dots N\}}$ of N data points consists of iid RVs. Then one can pick a subset of size $n < N$ and *artificially superimpose* a drift with velocity c to obtain a modified time series[8]

[7] It is still possible that a data set is not composed of iid entries and yet passes this test. Thus the example discussed here only works in one way, in that it can be used to *reject* the hypothesis of iid entries, but not to *confirm* it.

[8] Note that the range of numerical values of c should be chosen according to the mean spacing of the entries of the data set $\{x_i\}_{i\in\{1,\dots,N\}}$.

$$x_n = \eta_n + cn. \tag{3.75}$$

If one keeps the original data set, one can sample a large number of subsets like the one above, see eq (3.74). Note that now, since one added n units of drift to the n^{th} RV, a weak ordering has been imposed. In fact, the family of RVs in eq (3.75) now obeys LDM statistics. In particular, if one again records the relative number of times that the n^{th} RV is a record, $\hat{p}_n(c)$ and also the relative number that both entries n and $n-1$ are records, $\hat{p}_{n,n-1}(c)$ (see appendix D for details), one can probe the correlations between record events from the modified times series (3.75) in analogy to eq (3.40) by considering

$$\hat{l}_{n,n-1}(c) = \frac{\hat{p}_{n,n-1}(c)}{\hat{p}_n(c)\hat{p}_{n-1}(c)}. \tag{3.76}$$

The hat $\hat{}$ stands as a reminder that as opposed to the numerical studies in section 3.3, here only subsets of one large data set are discussed. An application of this test, known as 'heavy tail indicator' (HTI) [53], is illustrated in fig 3.9. There, two 'data sets' were created using the `gsl` random number generators for two different distributions. Was was drawn from the symmetric Lévy stable distribution with tail parameter $\mu = 1.3$, the other from a Gaussian distribution with standard deviation $\sigma = 1$. Then $\hat{l}_{n,n-1}(c)$ was computed for the HTI. The number s of subsets used, called 'internal statistics', was $s = 10^5$. The inset of fig 3.9 shows comparison of the cumulative distributions (symbols for the cumulative distribution from the data, lines for the exact expressions) allows to tell that one distributions has heavier tails than the other, but the signal is not as clear as the one presented in the main plot. Also there is no way of rejecting the hypothesis of a Gaussian with a larger σ being responsible for the wider curve. However under the HTI, increasing the standard deviation of the Gaussian produces a signal rather different from the one observed for the heavy tailed distribution. Thus using the HTI, the hypothesis that a Gaussian with a larger standard deviation is responsible for the first data set can be rejected.

In fig 3.9, the curves obtained for the HTI follow the exact numerics, but also show some degree of fluctuations. To systematically investigate the effect of these fluctuations, simulations of the HTI for a large number S of independent data sets (called 'external statistics') were performed. One notes that while for correlations greater than unity (attractive), the fluctuations can be quite substantial, for the repulsive case, they are quite small, see fig 3.10.

In the left panel of fig 3.10, two different sizes of data sets are compared ($N = 64$ versus $N = 128$) while $n = 16$ is kept fix. The fluctuations are greater for the smaller value of N. The way in which the ratio of subset size to data set size n/N affects the amount of fluctuations is investigated in the inset of fig 3.11. In particular it is shown for three different distributions that the effect of increasing internal statistics s saturates (main plot). How this saturation depends on the ratio n/N is investigated in the inset: The smaller n/N, i.e. the smaller n for N fixed, the smaller the fluctuations. Thus to minimize fluctuations, one should choose n small.

However, since in the last section, it was shown that the correlations only show their limiting behavior (attractive for heavy tails, see section 3.3) for large n, there are two contradictory constraints on n. Which values are to be chosen for a given application must be judged in each case, but for the examples presented in the next subsection, the ratio of $n/N = 1/4$ was used.

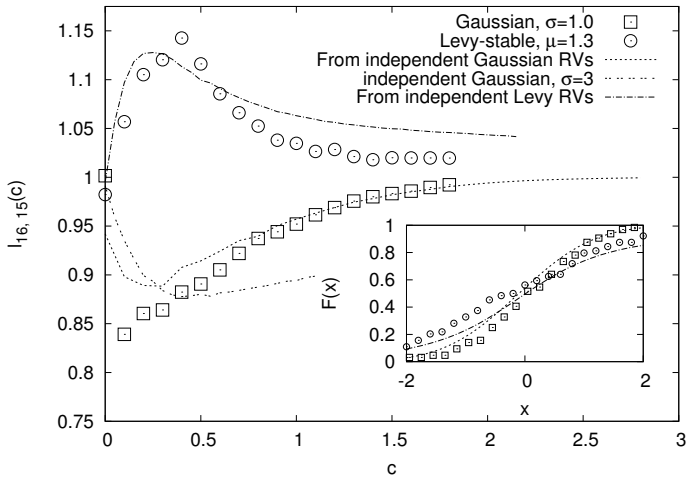


Figure 3.9.: Main plot show the application of the HTI for two data sets, one drawn from a symmetric Lévy-stable distribution with $\mu = 1.3$, the other from Gaussian distributions $\sigma = 1$ and $\sigma = 3$. The curves obtained from the HTI (symbols) closely follow the exact numerics obtained from arbitrarily many RVs (lines). The inset shows a comparison of cumulative distributions.

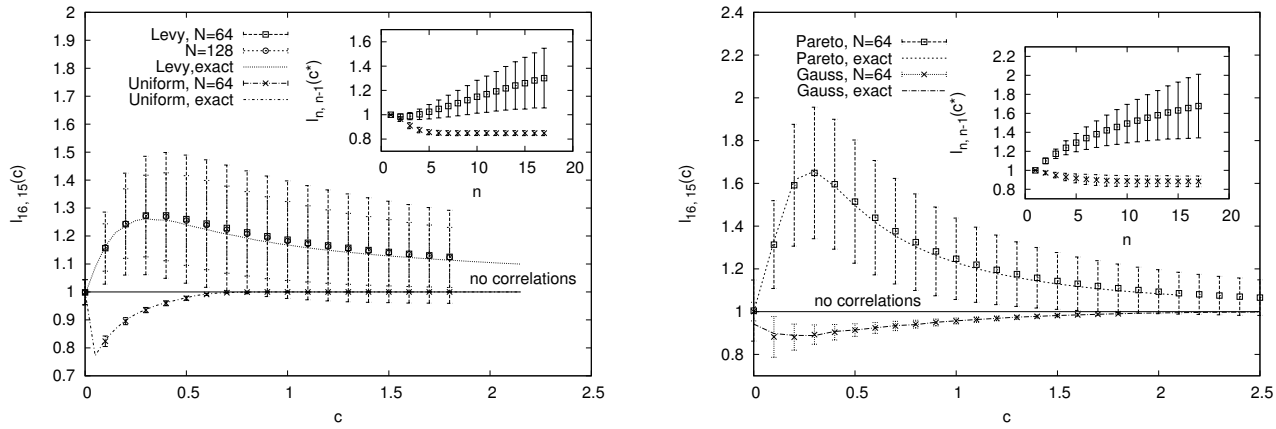


Figure 3.10.: Sample-to-sample fluctuations of $\hat{l}_{n,n-1}(c)$ obtained for different distributions. The left figure compares symmetric Lévy-stable distribution ($\mu = 1.3$) and uniform distribution on $(0, 1)$. The main plot shows that the fluctuations for the heavy tailed case are quite substantial while those for the uniform distribution hardly show. The right figure compares a Pareto-distribution ($\mu = 2.0$) to a standard Gaussian. For the Gaussian, the fluctuations are also quite small, albeit larger than for the uniform case. The insets of both figures compare the n-behavior of signal and fluctuations for the respective distributions at fixed drift velocity c^* chosen such that the deviations of the HTI from unity were maximal. In the heavy-tailed case, the signal increases with n while remaining constant in the other examples.

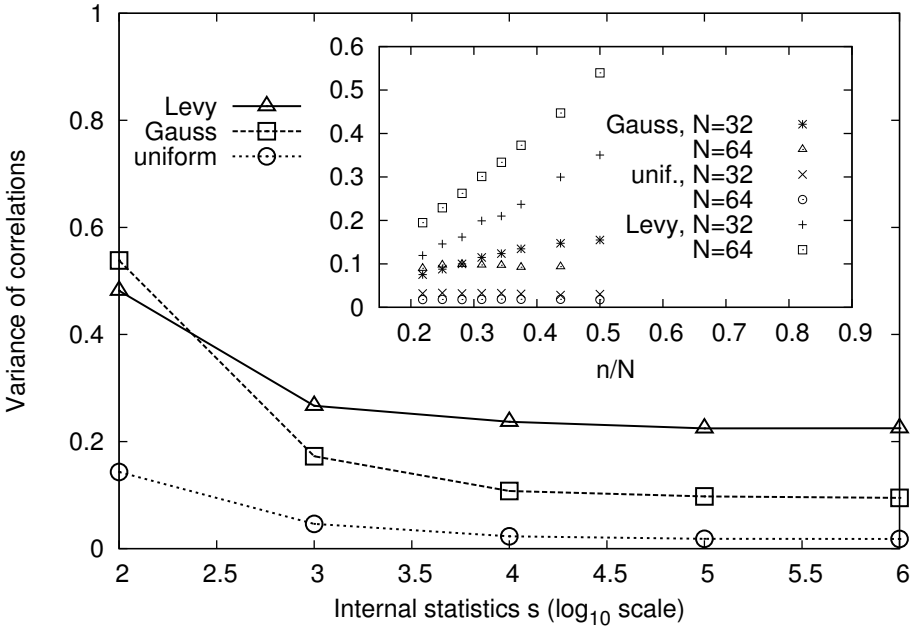


Figure 3.11.: The main plot shows that for all three distributions considered (standard Normal, Lévy-stable with $\mu = 1.3$ and uniform on $(0, 1)$), increasing the internal statistics s decreases the fluctuations measured at the value of c such that the mean signal deviates maximal from unity. Eventually however, fluctuations saturate. The value to which they saturate is given in the inset as function of n/N. The fact that in particular the two curves for the Lévy case have different slopes indicates that this is not the correct scaling combination, but here it suffices to show that in order to reduce fluctuations, n should be small.

Note however than there are still quite sizable sample-to-sample fluctuations to be expected, see fig 3.10. These fluctuations were numerically observed to be particularly strong if the underlying distribution $f(\cdot)$ had heavy tails. This means that the test presented here cannot be used to *reject* the hypothesis of heavy tails: If the HTI yields a signal below unity, this might very well be due to a fluctuation. On the other hand, the fact that fluctuations from underlying distributions without heavy tails were comparably small implies that a HTI signal strongly above unity is unlikely to be due to a fluctuation 'from below' and is therefore indicative of heavy tails underlying the data set. In this sense, the test only works in one way.

3.4.3. Examples

Identifying heavy tails underlying a time series is important because they imply a drastic increase in the probability of extreme events [22, 137]. In particular, if these heavy tails are of the power-law type ($f(x) \propto x^{-\mu}$ for large x) with exponent $\mu \leq 3$, the presence of heavy tails leads to unusual behavior of the system under consideration, e.g. super-diffusive behavior [16] or small-world effects in real-world networks [12, 107]. Other applications range from paleontology [106] to animal foraging behavior [40,125,141], where in particular the last example is regarded as quite controversial [124] due to a possible bias in the methods used for parameter estimation [39, 119] toward the tail exponent corresponding to a search strategy that has been

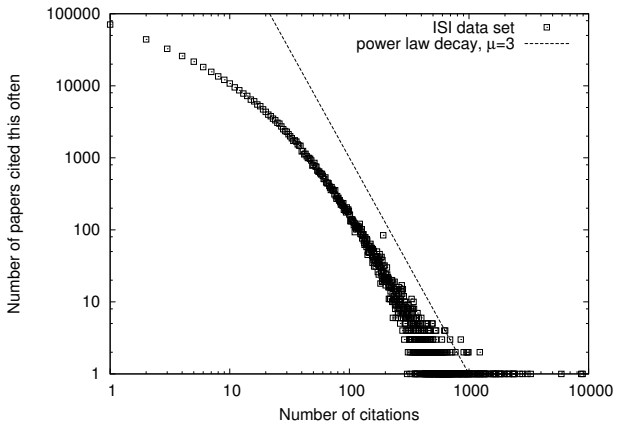


Figure 3.12.: . In this double-logarithmic plot, the citation distribution decays linearly in the number of citations. The line indicates decay like x^{-3}. Redrawn from [123], data supplied by S. Redner, [122].

proposed to be optimal [92, 121], underlining the need for distribution-free tests for heavy tails. In the light of all these examples, it is clear that there is a host of empirical data sets that are well established to possess heavy tails. To show that the test presented in this section recovers these heavy tails, the famous ISI data set on citation distributions [123] is considered: It consists of citations for each of the 783339 papers published in 1981 and cited between 1981 and June 1997 that are part of the ISI data base. Due to its size, it was possible to establish the existence of a power law tail by a conventional test, where it was found that the tail has a parameter of $\mu \approx 3$ for large numbers of citations. The existence of such a tail can also be verified by inspection, see fig 3.12. However only about 500 paper were in the regime of power-law tails, whereas most of the entries of the data set could best be fitted to a stretched-exponential decay, see [123] and references therein. Nonetheless, using the test presented here, it is possible to recover these heavy tails. As can be seen in fig 3.13, the test indeed does show attractive correlations increasing in n, thus recovering these heavy tails. However to do so, the test only required a subset of $N = 64$ entries of the original ISI data set. To show that this is not a coincidental result, three independently chosen subsets are picked from the data and evaluated, with the corresponding curves of the HTI presented in the left of fig 3.13.

Robustness under removal of outliers

One major shortcoming of conventional tests, in particular of maximum-likelihood methods, is that they are not very robust to removal of outliers [23]. If one believes such an outlier to be spurious and wishes to perform a maximum likelihood test without it, the results quickly loose statistical significance. The test presented here, however, only uses a subset of the data from the start, thus one expects that it will be more robust with respect to removal of outliers. This robustness is illustrated on the right of fig 3.13, where the HTI test was performed for data sets with the highest and second-highest entry removed. The resulting curves still show attractive correlations indicative of heavy tails, but in one case the growth-signal expected was lost. Upon closer examination of that particular data set, it turned out that the removed data point was more than a factor 10 greater than the next largest, making this data set

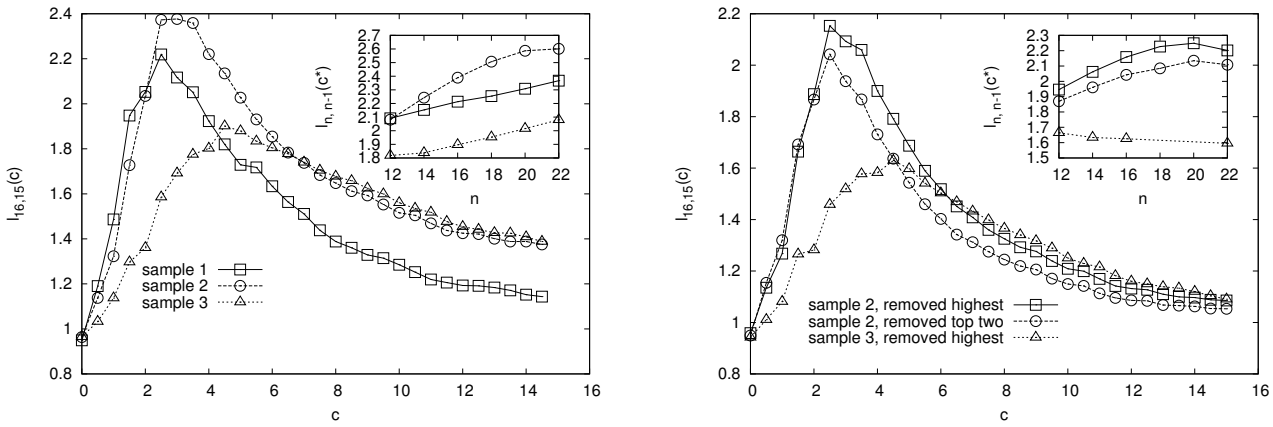


Figure 3.13.: Results of the HTI test applied to three different subsets of size $N = 64$ of the ISI data set. These sample subsets were picked uniformly at random picking 64 entries out of the 783339 of the full ISI data set. The left panel shows that the heavy tails underlying the data set can be recovered even from these small sub-samples in all three examples. The panel on the right shows that the test presented here is robust with respect to removal of outliers. The insets show the growth signal at fixed values of $c = c^*$ chosen for maximum deviation ofthe signal from unity. These values can be read off of the main plots.

somewhat pathological. The other data sets considered on the right of fig 3.13 still show positive correlations growing with n even when both the largest and second largest data point are removed. Closer examination of the data set showed that the largest entry was still about a factor of 8 greater than the second largest.

Application to the E. coli data set

As was just shown, this test is particularly useful when the data set considered is small. This is the case whenever each data point is quite costly or hard to obtain, which is true in studying the fitness effects of mutations, where each mutation corresponding to one entry of the data set must be created (see e.g. the *E. coli* data set introduced in section 2.2.2). Nonetheless in these cases, it is of particular interest to test the distribution of fitness effects for heavy tails, for if improvements in fitness were drawn from a heavy tailed distribution, this would clearly have strong implications for the adaptive dynamics, as fitness would improve in jumps of considerable magnitude, see e.g. [101].
The distribution of fitness effects [110] has been subject of intensive study [97, 111]. Based on EVT arguments, it has been theoretically predicted that the distribution (at least for Fisher's geometric model) should be an exponential [75, 97, 111, 112]. This, however, is contradicted by empirical studies [126], where it is found that 'Beneficial Fitness Effects Are Not Exponentially Distributed'. Thus to some extent the question of the shape of the distribution of fitness effects remains open.
Fig 3.14 shows an application of the test derived in this chapter to the last of *E. coli* data sets introduced in section 2.2.2, the one that consists of resistance measurements for a number of organisms that are exactly one mutation away from the WT. Even though in this data set,

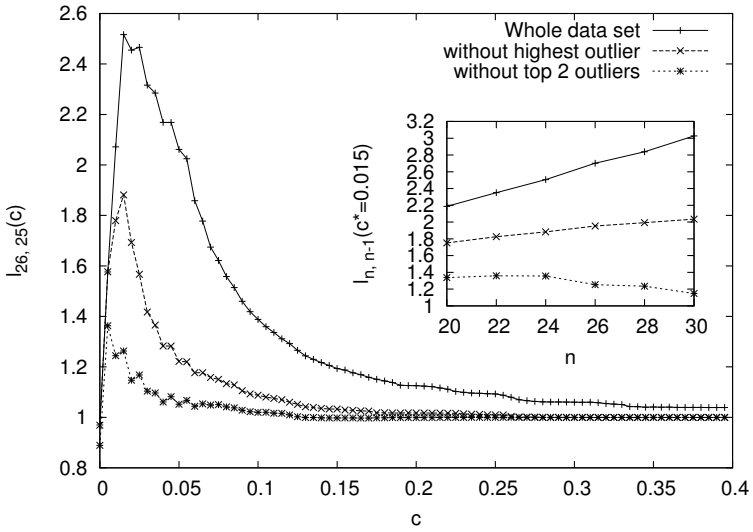


Figure 3.14.: The data set of 45 data points shows large attractive correlations indicative of heavy tails. These correlations grow (measured at $c^* = 0.1$) with the length of the sub-sequence n, showing that the positive correlations are not due to an exponential distribution. Removal of the largest entry of the data set, which is about three times larger than the next, results in a diminished signal which, however still clearly indicates heavy tails in the distribution.

not fitness but antibiotic resistance is measured, there is clearly a strong relation between the two. Furthermore the arguments predicting an exponential distribution for the fitness effects of mutations remain valid. At any rate, in the light of the growing number of multiply resistant bacteria in clinical isolates, finding heavy tails in the resistance distribution of mutations is of an interest of its own.
These heavy tails are found by the test, see fig 3.14. Even when the largest of the data points are removed, the heavy tails are still observed.

3.5. Conclusion

In this section, the statistics of records from random variables beyond the standard case of iid RVs was considered. While only the results presented in section 3.2.2 will later have direct implications for fitness landscapes, the field of record statistics has recently seen a surge of interest across many different models [86, 89, 96, 129, 130, 149, 151]. The counter-intuitive behavior observed numerically, i.e. that in some cases, record events attract, led to study the correlations between record events. These correlations were shown to have quite a non-trivial behavior that could roughly be described in terms of the universality classes of EVT in section 3.3.
These results on the correlations were used to set up a test for heavy tails presented in section 3.4, which works particularly well for small data sets on the order of a few dozen entries and is potentially more robust with respect to removal of outliers than conventional methods

used for detection of heavy tails[9]. Thus it was well suited for application to the data set from Cefotaxime resistance of the bacterium *E. coli* discussed above, where the data set was small since generating each point required generating one novel mutant and measuring its antibiotic resistance. Robustness with respect to removal of outliers was required to support the conclusion that the resistance effects were drawn from a heavy tailed distribution, since this is a very strong claim. So the side track of record statistics could be merged into the general subject of this thesis, in an unexpected, nonetheless interesting way.

The sequence $\{x_n\}$ was at times referred to as a time series. If this term is taken seriously, i.e. if the length of the time series n is indeed interpreted as a time point, the record rate p_n can be interpreted as 'the probability of a record occurring at time n'. Then, since the record process is Markovian, it is straightforward to extend it to a setting where the entries x_n and x_{n+1} of the time series are separated by waiting times τ_n, with $\{\tau_n\}$ a family of iid RVs and waiting times density $\psi(\tau)$ by standard arguments of renewal theory [102]. In particular when $\psi(\tau)$ is such that the mean waiting time $\langle\tau\rangle$ diverges, this leads to modifications in the behavior of the quantities considered in this chapter. These results are relegated to appendix B, since even though the resulting expressions seem to be new, they are obtained by a straightforward application of standard methods.

The results most important for the remainder of this thesis are those concerning the ordering probability $P_n(c)$ presented in section 3.2.2, as will be seen in the next chapter, which will return to the main theme of this thesis, i.e. the study of fitness landscapes.

[9]Note however that it only allows to *confirm* the hypothesis that heavy tails are present, not to *reject* this hypothesis, see discussion in section 3.4.

4. Accessible paths

In this chapter, the question of accessible paths traversing the whole fitness landscape is discussed for theoretical models of fitness landscapes as well as the empirical data.

4.1. Introduction

On any given fitness landscape, there is a fixed number n of paths of length L that start in the antipodal sequence, traverse the whole state space, end up in the global maximum and never encounter a step of declining fitness. This number varies from realization to realization.
As was pointed out in the introduction, such paths, called accessible [146], can play a decisive role in the adaptive fate of a population in the strong selection/weak mutation regime of adaptive dynamics, see section 1.4.2. Furthermore, they can be used as a measure of the global effects of epistasis, which is a feature only defined locally, see section 1.4.1. This second aspect of studying accessible paths is useful because accessible paths are a topographic feature independent of any adaptive dynamics taking place and can thus be used to classify an empirical fitness landscape. As will be shown in this section, accessible paths give some hints to the importance of epistasis in the FL studied, independent of the regime of evolutionary adaptation.
If on many realizations of FLs the number of accessible paths is counted, this histogram, suitably normalized, can be interpreted as the probability $\pi_L(n)$ of finding a given number n_L of accessible paths in a given FL of dimension L, provided, of course, that the FL considered is drawn from the same model as those that entered the histogram. Fig 4.1 shows the result of numerical simulations for the House of Cards model as introduced in section 2.1.1. The shape of $\pi_L(n)$ is typically found also for FLs derived from other models. This shape shows that in order to describe the distribution, the mean number of accessible paths

$$\langle n_L \rangle = \sum_{n=0}^{L!} n\pi_L(n) \tag{4.1}$$

does not suffice. In the examples shown in fig 4.1, one sees that many realizations have no accessible paths at all: In the main plot of fig 4.1, $\pi_L(n)$ decays roughly like an exponential for $n \geq 1$ but $\pi_L(0)$ is actually higher than what would be obtained by simply extrapolating the exponential to zero. Thus the probability $\pi_L(0)$ of finding no accessible paths at all needs to be considered as well. The inset shows $\pi_L(0)$ as function of L for two different conditions on the FL: The lower curve only considers those realizations, where the antipodal sequences was also the global minimum, the upper curve considers all realizations.

With the numerical methods used to count n_L on each instance of a FL (see appendix D for details on the numerical procedures used), each path is considered. The length of each path is just the sequence length and thus grows linearly in L, but the number of paths is given by

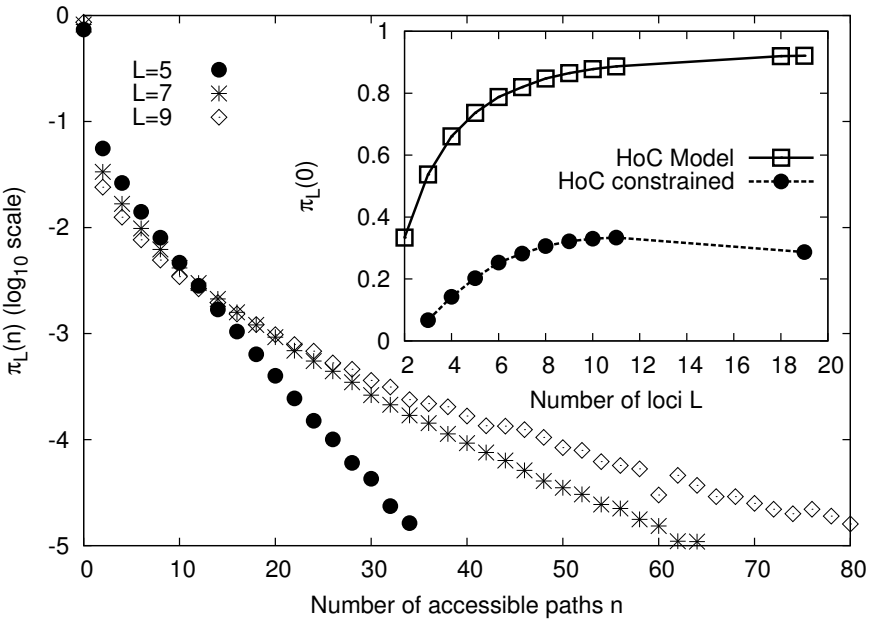


Figure 4.1.: Numerical simulations of the full distribution $\pi_L(n)$ of accessible paths (main plot) and the probability $\pi_L(0)$ of finding no accessible path (inset) for the HoC model. The inset shows the fraction of realizations without any accessible paths as function of L. The curves were averaged over 10^5 realizations.

$L!$ and thus grows faster than exponential. This means that sequence lengths on the order of magnitude relevant for biological systems cannot be explored numerically. Thus for numerical simulations, one faces a problem similar to that with empirical data: Only small sequence lengths are accessible to explicit explorations and the large-L behavior must be inferred. The difficulty in extrapolating the large L-behavior from curves for finite L is illustrated by an error made by Carneiro and Hartl [19]. In their publication, which introduced the use of $\pi_L(0)$ as additional measure to characterize $\pi_L(n)$, they traced the curves for $\pi_L(0)$ in the inset of fig 4.1 up to $L = 13$ and conjectured that without condition on the antipodal sequence (upper curve), $\lim_{L\to\infty} \pi_L(0) = 1$ and with condition (lower curve) $\lim_{L\to\infty} \pi_L(0) \sim 1/3$. The latter conjecture can be falsified by continuing the simulations to larger L, see the inset of fig 4.1 where in the lower curve, $\pi_L(0)$ shows non-monotonic behavior, reaching a maximum and declining eventually, see also [51]. The conjecture on the limit of $\pi_L(0)$ without condition on the AS, on the other hand, seems to be correct and is backed by several points that will be made in this chapter.

4.2. General arguments

Before entering into a detailed study of the models and empirical data, it is useful to consider the influence of the topology of the state space on $\langle n_L \rangle$ and $\pi_L(0)$.

4.2.1. Expected number of accessible paths

The expected number of accessible paths, $\langle n_L \rangle$, is perhaps the easiest property to study on high-dimensional fitness landscapes because it is essentially a one-dimensional object. This is due to the linearity of the expectation: Since there are $L!$ direct paths connecting any two points at Hamming distance L (like the AS and the GM), these paths can somehow be labeled 1 through $L!$. Then n_L can be written as

$$n_L = \sum_{l=1}^{L!} \mathbb{1}_{\text{path } l \text{ accessible}}, \tag{4.2}$$

where again the indicator function is 1 whenever the condition is met and zero otherwise. Taking the expectation on both sides of this equation, noting that by linearity of the expectation, it interchanges with the sum and using the standard identity that for any event X, $\langle \mathbb{1}_X \rangle = \text{Prob}[X]$, one obtains

$$n_L = L!\text{Prob}[\text{One arbitrarily chosen path is accessible}] \equiv L!P_L. \tag{4.3}$$

Note however that this is no longer true for higher moments of n_L, which is why these have not been studied so far. The probability P_L that any one given path is open[1] in eq (4.3) is the only object that will have to be computed separately for each model. It is equal to the probability that the L fitness values encountered along this path are ordered. Thus in order to compute the expected number of accessible paths, only a one-dimensional chain of L fitness values must be considered, in stark contrast to the higher-dimensional problem of even computing the second moment of n_L.

4.2.2. Probability of no accessible paths

The probability that no paths at all are accessible is harder to obtain, since all paths have to be considered. Furthermore, paths are not independent. Rather, the larger the dimension L of the hypercube is, the larger is also the number of combinations in which the paths can interact with each other, see fig 4.2. One way to avoid this problem is turning to another state space topology that yields results similar to those on the hypercube [15] but is easier to treat, see e.g. [13, 139]: The Cayley tree or Bethe lattice. From a starting node labeled 0 extend $\lambda + 1$ edges, each leading to another node. These $\lambda + 1$ 'nodes', corresponding to the states of the binary sequence, form a shell of distance 1 'around'[2] the starting node, and each of them is again the origin of λ edges, leading to the $\lambda(\lambda+1)$ states in the shell of distance 2 and so forth up to a given finite depth δ, see fig 4.3. Then, if each node on this axillary state space is given a fitness value, the same questions asked on the Boolean hypercube can also be treated on the Cayley tree. Note that in the last shell at depth δ, there are $(\lambda + 1)\lambda^{\delta-1}$ states and thus the same number of distinct paths, each leading to one of these states.

[1] This can be understood by the symmetries of state space: By averaging over all realizations of the FL with the appropriate weights, all paths are equivalent again, thus for the expected value, it suffices to consider an arbitrary path.

[2] One way to graphically represent this state space is to draw these states concentrically around the starting node.

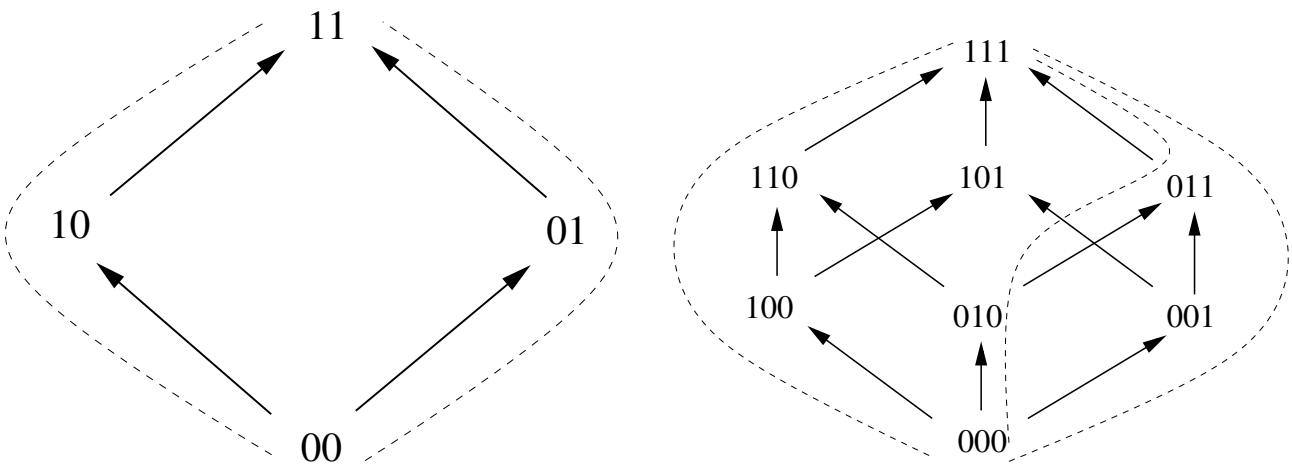


Figure 4.2.: The left panel shows the two paths on an $L = 2$ landscape. Clearly the two paths are independent and $p_L(0)$ can be computed for many models, while in the right panel, some of the paths are not independent anymore, see the dashed lines in each panel and in particular the two rightmost dashed lines in the right panel, which share the edge $011 \to 111$. As L increases, these interdependencies of paths become more and more important.

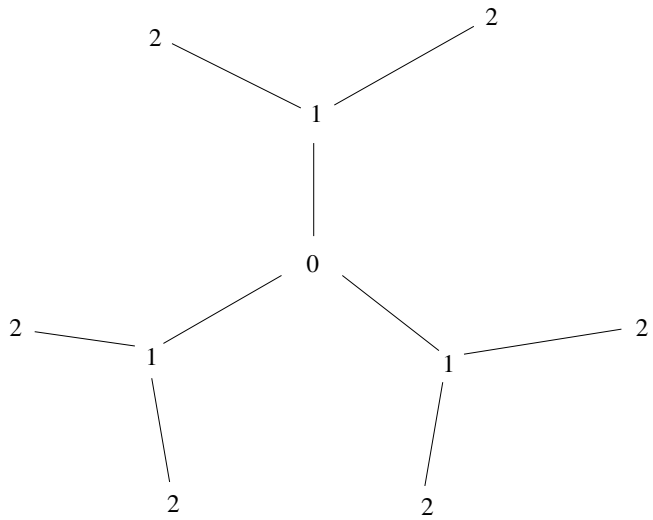


Figure 4.3.: Cayley tree of coordination number $\lambda = 2$ and maximum depth $\delta = 2$. Note that since the state space is a tree, no two paths share any edges once they have separated. Since the three (or in general $\lambda + 1$) subbranches originating at site 0 are independent, the states in a given shell k are not labeled k_1 through k_λ but just by the shell to which they belong.

In this state space, paths can still share edges but they will never merge once they have separated, as opposed to the original state space, see the right panel of fig 4.2. Paths then originate at the starting node and move along edges from one shell to another. The equivalent of an accessible path reaching the global optimum is then a path reaching the final depth δ with the choice $\delta = L$. The other parameter of the Cayley tree, the coordination number λ, still has to be fixed. There are two reasonable choices possible. One is to set $\lambda = L - 1$ such that there are L possible paths originating in the starting node 0. This however has the drawback that at final depth $\delta = L$, for $L \gg 1$ there will be approximately $L^L \gg L!$ paths leading to the final shell, much more than on the hypercube. Furthermore along each path crossing the hypercube, the number of edges leading towards the GM decreases with each step. Thus it seems more appropriate to set λ such that the total number of possible paths is $L!$,

$$(\lambda + 1)\lambda^{L-1} \approx \lambda^L \equiv L! \text{ or } \lambda = \lfloor (L!)^{1/(L)} \rfloor \tag{4.4}$$

since λ grows with L such that for L large enough, the term in λ^{L-1} can be ignored. Because λ is necessarily an integer number, only the integer part of the asymptotic solution is taken, as indicated by the $\lfloor \cdot \rfloor$ parentheses. Using Stirling's approximation for large L, $L! \approx \sqrt{2\pi L} L^L e^{-L}$, the coordination number λ from eq (4.4) becomes asymptotically

$$\lambda \sim \left\lfloor \frac{L}{e} \right\rfloor . \tag{4.5}$$

Thus the coordination number λ behaves approximately linear in L with a slope of $e^{-1} = 0.367879\ldots$, see also fig 4.4.
The reasoning leading to eq (4.3) for the expected number of accessible paths remains valid. Since in both cases, L steps have to be taken until the goal is reached, the probability for the corresponding fitness values to occur in order of increasing value, P_L, remains unchanged[3]. Thus by setting λ such that the *total* number of paths are equal in both cases, the *expected* number of accessible paths matches automatically as well.

Now it is possible to set up a recursion equation for $\pi_L(0)$ on this configuration space. One can adapt the arguments in [139], there in the context of percolation theory, and generalize them suitably. The probability of not getting to shell L factorizes into a product over the probabilities $\tilde{\pi}_L(0 \text{ via } i)$ of not reaching shell L via edge i

$$\pi_L(0) = \prod_{i=1}^{\lambda+1} \tilde{\pi}_L(0 \text{ via } i) = \tilde{\pi}_L(0 \text{ via } 1)^{\lambda+1} \tag{4.6}$$

by symmetry of the graph. If the first step is possible, the question remains whether shell L can be reached or if all paths are blocked further down in the tree. This gives rise to a fix point equation the precise form of which depends on the model to be considered, see later sections of this chapter.

With these general tools, one can analyze the behavior of $\langle n_L \rangle$ and $\pi_L(0)$ for fitness landscapes drawn from the different models.

[3]Provided, of course, that fitness values on both configuration spaces are drawn from the same model.

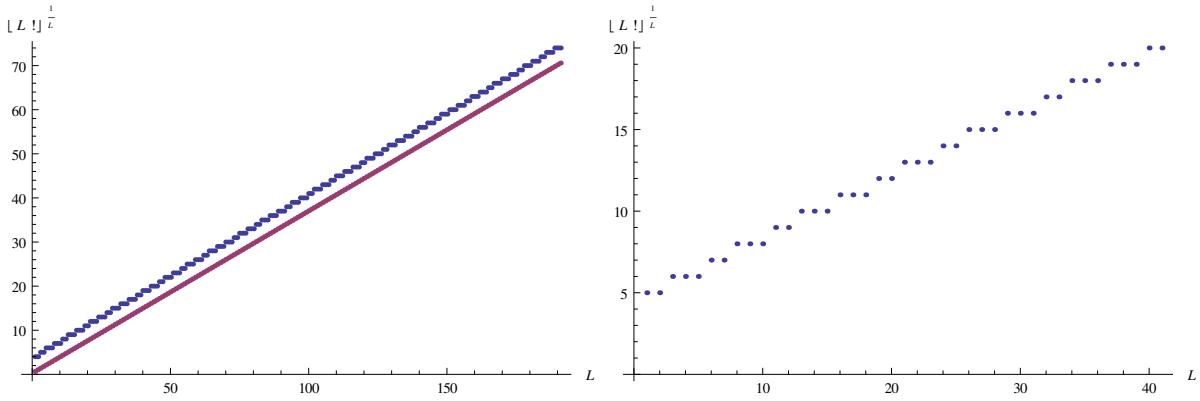


Figure 4.4.: The left panel shows the coordination number λ as given by eq (4.4) as function of L (blue). Note that λ appears to be linear in L with a slope of $0.367879\ldots$ as indicated by the parallel line (red). A closer inspection of λ shown in the right panel reveals a slightly more complex structure, which is due to the fact that only the integer part of $L^{1/L}$ is considered. The plots were obtained using Mathematica.

4.3. HoC and RMF models

In the House of Cards and Rough Mt Fuji models, the fitness values of each of the 2^L states are a combination of a contribution deterministically increasing towards the GM (or the initial node 0 on the Cayley tree) with slope $c \geq 0$ ($c = 0$ in the HoC model) and iid random contributions with common probability density $f(\cdot)$. To compute the expected number of accessible paths, the results on the ordering probability of sequences of RVs given in section 3.2.2 can be used to obtain a fairly general picture of the behavior of $\langle n_L \rangle (c)$. $\pi_L(0)$ on the other hand could not be studied analytically for the original configuration space $\{0,1\}^L$. However, extensive numerical studies together with heuristic arguments based on the analytically obtained scaling behavior of $\langle n_L \rangle (c)$ and a recursion relation for $\pi_L(0)$ on the auxiliary configuration space defined above (the Cayley tree of depth L and with coordination number $\lambda = \lfloor (L!)^{1/(L-1)} \rfloor$) will allow to make a conjecture about the general behavior of this probability for large L for arbitrary c and any underlying probability density $f(\cdot)$.

4.3.1. Expected number of accessible paths

To compute $\langle n_L \rangle$ for these two models, one first observes that any direct path between the AS and the GM has exactly one fitness value from each shell of given Hamming distance l from the GM. Then the probability that one given path is accessible is simply the probability that L RVs drawn according to

$$y_l = lc + x_l, c \geq 0 \text{ with } \{x_l\} \text{iid RVs} \tag{4.7}$$

occur in order of increasing magnitude. Thus P_L is in this case given by

$$P_L(c) = \text{Prob}[y_1 < y_2 < \cdots < y_L] \tag{4.8}$$

as studied in chapter 3, where the length of the sequence of RVs was denoted by n rather than L.

For $c \equiv 0$, $P_L(c)$ was computed exactly in eq (3.18) as $1/L!$. Thus in the HoC model, the expected number of paths is

$$\langle n_L \rangle = \frac{L!}{L!} = 1, \tag{4.9}$$

see [84] and [51]. Thus for HoC landscapes, on average one of the $L!$ paths will be accessible for any sequence length L. Comparing this finding to the full distribution of accessible paths obtained numerically for the HoC model, see fig 4.1, it becomes clear that a landscape with the expected number of accessible paths is not a *typical* landscape as such a landscape typically has no accessible path at all. This was the initial reason for considering the probability of having no paths, $\pi_L(0)$, as additional object.
If the distribution $f(\cdot)$ underlying the iid RVs $\{x_n\}$ is the Gumbel distribution, $P_L(c)$ could be obtained exactly for arbitrary c, see eq (3.21)[4]. For large L and finite c, an exact asymptotic expression for $P_L(c)$ was obtained in eq (3.24), where it was found that $P_L(c) \propto \exp(-L \ln(1 - e^{-c}))$. Since the range of c for which $P_L(c)$ shows non-trivial behavior scales like $\ln(L)$ and is thus large for large L, one can use the approximation $\ln(1 - e^{-c}) \approx -e^{-c}$ and has

$$P_L(c) \propto \exp(-Le^{-c}). \tag{4.10}$$

Thus $P_L(c)$ essentially decays exponentially in L. Inserting this into eq (4.3), one sees that for large L, the *combinatorial* proliferation of the total number of paths will out-compete the *exponential* decline in probability that one given path is accessible. Thus for the exactly solvable case of the Gumbel distribution in the RMF model, the expected number of accessible paths will grow without bounds.
While a general expression for arbitrary c could not be obtained, a perturbative approach for $0 < c \ll 1$ gave the expression

$$P_L(c) \approx \frac{1}{L!} + \frac{1}{(L-2)!} I_f \tag{4.11}$$

with $I_f = \int \mathrm{d}x f^2(x)$ a positive constant depending on the underlying probability density $f(\cdot)$, see eq (3.32). Plugging this asymptotic form into eq (4.3), one sees that the expected number of accessible paths behaves for small c like

$$\langle n_L \rangle (c) \approx 1 + L(L-1)cI_f. \tag{4.12}$$

For $c \equiv 0$, one finds the result for the HoC model derived above, while for positive c, $\langle n_L \rangle$ grows with L without bounds. Thus in the large L-limit, any small positive drift will have a very pronounced effect.
For large L and large c, $P_L(c)$ behaves like

$$P_L(c) \approx \exp[-L\epsilon(c)] \text{ with } \epsilon(c) > 0 \tag{4.13}$$

[4]Like in chapter 3 the important parameter is actually $\theta \equiv c/\sigma$ where σ is the standard deviation of the parental distribution $f(\cdot)$ (if it exists) or some other means of estimating the scale on which RVs drawn from $f(\cdot)$ vary (if it does not). However, as all simulations performed use RVs with a standard Gaussian distribution $\sigma = 1$, thus this distinction is not necessary since $\theta = c$.

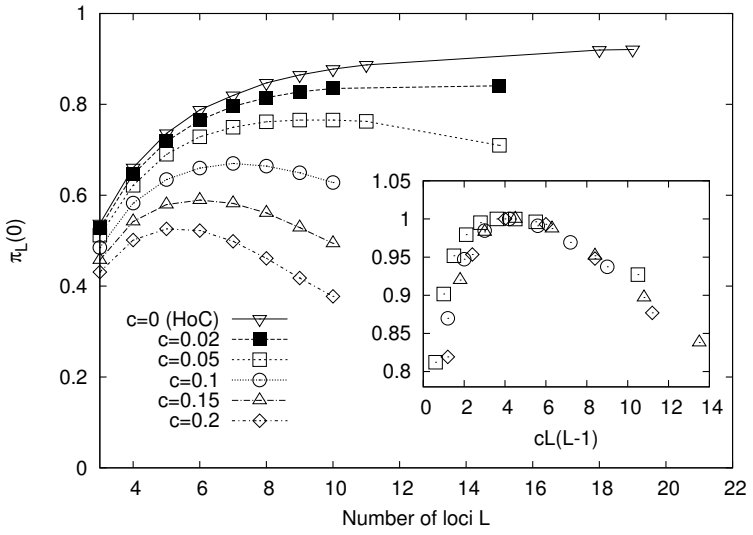


Figure 4.5.: The figure shows that behavior of $\pi_L(0)$ from the RMF model for different values of c. For $c \geq 0.05$, the curves in the main plot show increase up to a certain sequence length followed by a decline. The inset suggests that this non-monotonic behavior is universal for positive c: All curves for $c \geq 0.05$ are plotted as function of $cL(L-1)$ and normalized in height. They peak at a value of $cL(L-1) \sim 4$, which indicates that as c decreases, the sequence length for which the curves heave a maximum increases but remains finite for any positive c. For the iid part of the fitness values, a Normal distribution with mean $\sigma = 1$ was used. The curves were averaged over 10^5 simulations.

as shown in section 3.2. Again, substituting this asymptotic expression into eq (4.3), one sees that for large enough L, the combinatorial proliferation of the total number of paths outweighs the exponential decline of the probability that one given path is accessible, meaning that in the limit of large c, $\langle n_L \rangle (c)$ grows with L.
While it is still possible that the expected number of accessible paths decreases with L for a value of c between the ranges of validity of the asymptotic expressions given by eq (3.32) or (3.34) for small and large c respectively, this would come as a great surprise and contradict numerical simulations, specially in the light of the results for the Gumbel case which are valid for all c, see [52]. Thus one can conclude that the behavior of the HoC model for large L differs fundamentally from that of the RMF model even for arbitrarily small drift c, at least with respect to $\langle n_L \rangle$.

4.3.2. Probability of no accessible paths

For the original configuration space $\mathcal{C}_L = \{0,1\}^L$, due to the complicated inter-dependencies of paths mentioned in section 4.2 only numerical simulations of these two models were possible. Fig 4.5 shows numerical simulations of $\pi_L(0)$ for various values of c.

The inset of fig 4.5 shows the curves, divided by their respective maximum value, plotted as function of $cL(L-1)$. In this scaling variable, the same as for $\langle n_L \rangle (c)$, see eq (4.12), the curves attain their maximum at the same point, $cL(L-1) \sim 4$. This scaling could not be proven to hold for $\pi_L(0)$ in general, but as seen in the inset of fig 4.5, it clearly holds at least for those

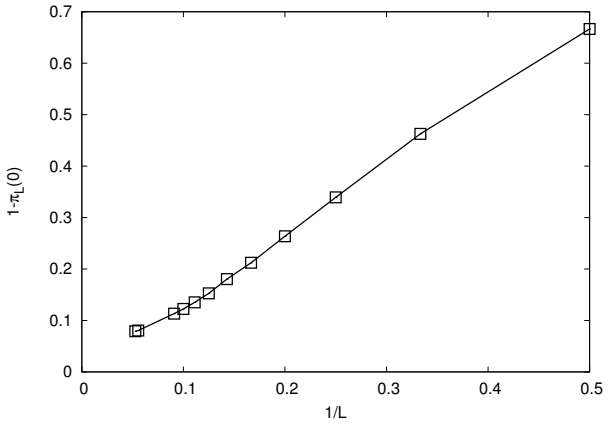


Figure 4.6.: This plot shows numerical simulations of $1-\pi_L(0)$ as function of $1/L$ for the HoC model ($c = 0$). While it is not possible to infer from this curve that $\pi_L(0)$ tends to unity for this model, it clearly has a limiting value above 0.9. The iid part of the fitness values were drawn from a standard Normal distribution, the curves were averaged over 10^5 realizations.

examples presented there. Assuming that it holds generally, one can make predictions about the limiting value of $\pi_L(0)$ for large L for arbitrary c based on simulations for values of c for which a limiting regime appears to be attained already for moderate values of L. If $cL(L-1)$ is the indeed the relevant scaling parameter, then the limit of large L at fixed c is equivalent to that of large c at fixed L. In the latter limit, clearly $\pi_L(0) \to 0$. For the HoC model, extrapolation to large L is more difficult. However, assuming that $\pi_L(0)$ obeys for this model the same scaling as for the RMF model, one can obtain a lower bound for the limiting value of $\pi_L(0)$. If the same scaling as for the RMF model applies, $\pi_L(0)$ will continue to increase for all finite L, since the condition $cL(L-1) \sim 4$ with $c = 0$ is not met by any finite L. Thus for the HoC model, $\pi_L(0)$ is monotonic for all finite L. Fig 4.6 shows

$$1 - \pi_L(0) \equiv \text{Prob}[n_L \geq 1] \tag{4.14}$$

as function of $1/L$. While it is not clear that $\lim_{L^{-1}\to 0}(1-\pi_L(0)) = 0$, one already sees that $\lim_{L\to\infty} \pi_L(0) \geq 0.90$.

On the auxiliary configuration space defined in section 4.2 above, the Cayley tree of depth L and coordination number $\lambda = \lfloor L!^{1/(L-1)} \rfloor$, the question of $\pi_L(0)$ seems more amenable to analytic computation. The method presented in [139] for the percolation problem, which led to eq (4.6), can here be used to set up a recursion equation for $\pi_L(0)$. If one denotes the conditional probability that there are no accessible paths to depth L *given* that the starting node 0 has fitness x_0 by $\pi_L(0|x_0)$ and the probability density of x_0 by $f_0(\cdot)$, one can rewrite $\pi_L(0)$ as

$$\pi_L(0) = \int \mathrm{d}x_0 f(x_0) \pi_L(0|x_0). \tag{4.15}$$

With a condition on the fitness x_0 of initial site, the branches of the Cayley tree become stochastically independent and $\pi_L(0|x_0)$ factorizes into the probabilities that depth L cannot

be reached through any of these branches. Denoting the probability that depth L cannot be reached via branch i given the fitness x_0 of the starting site by $\tilde{\pi}_{L-1}(0 \text{ via } i|x_0)$ and using the fact that this probability is the same for all branches i by symmetry, eq (4.15) becomes

$$\pi_L(0) = \int \mathrm{d}x_0 f_0(x_0) \tilde{\pi}_{L-1}(0 \text{ via } i|x_0)^{\lambda+1}. \tag{4.16}$$

With this condition, the event that no accessible paths lead from the starting site 0 to depth L via branch i can be given by two disjoint events: *Either* state 1, the first state in branch i, has lower fitness x_1 than state 0 *or* state 1 has higher fitness but none of the paths starting there lead to depth L. The latter event then has probability $\pi_{L-1}(0|x_1)$ since with the condition on x_1, the previous history of the process, i.e. the number of steps taken so far, has no influence on the probability of reaching depth L. $\pi_{L-1}(0|x_1)$ again factorizes with respect to the λ different branches originating in state 1 and eq (4.6) finally takes the form

$$\pi_L(0) = \int \mathrm{d}x_0 f_0(x_0) \left\{ \int^{x_0} \mathrm{d}x_1 f_1(x_1) + \int_{x_0} \mathrm{d}x_1 f_1(x_1) \tilde{\pi}_{L-1}(0 \text{ via } i|x_1)^{\lambda} \right\}^{\lambda+1}. \tag{4.17}$$

Expressing the left hand side of eq (4.17) by the right hand side of eq (4.16) yields

$$\int \mathrm{d}x_0 f_0(x_0) \tilde{\pi}_L(0 \text{ via } i|x_0)^{\lambda+1} = \int \mathrm{d}x_0 f_0(x_0) \left\{ \int^{x_0} \mathrm{d}x_1 f_1(x_1) + \int_{x_0} \mathrm{d}x_1 f_1(x_1) \tilde{\pi}_{L-1}(0 \text{ via } i|x_1)^{\lambda} \right\}^{\lambda+1}. \tag{4.18}$$

This can be interpreted as an implicit[5] fixed-point equation. Clearly one solution of this is $\tilde{\pi}_l(0 \text{ via } i|y) = 1$ for all l, y, which corresponds to $\pi_L(0) = 1$. For $\lambda \to \infty$, another solution of eq (4.18) is $\tilde{\pi}_l(0 \text{ via } i|y) = 0$ for all l, y. This solution corresponds to $\pi_L(0) = 0$. For large L, these two seem to be the only possible solutions, since if L is so large that for any $\tilde{\pi}_L(0 \text{ via } i|x_0) < 1$, $\tilde{\pi}_L(0 \text{ via } i|x_0)^L \approx 0$, the term in integrand on the right hand side of eq (4.17). It was not possible to compute a critical value of drift c dividing these two solutions, but comparison to the behavior of $\langle n_L \rangle (c)$ suggests that this critical value might be $c = 0$. Numerical simulations performed on this state space, show a behavior similar to that for the original problem on the Boolean hypercube, see fig 4.7.

4.4. Neutral model

In the neutral model as represented by site percolation (see section 2.1.4), both properties $\langle n_L \rangle (p)$ and $\pi_L(0)$ are well studied, see e.g. [139] and references therein and [57]. Thus mostly known results will be restated here, for the largest part closely following [139].

[5] implicit because the object of interest, $\tilde{\pi}_L(0 \text{ via } i|x_0)$ only appears under the integrals

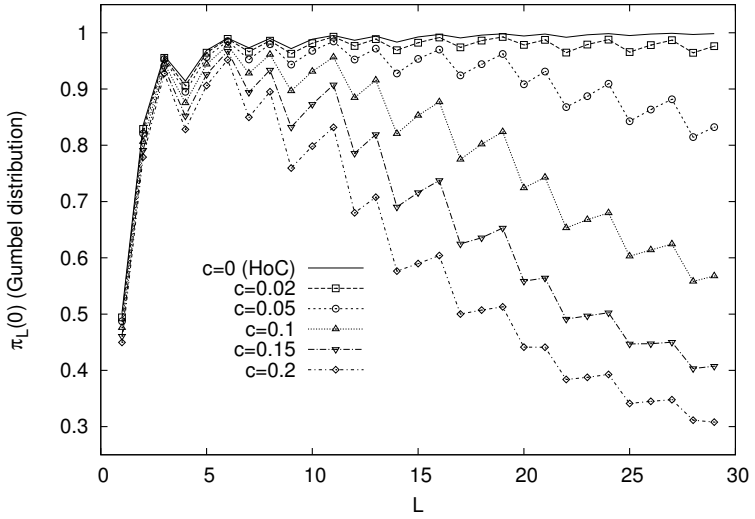


Figure 4.7.: Simulations of $\pi_L(0)$ as function of L on the Cayley tree with coordination number $\lambda = \lfloor L!^{1/(L-1)} \rfloor$. Because λ has jumps, the curves have a characteristic structure (cf. fig 4.4) but follow the overall behavior observed for the RMF model on the original state space. Simulations were obtained from a Gumbel distribution with drift c measured in units of that distribution's standard deviation

4.4.1. Expected number of accessible paths

The reasoning leading to eq (4.3) remains valid for the neutral model, where P_L is simple to compute. Since the states are independently of each other either viable with probability p or not viable with probability $1-p$, the probability that one given path is accessible is simply the probability that L consecutive states are viable and thus

$$\langle n_L \rangle = L! p^L. \tag{4.19}$$

Note that even though for any $p < 1$, $P_L(p)$ decays exponentially in L, the combinatorial proliferation of possible paths always wins and thus $\langle n_L \rangle$ grows without bounds in L for fixed p. A closer examination reveals that if $p = p_c(L) \equiv L!^{-1/L}$, $\langle n_L \rangle = 1$, thus if p remains constant or decays more slowly than $p_c(L)$, the expected number of accessible paths diverges. On the Boolean Hypercube $\{0,1\}^L$, the critical value for p is therefore L-dependent and given by this $p_c(L)$.

On the Cayley tree, this model has been discussed in detail in the literature. The mean number of accessible paths was originally computed in [49]. Each node of the tree is either occupied (with probability p) or empty (with probability $1-p$). This means that in each new shell (with the exception of the first), there will be on average $p\lambda$ viable states leading forward to the next shell, from each of which there will be again an average number of $p\lambda$ paths emanating towards higher shells. Then the mean number of paths reaching depth L for given p and λ is simply $\langle n_L \rangle = p(\lambda+1)(p\lambda)^{L-1} \approx (p\lambda)^L$. In the limit of $L \to \infty$, three different cases are possible:

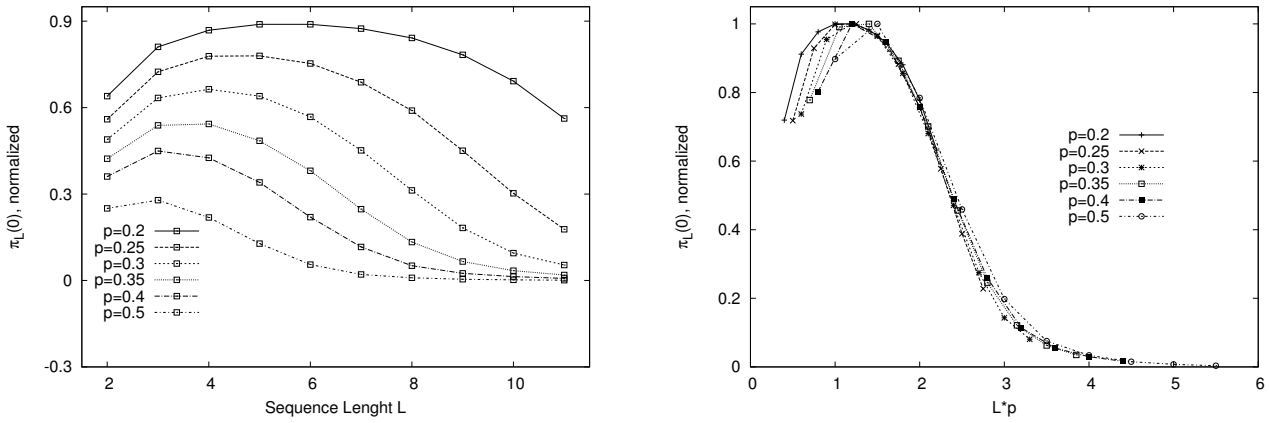


Figure 4.8.: Simulations of $\pi_L(0)$ for different values of p on the Boolean hypercube of dimension L. For all values of p simulated here, $\pi_L(0)$ eventually decays to zero as can be seen in the left panel. The right panel shows a collapse of the data as function of the product of sequence length L and probability p.

$$\langle n_L \rangle = \begin{cases} 0 & \text{for } p < \frac{1}{\lambda} \\ 1 & \text{for } p = \frac{1}{\lambda} \\ \infty & \text{for } p > \frac{1}{\lambda} \end{cases} \tag{4.20}$$

which allows to infer that $p_c = \lambda^{-1}$ is the critical value of p for a given coordination number. Since $\lambda = \lfloor L!^{1/L} \rfloor$ as discussed above, this critical $p_c(\lambda)$ has essentially the same L-dependence as $p_c(L)$ found for the original state space.

4.4.2. Probability of no accessible paths

Like for the RMF model, $\pi_L(0)$ could on the original state space only be studied numerically. The picture is qualitatively similar to that obtained in the RMF case with the exception that none of the curves corresponds to that for the HoC model: for all values of p tested, the curves eventually decline, see fig 4.8.

On the Cayley tree, computing $\pi_L(0)$ is a classical problem first treated by Flory [49]. Here, the presentation given in [139] will be extended to the problem at hand. The event that the starting site 0 is not connected to the starting site is given by the two mutually exclusive subevents that *either* site 0 is not occupied (with probability $1-p$) *or* site 0 is occupied but not connected to depth L through any of the $\lambda + 1$ branches originating there. The probability that a given branch is not connected to infinity is called Q. Thus $\pi_L(0) = 1 - p + pQ^{\lambda+1}$. If however, as was done in the simulations shown here, the starting site 0 is conditioned to be occupied, one has

$$\pi_L(0) = Q^{\lambda+1}. \tag{4.21}$$

In the limit $L \to \infty$, the graph becomes translationally invariant and the probability that a given site does not lead to depth $L = \infty$ is the same for all sites. Then, by essentially the same reasoning as above, Q obeys the fix point equation

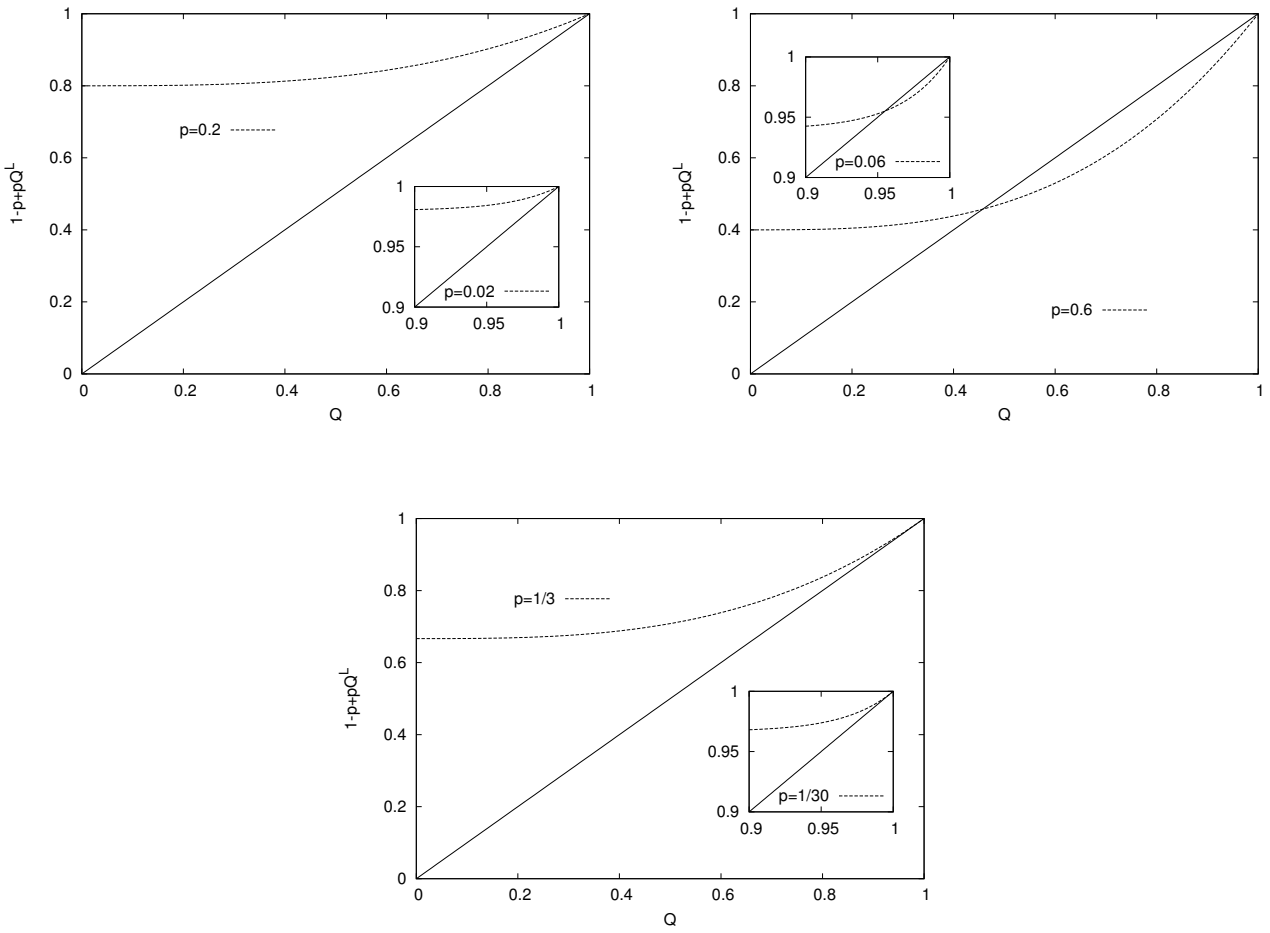


Figure 4.9.: Left hand side (solid line) and right hand side (dashed line) of eq (4.22) for $\lambda = 3$ (main plots) and $\lambda = 30$ (inset). On the top left ($p < 1/\lambda$), there is only one solution at $Q = 1$, while on the top right ($p > 1/\lambda$), there are two solutions. In the middle at the bottom ($p = 1/\lambda$), there is also just on solution, but the two curves have the same slope at $Q = 1$.

$$Q = 1 - p + pQ^{\lambda}. \tag{4.22}$$

This has obviously the solution $Q = 1$, which, substituted into eq (4.21) means $\pi_L(0) = 1$. This solution belongs to the regime below some critical value $p_c(\lambda)$. For sufficiently large values of p, however, another solution appears, as can be seen in fig 4.9 where the left hand side and right hand side of eq (4.22) are plotted independently of each other for the three different regimes of p ($p < p_c$, $p = p_c$, and $p > p_c$) and two values of λ. From fig 4.9, one sees that for one (λ-dependent) critical value of p_c of the parameter p, the two solutions must coincide. For that value of p, both curves must have the same slope at $Q = 1$. This condition yields

$$\frac{\mathrm{d}}{\mathrm{d}Q}Q\bigg|_{Q=1} = 1 = \frac{\mathrm{d}}{\mathrm{d}Q}(1 - p_c + p_c Q^{\lambda})\bigg|_{Q=1} \equiv \lambda p_c(\lambda) \tag{4.23}$$

and thus $p_c(\lambda) = 1/\lambda$ as for the mean number of accessible paths, see eq (4.20).

This regime of $p < p_c(\lambda) = 1/\lambda$, however, will vanish for large sequence lengths L, as $\lim_{L\to\infty} p_c(\lambda(L)) = 0$. Since $L \to \infty$ implies $\lambda \to \infty$ and $Q \leq 1$ as it is a probability, the other solution of eq (4.22) becomes $Q \to 1-p$ for $L \gg 1$. Substituting this back into eq (4.21), one sees that any solution $Q < 1$ will for $L \to \infty$ lead to vanishing $\pi_L(0)$ and thus for any $0 < p \leq 1$ the probability of no accessible path is given by $\pi_L(0) = 0$. One obtains again the same statements as for the RMF case, that for almost all parameter values, there will be accessible paths on this FL.

4.5. The LK model

In the HoC model, the computation of local properties like the probability that a given state is a maximum (Prob [state i is max] $= 1/(L+1)$ since any one of the $L+1$ states has equal probability of being the highest, see [76]) and thus the expected number of maxima ($2^L/(L+1)$ by the same reasoning that led to eq (4.3)) are more than easy to compute. In the RMF model, computations are harder but essentially straightforward extensions of the same ideas.
In the LK model, due to its complicated statement, even local properties can be fairly difficult to compute, see the works by Weinberger [144], Durrett and Limic [38] and others [44, 90] in contrast to the simple argument for the HoC model presented above. Choosing an alternative state space like the Cayley tree is also not as straightforward as in the other models since, following classical approaches to spin glasses on the Cayley tree or Bethe lattice [13], the coordination number would be set by the interaction parameter K and one would not have the freedom to match the coordination number λ such that the expected number of accessible paths coincides with that on the Boolean hypercube. Thus for both $\langle n_{L,K} \rangle$ and $\pi_{L,K}(0)$, only numerical simulations were possible.

4.5.1. Expected number of accessible paths

The simulations in fig 4.10 show the expected number of accessible paths as function of K for different values of L. For $K = 0$, all $L!$ paths are accessible, while for $K = L-1$ (the last point of each curve), there is on average one accessible paths consistent with the HoC model indicated by the dashed line. For large K, the lines in the semi-logarithmic scales used in fig 4.10 become parallel, which implies that for fixed K and large L, $\langle n_L \rangle$ decays exponentially in K. This is corroborated by the inset which shows a detail of the main plot for $L = 7, 8$ and $L = 9$ (symbols) compared to exponential decay (lines). On the other hand, for fixed $K \geq 4$, the curves for different L appear equi-distant, which would imply exponential growth with L. If one connects the dots corresponding to $K = L/2$ (as done by the line labeled 'exponential growth' which follows the curve $0.3e^x$), one sees that in this particular scaling of K with L, $\langle n_L \rangle$ also grows approximately exponentially. The fact that both the $K = 1$ and $K = 8$-points are above the exponential growth curve might suggest that $\langle n_L \rangle$ for $K = L/2$ is slightly curved, meaning super-exponential growth.
For $L-K$ fixed, e.g. by connecting the last (corresponding to $L-K = 1$) or second to last (for $L-K = 2$) points of the curves, one sees that $\langle n_L \rangle$ is in this regime approximately constant, just like in the HoC ($K = L-1$) case only at a higher value.

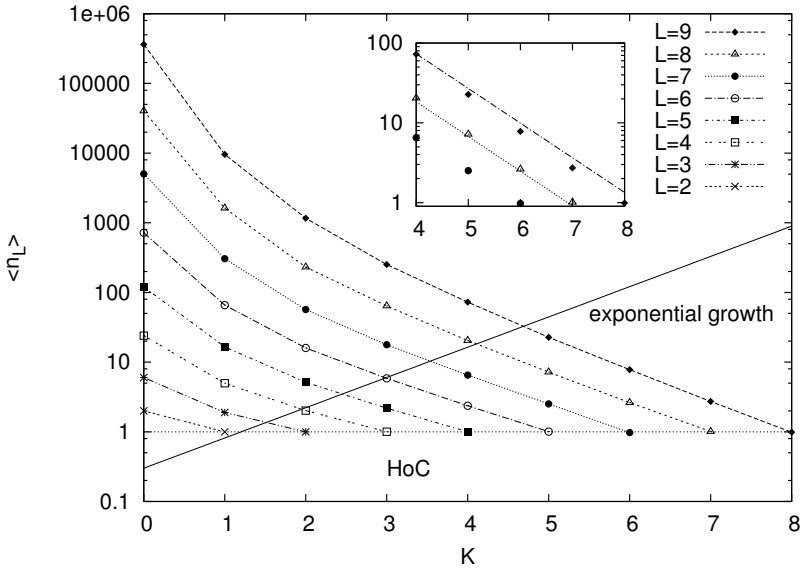


Figure 4.10.: The expected number of accessible paths as function of K for various values of L. The inset shows a detail of the main plot. One sees that the curves are approximately parallel and decay exponentially for large K (indicated by the lines in the inset). In the main plots, on recovers the HoC result (dashed line) for $K = L - 1$ and an exponentially growing curve can be fitted through the four points that correspond to $K = L/2$. In parts, this behavior has already been observed in [84].

4.5.2. Probability of no accessible path

The behavior of $\pi_{L,K}(0)$ depends strongly on the way in which the interaction parameter K depends on the sequence length L. This behavior was discussed in [51] in the context of the *A. niger* landscape. See also [113] for an investigation of the influence of K on the outcome of populations adapting in dynamical regimes different from those discussed here.
If the difference $L - K$ is fixed, $\pi_{L,k}(0)$ increases for all values of L considered, as can be seen in fig 4.11. This phenomenon can heuristically be explained by noting that as L grows, the fraction of non-interacting sites $L - K$ of the whole sequence length L decreases such that eventually the difference between $L - K > 1$ fixed and $L - K = 1$ will become negligible. Thus for large L all LK models with fixed number of non-interacting sites resemble the $K = L - 1$ case (which is equivalent to the HoC model, see section 2.1.2).

This reasoning can also be inverted. If K is set to a fixed number, then, as L grows, $\pi_{L,K}(0)$ should asymptotically show the behavior as the model for $K = 0$. The simulations in fig 4.12 show precisely this behavior with an interesting and unexpected deviation for $K \leq 2$ (a regime that is not believed to be of biological relevance, see [51]). When keeping the *ratio* of interacting sites K/L fixed, heuristic explanations cannot be applied. Numerical simulations show that $\pi_{L,K}(0)$ in this case behaves much like for the models considered in the previous sections. After an initial increase, it finally peaks and eventually a clear decline sets in.
For fixed value of K, the landscapes appear to become smooth for large values of L, as can be seen in the inset of fig 4.12, where $\pi_{L,K}(0)$ declines monotonically with L. This is however only true for values of $K \geq 3$. On the right of fig 4.12, where the inset on the left is shown with

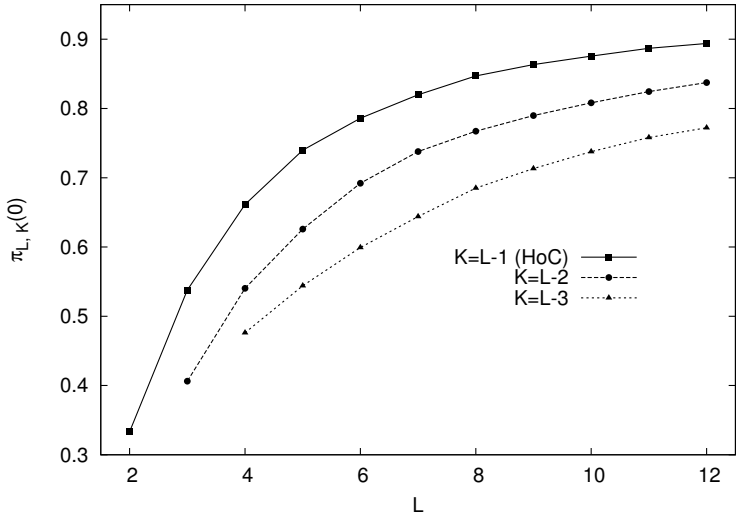


Figure 4.11.: Simulation results for $\pi_{L,K}(0)$ for different values of $L-K$. Simulations were averaged over 10^5 realizations with RVs drawn from a standard Normal distribution.

more values of K, it is seen that for $K = 2$, $\pi_{L,K}(0) \sim 0.55$ and does not appear to change much, while for $K = 1$, it actually *increases*. This numerically observed behavior is fairly surprising since for $K \geq 3$, i.e. in cases with more interactions and thus more epistasis than for $K \leq 2$, $\pi_{L,K}(0)$ is actually much lower. One possible interpretation of this is that for $K = 1$, the landscapes arranges itself such that the path to the GM is blocked. This non-monotonic behavior in K was a surprise specially in the light of the well-ordered K-behavior observed for $\langle n_L \rangle$ in fig 4.10.
The fact that this sudden and drastic change of behavior occurs at $K = 2$ has been tentatively conjectured to be a sign of a transition similar to the $\mathcal{P} - \mathcal{NP}$ transition in computational complexity in related spin glass problems [99] which is also present in the LK model [145].
Note however that in [145] it was found that the transition in computation complexity only occurs for the LK model with randomly chosen neighborhoods (the case considered here) while for adjacent neighborhoods, this is not the case, see also [6].

4.5.3. Full distribution of accessible paths

To better understand the behavior of $\pi_L(0)$ observed on the LK model for fixed value of K in the last section and in particular the counter-intuitive behavior for $K \leq 2$, here the full distribution $\pi_L(n)$ for $K = 1, 2$ and $K = 3$ is studied by numerical simulations. The simulations presented in figs 4.13 - 4.15 show that for $K = 1$, the distribution of accessible paths is dominated by peaks corresponding to substantial fractions of the number of paths being accessible (note the peak at 1 in fig 4.13 indicating that all paths are accessible). While the position of the peaks can be explained by observing that the peak closest to that corresponding to all paths being accessible corresponds to one 'broken link' in the middle of the graph with all other links present and so on, the reason for their magnitude remains unknown.
As K increases, the distribution becomes more and more similar to those observed for the RMF and HoC models (see e.g. fig 4.1). This behavior was originally observed in the Diploma thesis of Alexander Klözer [84].

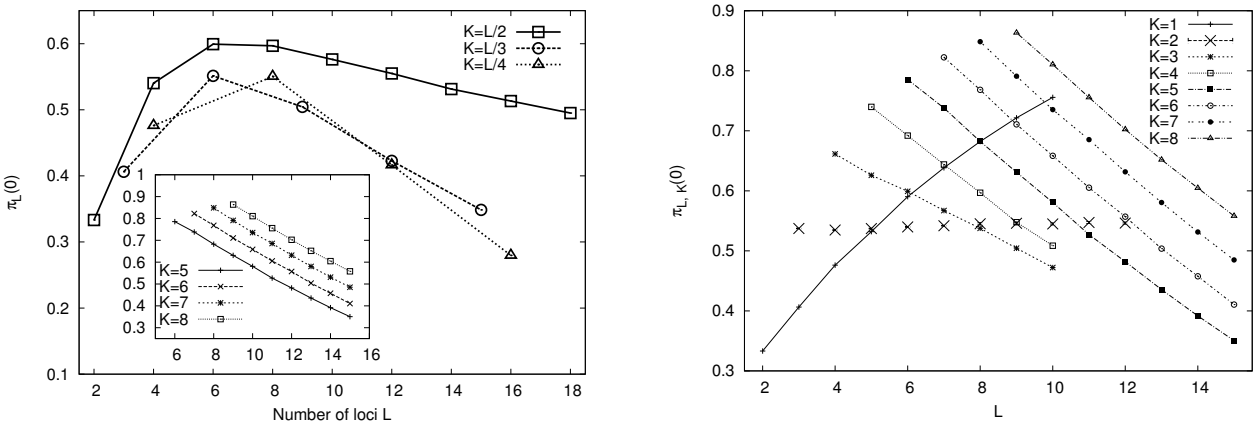


Figure 4.12.: The left panel shows the results of numerical simulations for various values of L/K (main plot) and for fixed K (inset). In both cases, $\pi_L(0)$ ultimately declines as L grows. For the case of L/K fixed, where none of the heuristic explanations given for the behavior of the other two cases applies, one observes a behavior very similar to that of $\pi_L(0)$ in the RMF model. The right panel shows an extension of the inset of the left panel. Smaller values of K are considered. For $K \geq 3$, all curves essentially show the same behavior. For $K = 2$ (no lines), however, $\pi_L(0)$ is almost constant for most of the L values considered and has a values around 0.5. Most surprising is the curve for $K = 1$. In this case $\pi_L(0)$ actually *increases*.

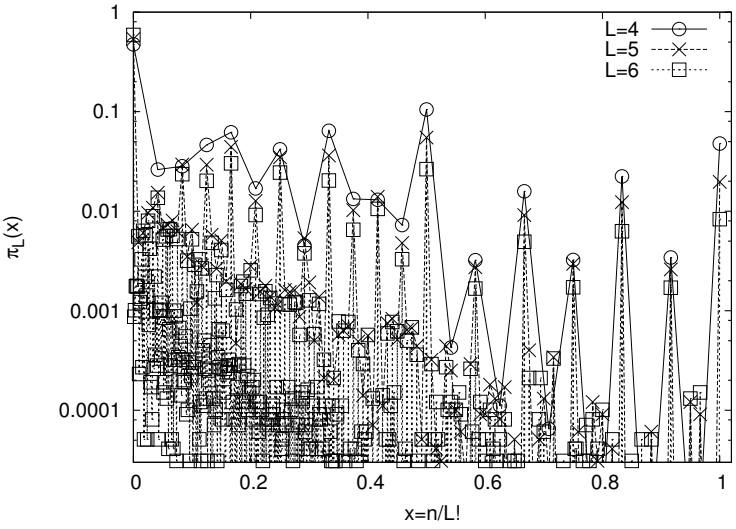


Figure 4.13.: Full distribution of the number of accessible paths in the LK model for $K = 1$, shown as function of the fraction of the total number of paths $x \equiv n/L!$. Note the peaks at $x = 1$ and $x = 0.5$.

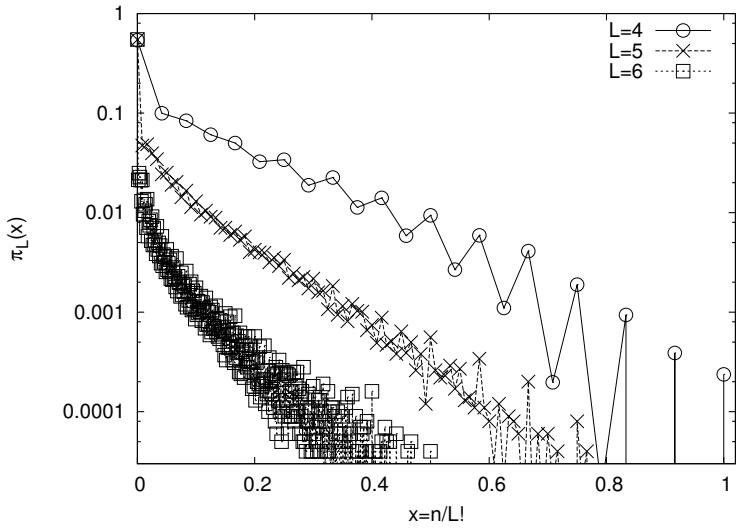


Figure 4.14.: Full distribution of the number of accessible paths in the LK model for $K = 2$, shown as function of the fraction of the total number of paths $x \equiv n/L!$. While the peaks are still discernible, their importance seems to have diminished.

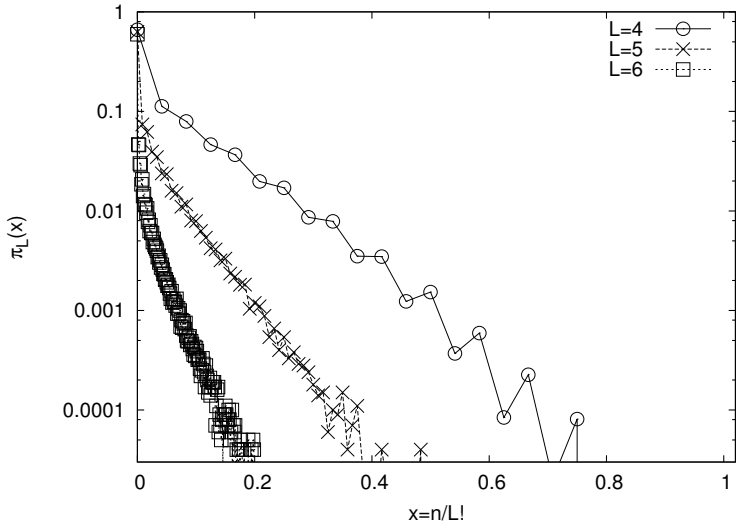


Figure 4.15.: Full distribution of the number of accessible paths in the LK model for $K = 3$, shown as function of the fraction of the total number of paths $x \equiv n/L!$. The peaks seem to have lost their importance in the sense that the distribution now looks smooth like those for the RMF or HoC model, see fig 4.1.

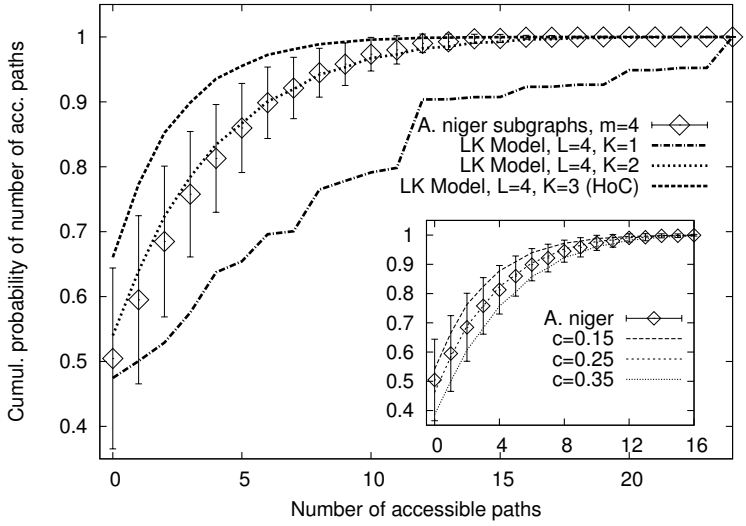


Figure 4.16.: The cumulative distribution of the number of accessible paths as found in the *A. niger* data set (symbols) compared to the curves from the LK (main plot) and RMF model (inset). Error bars on the empirical data were obtained by a resampling.

4.6. Comparison to empirical fitness landscapes

One of the main questions to be answered in this thesis is to what extent the behavior of these models can be found in empirical landscapes. The answer to this question is at first sight disappointing because even when using the subgraph method introduced in section 1.5.3 and validated in appendix C, the error bars on the expected number of accessible paths are such that both RMF (with $c > 0$) and LK model (and neutral model, even though this is not shown here as it will become clear that the dominant mechanism for blocking paths works via by the fitness *values*) can be fitted to the data[6], see figs 4.16 and 4.17. Thus none of the models considered here can be singled out as describing the empirical FLs best. Consequently, none of the models could be identified as the new standard model from which predictions about the features of empirical FLs are to be derived.
Upon closer examination, however, one sees that interestingly one model can be clearly rejected: The HoC model clearly does not adequately describe the expected number of accessible paths, see fig 4.16 and 4.17. In fig 4.16, the cumulative distribution of the number of accessible paths on subgraphs of size $m = 4$ is shown. They match (surprisingly, given simplicity of the models) well with both the LK and RMF models for parameter values corresponding to *intermediate* ruggedness. The same is true for the expected value of the number of accessible paths as function of sequence length (or subgraph size for the data) shown in fig 4.17.

This is a strong result, since the HoC model was the only one studied here that predicted $\pi_L(0)$ to monotonically grow as $L \to \infty$, while all other models predicted that with probability (close to) 1, there will be paths accessible in the limit of adaptation acting genome wide. In the light of eqs (4.3), (4.10) and (4.13), one can conjecture that the reason for the similar behavior across all models is caused by the fact that the probability P_L that one given path is accessible decays exponentially while the total number of possible paths grows as $L!$ with sequence length.

[6]because of the resolution of the data set

m	$\langle n \rangle_{\text{leth}}$	$\langle n_m \rangle$
2	1.61 (1.72)	0.82
3	4.05 (4.22)	1.34
4	12.53 (13.19)	2.01
5	55.32 (48.81)	3.16
6	246.0 (201.16)	6.07

Table 4.1.: Table of the mean number of accessible paths as function of subgraph size m if only lethals states can block a path (middle column) and if both lethal states and fitness values can block a path (right column).

If this were the reason, this would also apply to genome wide evolution as the topology of the state space is largely independent of the evolutionary dynamics[7] and the findings of this study would universally apply.
As was mentioned in section 2.2.1, the *A. niger* data set also contains some lethal states. To check the influence of these lethal states on accessibility, table 4.1 lists the number of accessible paths $\langle n_m \rangle_{\text{leth}}$ expected if *only* lethal states can block a path and compares them to the expected number of accessible paths computed from the data set if both fitness value and lethality can block a path. The values shown for $\langle n_m \rangle_{\text{leth}}$ are those found in the data set and (shown in parenthesis) those expected under the model for lethality presented in section 2.2.1, where one mutation was singled out as conferring a much higher chance of lethality if present in a given state. The values in parenthesis follow those found in the data set quite closely, indicating that the model does capture an essential part of the mechanism by which these lethals arise, consistent with table 2.1. However an explanation in terms of the biochemical function of that mutation has not been found so far.
From table 4.1, it is obvious that path accessibility is most strongly influenced by fitness values, since for example at $m = 4$, of the $4! = 24$ possible paths, about 12 are accessible if only lethals are considered as blocking a path (corresponding to a reduction by about one half), whereas if in addition to lethality also fitness values are allowed to block a path, the number of accessible paths reduces from 12 to 2 by about a factor 1/6, a much greater influence than that of lethals alone, see also [51].

However none of the empirical data sets offered a size or resolution sufficient to also verify the prediction that $\pi_L(0)$ eventually decreases. This would be very interesting to observe, as $\pi_L(0)$ is at least as important for the adaptation of a population on the FL as is the expected number of accessible paths.

4.7. Conclusions

In this chapter, the distribution of accessible paths on FLs was studied. It was in particular found that all models - except for the HoC model - considered, the GM of the FLs was found to be highly accessible in the sense that the FLs show both increasing $\langle n_L \rangle$ as well as (eventually)

[7] and for $\mu N > 1$ or with recombination, when the constraint that the population can only move between nearest neighbors is uplifted by e.g. allowing double mutants , even fewer constraint apply. The same is true when the selection strength decreases, such that also slightly deleterious steps are possible for the population.

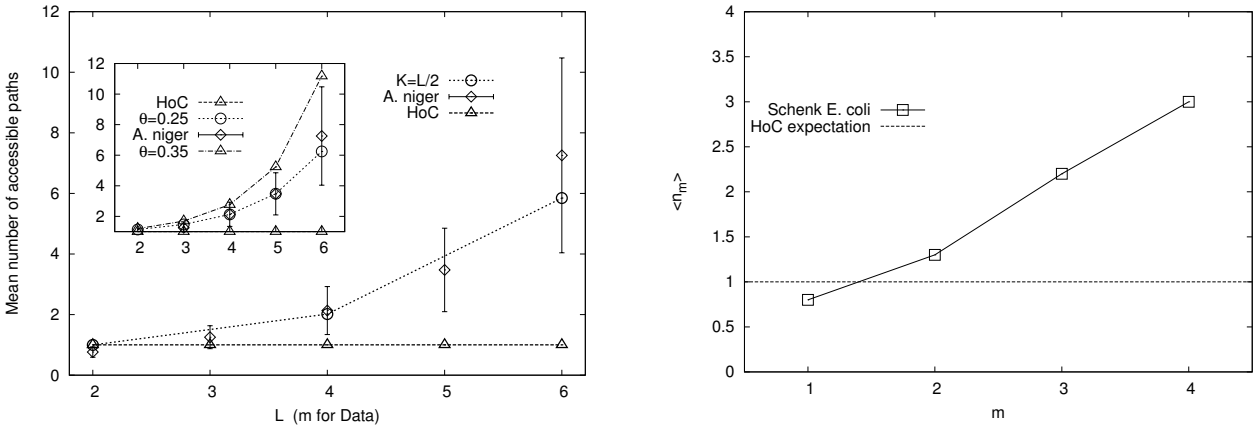


Figure 4.17.: The left figure shows the expected number of accessible paths of the *A. niger* landscape [30]. Note that for LK and RMF models, parameters can be chosen to match the empirical curves while the HoC model is rejected by the data. The right figure shows a similar comparison of the *E. coli* data set by Martijn Schenk (specifically the landscape made up of the mutations with big effect) introduced in section 2.2.2 on antibiotic resistance levels. Note that while no resampling procedure was possible since there was only one measurement for each point and hence no way of estimating a variance, the data set still clearly rejects the HoC hypothesis.

decreasing $\pi_L(0)$. The HoC model could be rejected by comparison to the empirical fitness landscapes of the model organisms *A. niger* and *E. coli*. In the RMF model, it was shown that even with an arbitrarily small positive c, the mean number of accessible paths grows with n while for the HoC $c = 0$, $\langle n_L \rangle = 1$ for all L. Thus a transition in the behavior of $\langle n_L \rangle$ occurs at $c = 0$. For $\pi_L(0)$, a similar transition in behavior was observed. While the value of c that separates these two regimes in the RMF models for the probability of finding no accessible paths could not be computed directly, comparison to the neutral model (see eqs (4.23) and (4.20)) suggests that also for the RMF model, this might be the same critical value $c = 0$ as found for the mean number of accessible paths. Thus it was found that for the vast majority of parameter values (in particular those that match the empirical data, see figs 4.16 and 4.17), the GM is accessible.

Supposing that the state space topology based explanation for this offered at the end of the last section is true, this would have important consequences for populations adapting on any fitness landscape and also beyond the strong selection/weak mutation regime considered here. For the LK-model, the results presented in this chapter indicate that for $L - K$ fixed, the resulting landscape will behave HoC-like ($\langle n_L \rangle$ constant, $\pi_L(0)$ increasing), while for K/L fixed and K fixed, it the FL will become similar to a smooth landscape for large L. This extends the range in which FLs from the LK model are known to 'essentially behave smooth' from the K-fixed regime (where this was known, see [76, 143] and references therein) to the case of L/K fixed.

The studies presented here and in particular the behavior of $\langle n_L \rangle$ also present a method of analyzing empirical FLs (or more generally energy landscapes such as spin glasses) independent of the adaptive dynamics. In considering how the expected number of accessible paths changes

with subgraph size m, one can tell whether a given empirical landscape is maximally epistatic or only displays an intermediate amount of epistasis, where the latter finding possibly implies that in the limit of genome wide adaptation, the landscape will essentially behave like a Mt Fuji landscape. With that respect, the distribution of accessible paths is just one more tool to quantify epistasis, see e.g. [19] for other measures of epistasis, but both $\langle n_L \rangle$ and $\pi_L(0)$ show a clear distinction between 'maximally' and 'less-then-maximally' epistatic landscapes, thus giving a direct interpretation of numerical values observed on empirical fitness landscapes, specially if the subgraph-method is used to obtain with the m-dependence of these objects an approximation of their respective L-behavior.

5. Basin of attraction of the global maximum

The statistics of the basin of attraction of the global maximum of the fitness landscape under greedy evolutionary dynamics (steepest ascent) is discussed.

Under greedy dynamics and with fitness values drawn from a continuous distribution as discussed in section 1.5.2, the fitness landscape separates into basins of attraction (BoA), each belonging to one (local or global) maximum. Clearly the size of the basin of attraction gives a hint at the importance of that maximum for the landscape. In particular if the basin of attraction of the global maximum comprises a substantial part of the FL (and is thus much larger than that of the local optima), this would imply that the FL is essentially dominated by its GM. In this section, results on the distribution of the GM for the different landscape models and the empirical FLs (from *Aspergillus niger* and *Escherichia coli*) are presented.

5.1. Model studies

The distribution of the BoA of the GM is studied relative to the whole state space for different models in this section. Analytically, this question has only been addressed in low-dimensional cases [18] and no attempt at a thorough analytic description is made here. Rather, the results of numerical studies are presented for different models. Both the full distribution and the expected size of the BoA of the GM are presented.

5.1.1. HoC and RMF models

Since the fitness of the global optimum is larger than that of any other maximum, the BoA of the GM consists at least of all L surrounding states. This bias imposed by the condition of considering the *global* maximum is still present even in the HoC model, as can be seen by the following consideration: The probability that a given state is a local maximum is $1/(L+1)$, since in the HoC model each of the $L+1$ states (the state considered plus L neighbors) has equal chances of being a maximum. Thus the expected number of local maxima is $2^L/(L+1)$, see e.g. [76]. Then the average size $\langle b \rangle_{HoC}$ of the basin of attraction of an arbitrary local optimum is simply given by dividing the total number of states by the expected number of local optima,

$$\langle b \rangle_{HoC} = \frac{2^L}{2^L/(L+1)} = L+1. \tag{5.1}$$

A given local optimum has to 'compete' for each of its L immediate neighbors with other local optimum and 'wins' only in a fraction of cases. On the other hand, as stated above, the BoA of the GM necessarily contains all L immediate neighbors and thus has a *minimal* size of $L+1$,

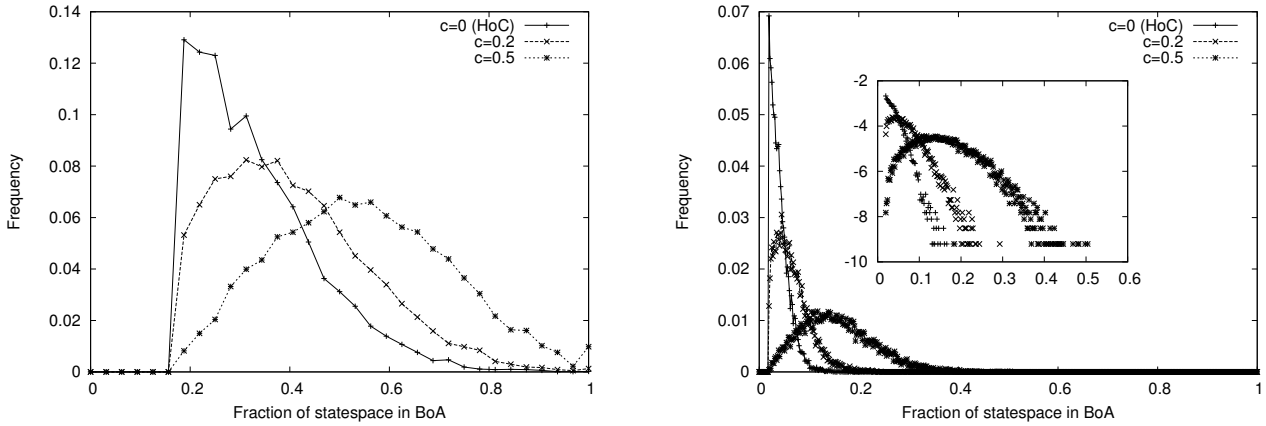


Figure 5.1.: Full distribution of the size of the BoA for $L = 5$ (left) and $L = 9$ (right) from numerical simulations of the HoC and RMF models. Note that the minimum size of the BoA is $L + 1$, thus the cutoff in this representation is at $(L + 1)/2^L$. In the semi-logarithmic scale used in the inset of the right panel, one sees the typical shape of the distribution for large values of c.

as can be seen by the sharp cutoff of the HoC curves in fig 5.1. Thus the expected size of the BoA of a local maximum is a lower bound for the size of the BoA of the GM,

$$\langle b_{GM} \rangle_{HoC} \geq \langle b \rangle_{HoC} = L + 1. \tag{5.2}$$

On the other hand, as can be seen in fig 5.1, the distribution of the BoA becomes in the HoC case very much concentrated around its lower cutoff at $L + 1$ as L increases. Thus it is not surprising that in numerical simulations, $\langle b_{GM} \rangle_{HoC}$ seems to stay close to the linear behavior found for the BoA of a local maximum in eq (5.1), see fig 5.2.

A similar reasoning for higher moments of the distribution was not possible for the same reasons that analytic considerations in chapter 4 were restricted to the mean number of accessible paths. The full distribution of the size of the BoA for the GM was obtained numerically in fig 5.1, where it is presented as function of the *fraction* of the size 2^L of the state space to allow for comparison between different values of L. In the two panels of fig 5.1, it is also seen that increasing the drift velocity c moves the distribution to the right. This can be explained, since for $c \to \infty$, the distributions are a sharp peak at 1, since with probablility one, in such landscapes there is only one maximum with the entire state space in its BoA.

5.1.2. LK model

For the LK model, analytic estimates for the number of local optima are available for the regime $1 \ll K \ll L$. [38, 44, 90] which rely on methods from [144]. For the LK model it is known that for fixed K and large L, the landscape is essentially dominated by the global optimum, see [76, 143] and references therein. Thus simply dividing the number of states by the expected number of local optima does yield the expected size of BoA of an arbitrarily chosen maximum. But since if the FL is dominated by the GM, one expects that the BoA of the GM will be much greater than that of an arbitrarily chosen local optimum, this does not contain much information about the BoA of the global optimum. However, the average size of BoA of an

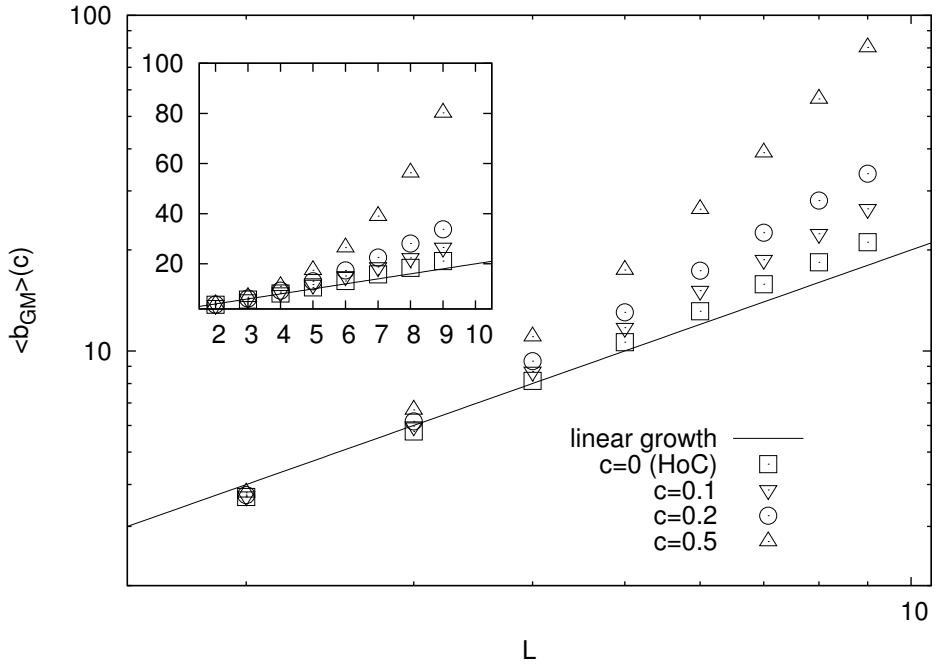


Figure 5.2.: The expected value of the BoA of the GM in HoC and RMF models for different values of drift c as direct plot (inset) and to double-logarithmic scales (main plot). While the linear form eq (5.1) seem to be a reasonable approximation in the direct plot, the double-log scale reveals that in the HoC case, $\langle b_{GM} \rangle$ increases slightly faster than linear. Nonetheless there is a clear distinction between the HoC behavior and that for $c = 0.1$ and greater,

arbitrary optimum can still be used as a lower bound on the size of BoA of the GM by dividing the number of states by the expected number of local optima. In [143], it was found that for fixed K and $L \to \infty$ and the neighborhood consisting of the adjacent states, the probability P_{opt} that a given states is a local optimum is given by

$$P_{opt} \approx \left[1 + 2\frac{K+2}{K+1}\ln(K+1)\right]^{1/2} \left[\left(\frac{1}{K+1}\right)^{1/(K+1)} \mathrm{erf}(\sqrt{2\ln(K+1)})\right]^{L} \tag{5.3}$$

with $\mathrm{erf}(\cdot)$ the standard error function. For K so large that $\mathrm{erf}\left(\sqrt{2\ln(K+1)}\right) \approx 1$, the expected number of local maxima $\langle N_{LM}\rangle$is given by

$$\langle N_{LM}\rangle \approx 2^L \sqrt{2\ln(K)} \left(\frac{1}{K+1}\right)^{L/(K+1)}. \tag{5.4}$$

Then, a lower bound for the size of the BoA of a local maximum is given by dividing the 2^L states by $\langle N_{LM}\rangle$ to yield

$$\langle b_{GM}\rangle > (K+1)^{L/(K+1)} \left(2\ln(K+1)\right)^{-1/2}. \tag{5.5}$$

For fixed K, this implies exponential growth of $\langle b_{GM}\rangle$ with L, which means that for large L, the LK model for fixed K behaves like a smooth landscape. Fig 5.3 shows a comparison of numerical simulations to the bound given by the equation above for the LK model. The numerical curves indicate that $\langle b_{GM}\rangle$ for fixed K grows exponentially with sequence length L (and thus in a similar way as the number of states 2^L in the configuration space). One observes that the bound is indeed much lower than the numerical curves, which means that the GM behaves differently from an arbitrarily chosen local maximum (supposing that the approximate bound, which was derived for the LK model with adjacent neighborhoods makes any statement at all about the model considered here numerically).

For all regimes other than that of fixed K, no analytic results for the LK model are available and all results presented for this model rely on numerical simulations alone.
Fig 5.4 shows the full distribution of the GM's BoA (again as function of the fraction of state space) for two different values of L and $K = 1$. There is a pronounced peak structure with much of the weight on configurations in which 75% or 100% of the state space belong to the BoA of the GM. The peak indicating that the entire state space belongs to the BoA corresponds to realizations where the GM is the *only* maximum of the landscape. The peak at 75% however, although it seems to belong to realizations with exactly one further local maximum, has thus far eluded explanation, as have the other peaks.
While the peak at 100% persists as the number of interacting sites is increased to $K = 2$, it carries very little weight and the curves change appearance, see fig 5.5. The structure is no longer dominated by individual peaks (which however remain visible as a feature of the distribution) but rather seems to approach the form observed for the HoC and RMF models. For $K = 3$, the peak corresponding to realizations in which the entire landscape is in the BoA of the GM vanishes and qualitatively the structure of the distribution of the BoA seems more similar to that observed for the RMF model, see fig 5.6. For $K = 4$, the shape of the distribution is very close to those obtained from the RMF model, see fig 5.7. In this case the

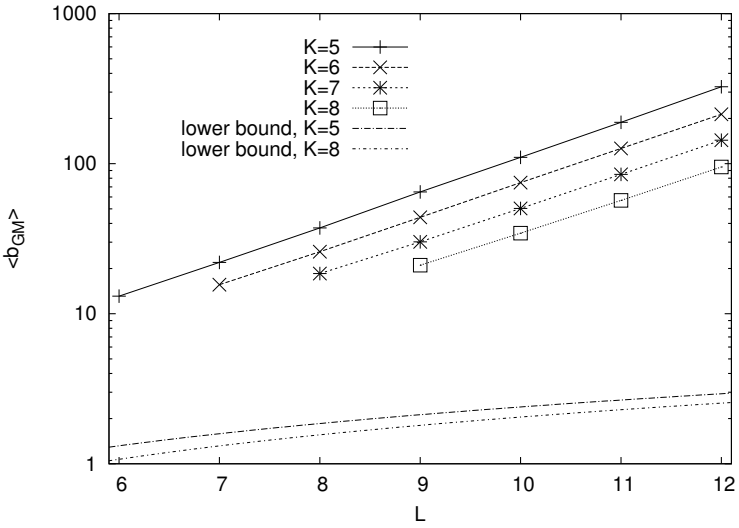


Figure 5.3.: For the LK-model in the regime of fixed K, $\langle b_{GM} \rangle$ is shown as function of L for various K. The line without points indicate the L-scaling of the lower bound in eq (5.5). $< b_{GM} >$ increases exponentially, but the lower bound is very far away from the curves, meaning that the GM behaves very differently from a randomly chosen maximum.

distribution of the size of the BoA from LK can be described by a RMF model with suitably chosen 'effective drift' c.

This change in behavior occurs in the same range of values of K for which $\pi_{L,K}(0)$ as considered in chapter 4 changed behavior strongly, see fig 4.12. Again one can speculate that this change in behavior is a sign of the transition in computational complexity of finding the height of the global maximum, which is further corroborated by the fact that for $K = 1$ and to some extent also for $K = 2$, in a non-vanishing number of realizations, the entire state space is the BoA of the GM, thus finding the GM just requires a greedy search (see figs 5.4 and 5.5). These speculations, however only apply to the regime of small K which appears to be of limited biological relevance.
The expected size of the BoA is shown in fig 5.8. For fixed difference $L - K$, the curves appear to be roughly linear in L and thus similar to the HoC-like case, see the inset on the left of fig 5.8. In the double-logarithmic scales in the main plot, however, one sees that the curves are not quite parallel. Nonetheless they seem linear in the double-log plot, which means that $\langle b_{GM} \rangle$ grows like a power law for $L - K$ fixed. In contrast, if K is scaled as fraction of L, the BoA clearly increases faster than linear with L. The main plot on the right of fig 5.8, again to double logarithmic scales, seems to show that the curves are not linear in the scale used there, which implies that $\langle b_{GM} \rangle$ grows faster than a power law, see also fig 5.3.
Thus the curves of $\langle b_{GM} \rangle$ behave in a way similar to the HoC case for $L - K$ fixed, while they behave quite differently for fixed ratio K/L. The contrast in behavior found here corresponds to the conjectured behavior change of the probability of finding no accessible path $\pi_L(0)$ found between HoC-type and less-than-maximally rugged FLs for the LK model discussed in chapter 4. There it was found that for the LK-model, any fixed difference $L-K$ behaves asymptotically like the HoC model where a fixed fraction L/K corresponds in the large L limit to a FL of

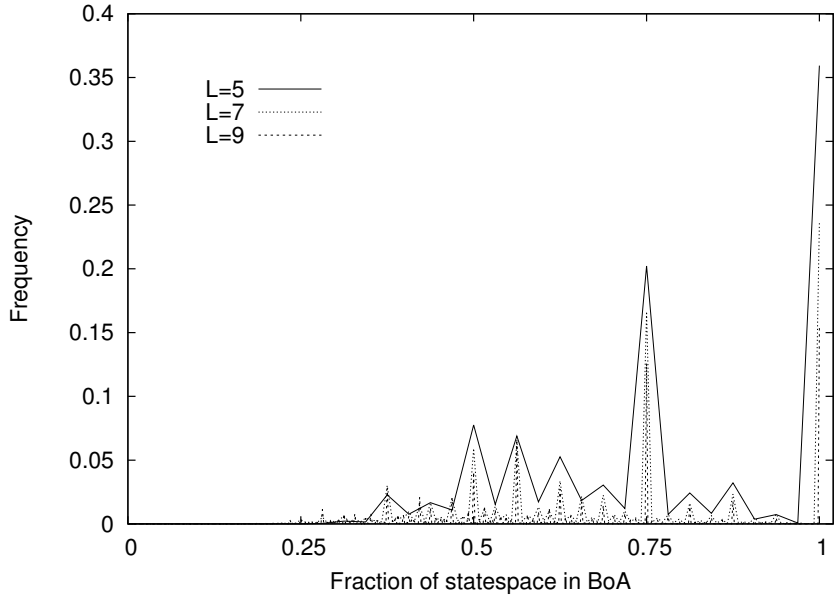


Figure 5.4.: The distribution of the fraction of state space belonging to the BoA for $K = 1$ and different values of L. Note that there are pronounced peaks indicating that realizations where 75% or even 100% of the state space belong to the BoA of the GM carry significant probability.

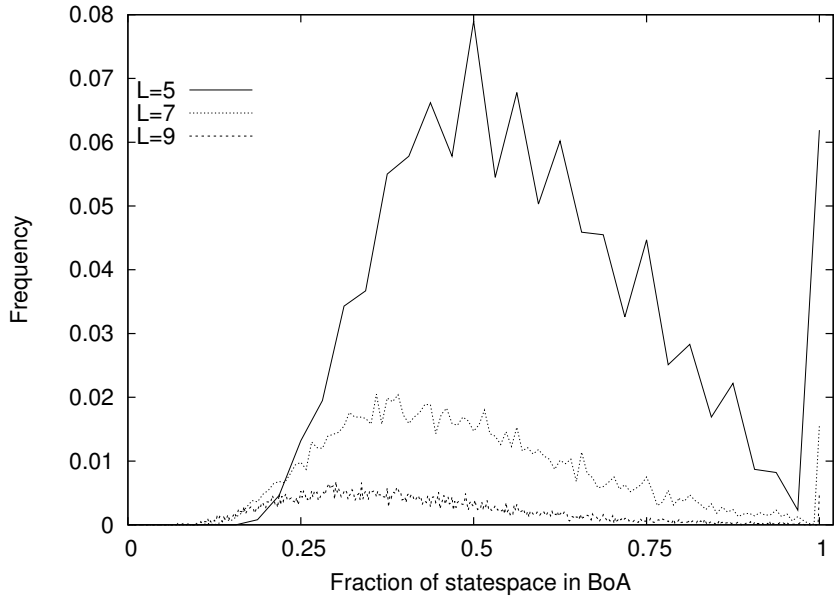


Figure 5.5.: The distribution of the fraction of state space belonging to the BoA for $K = 2$. Note that the peak at 100% is still present while the rest of the distribution approaches the form found for the RMF model, see fig 5.1.

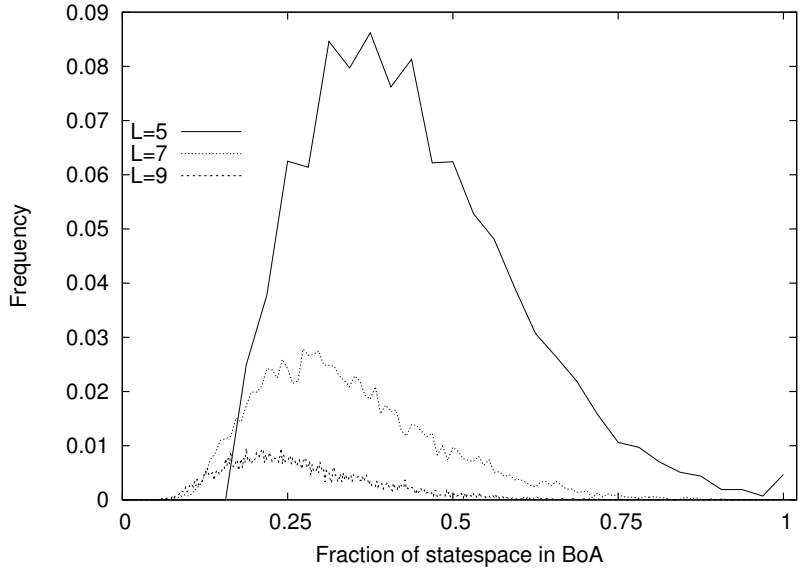


Figure 5.6.: For $K = 3$, the distribution strongly resembles those found for the RMF case: The overall structure is no longer dominated by individual peaks and comparison of the semi log plot to those of fig 5.1 suggests that a value for c can be found to resemble the shape of distribution observed for the LK model.

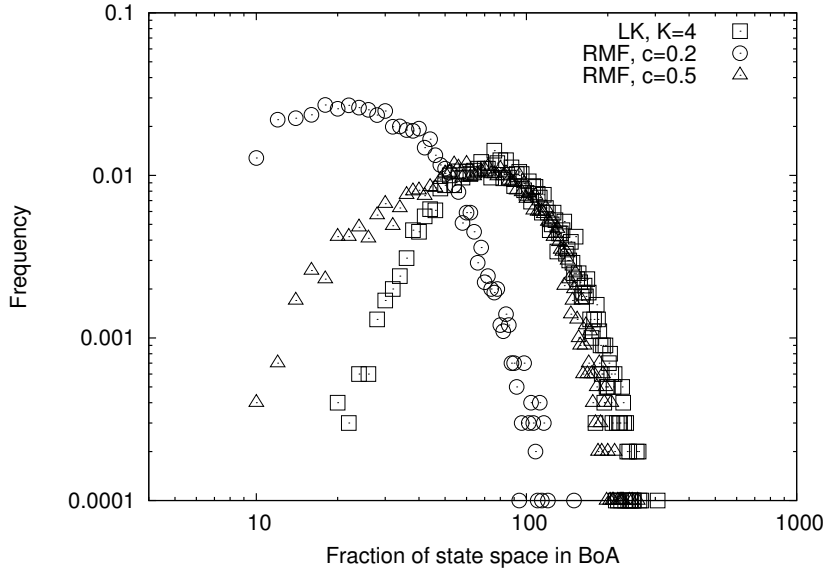


Figure 5.7.: Comparison between the full distribution of the BoA on landscapes from the LK and RMF model for different parameter values and $L = 9$. For the case of $K = 4$ shown here, the distribution from the LK landscapes seem similar to those from the RMF model. The points from the LK model almost seem to approach a Gaussian shape

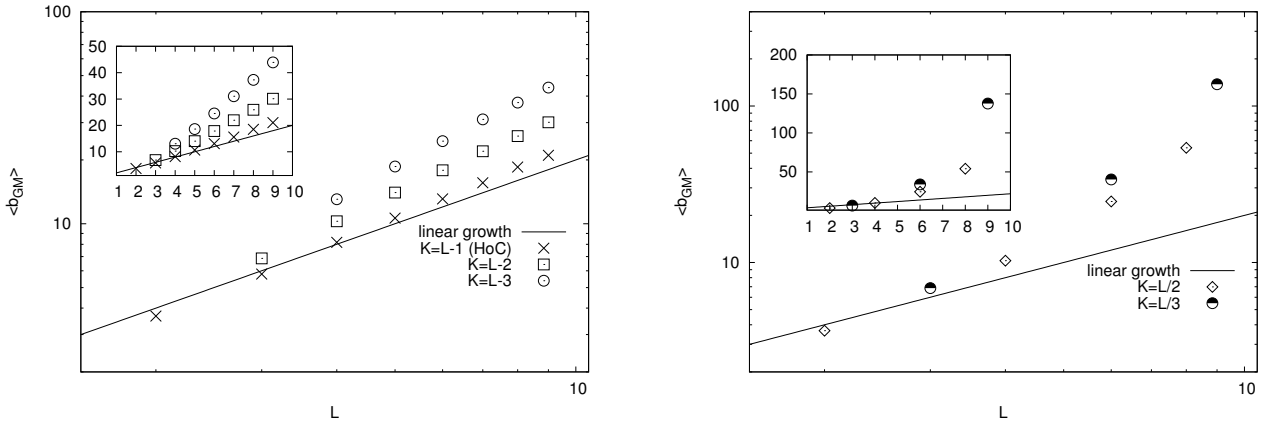


Figure 5.8.: $\langle b_{GM} \rangle$ as function of L for different values of K. In the left panel $L - K$ is fixed. While in the inset the curves seem linear with different coefficients, on the double logarithmic scales of the main plot, one sees that the curves for $L-K$ are not quite parallel. In the right panel, the fraction L/K is fixed. One clearly sees that $\langle b_{GM} \rangle$ grows much faster than linear. Also $\langle b_{GM} \rangle$ seems to show a curvature in this case.

intermediate ruggedness.

5.2. Study of the empirical fitness landscapes

Since in appendix C it is shown that the subgraph method can be used to obtain an estimate for the behavior of $\langle b_{GM} \rangle$, the behavior observed for the two models in figs 5.2 and 5.8 can be compared to empirical data. Fig 5.9 shows the expected size of the BoA as function of subgraph size m for the *A. niger* landscapes introduced in section 2.2.1. Even though the error bars allow in principle for a linear fit to the data (left panel), the curves to double-logarithmic scales presented in the right panel show that $\langle b_{GM} \rangle$ grows faster than linear, which implies that the empirical data are not of the HoC type but rather more strongly correlated. While it cannot be concluded that the mean size of the BoA grows faster than a power law (the error bars do not allow to say that $\langle b_{GM} \rangle$ is curved in the double logarthmic plot), at least it seems to grows with an exponent af about 1.5, which is faster than the behavior observed for the HoC curves (where it was closer to 1, see figs 5.2 and 5.8).
The curve for the *E. coli* data set shown in fig 5.10. on the other hand, shows a form compatible with a HoC hypothesis (for $\langle b_{GM} \rangle$, not so for the expected number of accessible paths, where the empirical findings are incompatible with the HoC hypothesis, see fig 4.17).

5.3. Conclusions

In this chapter it was found that the behavior of the size of the basin of attraction of the global optimum differs between fitness landscapes of maximal ruggedness (HoC type) and those with intermediate ruggedness. This distinction is very similar to that for the behavior of accessible paths discussed in chapter 4. However, due to a lack of analytical understanding of the behavior of $\langle b_{GM} \rangle$ even in the HoC case, this distinction cannot be made as strongly as that in chapter 4

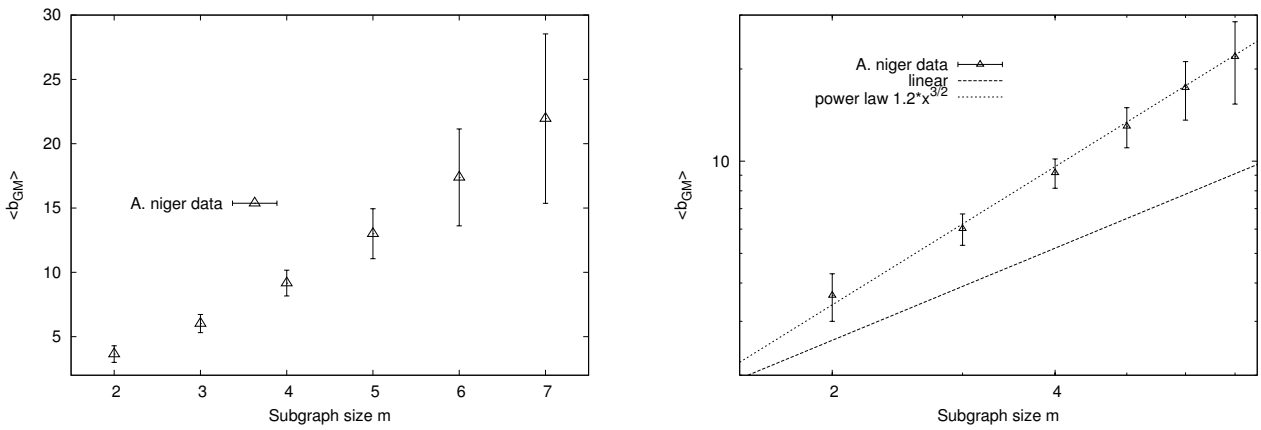


Figure 5.9.: Mean size of the BoA of the GM for subgraphs sizes between 2 and 6. The error bars were obtained by the resampling method. While in the left panel, linear behavior seems at first glance unlikely but cannot rejected due to the error bars, the double log plot on the right shows that curves are best represented by a power law with exponent of about 1.5.

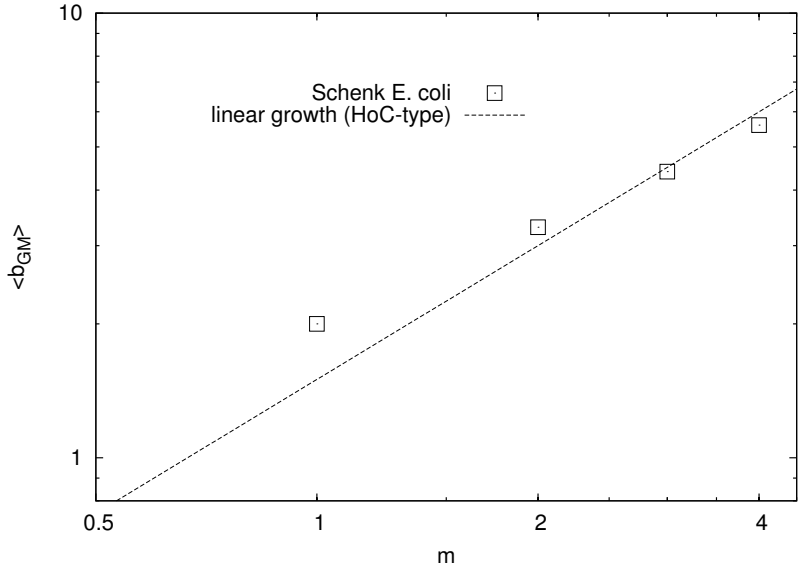


Figure 5.10.: The behavior of $\langle b_{GM} \rangle$ as function of the subgraph size m for $m = 1, 2, 3, 4$. On the double-logarithmic scale, the curves seem compatible with linear growth.

and only relies on numerical observations. Strictly speaking it is found here that $\langle b_{GM} \rangle$ seems to grow like a power law for HoC type models (see fig 5.8) and possibly faster than a power law for less-than-maximally rugged models of FLs. This pattern is consistent across the numerical simulations presented here, see fig 5.2.
In the LK model, the full distribution of the basin of attraction shown in figs 5.4 - 5.6 was shown to be dominated by peaks corresponding to realizations in which a substantial part of the landscape is part of the BoA of the GM. These peaks have eluded a full explanation this far. While the peak at 100% corresponds to realizations with only one maximum and the peak at 75% apparently to those with one additional local maximum, it is not easy to see why the presence of one further maximum should reduce the BoA of the GM by exactly 25% regardless of sequence length L. However this peak structure looses importance as K increases beyond $K = 2$ and for larger values of K, the distribution of the BoA has essentially the same shape as that from the RMF model for a suitably chosen value of c, see fig 5.7.
The analysis of the empirical *A. niger* landscape, see fig 5.9 seems to show a power-law increase of BoA size with an exponent larger than that observed for HoC-like models. This, much like the analysis of the expected number of accessible paths shown in the left panel of fig 4.17, is inconsistent with the hypothesis that the empirical landscape is of the HoC type. In contrast, the behavior of $\langle b_{GM} \rangle$ on subgraphs of the *E. coli* data set provided by Martijn Schenk showed a clear linear behavior in subgraph size, which implies that the HoC hypothesis cannot be rejected on these grounds, see fig 5.10.
The statements made about the behavior of $\langle b_{GM} \rangle$ in this chapter however are not as strong as those in chapter 4 because no analytical results on the behavior of $\langle b_{GM} \rangle_{HoC}$ for HoC landscapes were available beyond the lower bound presented in eq (5.1). A perturbative analysis within the RMF model around the HoC case $c = 0$ seems however feasible and is a very interesting open problem.

6. Conclusions

The main results of this thesis concern three different areas of science. The connection between these three at a first glance quite separate areas arose naturally in the course of investigations.

6.1. Record statistics

As was mentioned in chapter 3, the evolutionary adaptation in the weak mutation/strong selection regime of a population can be treated by considering it as a sequence of record values of the trait under selection. This connection has been observed many times before (see [87,111] and references therein, also [73,75,104] and references therein for the applications of the related field of extreme value theory to evolutionary biology). Here it led to study the record statistics of independent random variables with a linear trend.
The results presented in chapter 3 and separately published in [52,53,150] are part of a recent surge of interest in record and extreme value statistics (see e.g. [86,89,96,151] and references therein) beyond the standard setting of independent, identically distributed random variables [105]. The key results of chapter 3 can be summarized as follows:

- The complete universality of the ordering probability $P_n(c)$ with respect to the parental distribution $f(\cdot)$ of the entries observed in the iid case [105] ($c = 0$) breaks down for $c > 0$. For c small compared to the scale set by the parental distribution, $f(\cdot)$ only enters via a non-universal constant with no effect on the scaling with c, while for large c, the scaling of $P_n(c)$ with c depends strongly on $f(\cdot)$. The way in which the properties of the parental distribution determine the c scaling of $P_n(c)$ appear remotely similar to the way in which $f(\cdot)$ determines the universal distributions of *extreme value* statistics in the iid case [29]. This observation might be explained by the fact that the c-scaling of $P_n(c)$ for large c is, like the EVT behavior, determined by the tails of (the twofold convolution of) $f(\cdot)$. However, this observation might also just be an effect of the example distributions considered.

- Record events, which are statistically independent in the iid case, show non-trivial and quite counter-intuitive correlations for $c > 0$. The probability of observing a record event at a given time n was shown to depend on knowledge of previous record events. This dependency was surprisingly strong: The probability of a record *given* the occurrence of a record in the previous time step was in some cases observed to be well over twice the unconditioned record probability. Whether these correlations are such that record events attract or repel each other could asymptotically be classified depending on whether the parental distribution $f(\cdot)$ has heavy tails or bounded support respectively.

- It was shown that the relationship between heavy tails of the underlying probability density $f(\cdot)$ and attractive correlations between record events can be inverted: By randomly

choosing subsets of a given data set, artificially superimposing a linear trend and observing correlations between record events of the resulting 'time series', one has a record based distribution free test for heavy tails that works particularly well for small data sets. As proof of principle, this test was used to recover the well-known heavy tails of the distribution of paper citations [123]. Furthermore, it was used to show that the antibiotic resistance effects of mutations measured in the *E. coli* data set presented in section 2.2.2 are most likely drawn from a distribution with heavy tails.
This test only considers a subset of the data in each iteration and thus is potentially more robust to removal of outliers than existing methods such as maximum-likelihood estimators [23]. Furthermore, the test is distribution-free (or non-parametric), a property which allows to avoid controversies like the one that arose in context of animal foraging behavior [40, 141] because of possible biases in the estimator used [39]. With a non-parametric test, such a bias is not present.

6.2. Energy landscapes

The problems treated in chapter 4 and 5 arose naturally in the context of evolutionary biology but can be formulated independently of that specific context. If the fitness landscape is considered as a mapping from the Boolean hypercube $\{0,1\}^L$ of dimension L into the real numbers, it can be interpreted as a generalized energy landscape. In the context of spin glasses [100], topological properties of energy landscapes have been studied but, for the most part, these studies have been concerned with locally defined properties such as the question whether a given state is an extremum or a saddle point. While the size of the expected basin of attraction of the global maximum treated in chapter 5 has been considered before [18], it has not been in the center of attention thus far. The same is true for the distribution of accessible paths crossing the whole state space introduced in [19, 146], which arose quite naturally in the context of evolutionary biology. The latter question can be interpreted as an extension of the percolation problem [139] to a setting where the probability for a given step to be accessible depends on the energies associated with the two configurations to be connected.
Unlike the topological properties of spin glasses usually considered, the objects under investigation here are global in the sense that they are concerned with the GM rather than an arbitrarily chosen local maximum. For both the expected value of accessible paths and the mean size of the basin of attraction, it was shown or at least implied by numerical simulations that their behavior shows a marked difference depending on whether the landscape is of the HoC-type (including LK models with fixed value of $L-K$) or has an (effective) global trend towards the global optimum, regardless on whether this trend is put in explicitly as in the RMF model or arises implicitly like in the LK model with fixed value of L/K (for brevity these landscapes are referred to as 'with trend' even for those from the RMF model with infinitesimally small underlying trend). These changes can be summarized as follows:

- The expected number of accessible paths is independent of L if the landscape is of the HoC type. If on the other hand the landscape has a trend, the expected number of accessible paths grows with L.

- The probability $\pi_L(0)$ that none of the paths on the landscape is accessible increases

monotonically with L for HoC type landscapes, while for landscapes with an underlying trend, there is a sequence length beyond which $\pi_L(0)$ decreases with L and subsequently tends to zero.

- For HoC type landscapes, the expected size of the global optimum's basin of attraction grows with sequence length L slower (like a power law, numerical observations point to a power not much greater than 1) than for fitness landscapes with trend (where numerical simulations seem to indicate growth with a larger power or possibly even faster than any power law).

- On LK landscapes for fixed $K \leq 2$, the distributions considered (that of the basins of attraction and the number of accessible paths, see figs 4.13 through 4.15 and 5.4 through 5.6) show interesting peak structures. These structures vanish as K increases beyond 3 and for higher values, the distributions are very similar to those obtained from the RMF model. At the value of $K = 2$, the probability of having no accessible path in a given landscape shows a very interesting, counter-intuitive behavior as well, see fig 4.12.

The coherence between the two properties considered in this thesis allows to suppose that other properties will also show this division into two regimes.
It has been known for some time that there are non-trivial regimes of tunably rugged fitness landscapes like the LK model [76] or related models [143] where the global optimum dominates the landscape. These studies however only considered special cases of the models (such as LK-model with fixed value of K). In this thesis it could be shown that the global optimum of the landscape is of pronounced importance[1] even for fixed ratio L/K or landscapes with even an infinitesimally small underlying trend (RMF model for $c \ll 1$). On the one hand, this extends the parameter range for which this behavior has been observed (namely to the regime of K/L fixed) while on the other hand the findings for the RMF model allow to hope that similar results will hold for much more general energy landscapes, in particular when they implicitly give rise to an *effective* underlying trend with superimposed random contributions.

6.3. Evolutionary biology

The topological structures of the fitness landscapes considered in this thesis play a role for the adaptive fate of the population if and only if this adaptation takes place in the corresponding adaptive regime, i.e. if population size N, mutation rate μ and selection strength s have the appropriate values (see section 1.4.2). The statements that can be made on these grounds, although evident, will be listed in the following subsection.
Furthermore, it was established in this thesis that at least with respect to the topological features and models discussed in here[2], fitness landscapes can be classified as *either* HoC type *or* of the correlated type. These two classes were shown to behave very differently in the relevant limit of large L. Using the subgraph method established in appendix C, in this section it will be shown that most empirical fitness landscapes know to date can be classified as 'with trend'. If this were true in general, this could be an explanation for apparent inconsistencies

[1] even though saying that it 'dominates' the landscape appears to be a bit too strong a statement
[2] but quite possibly also for other properties of interest

between theoretical results and observations of the outcome of natural evolution, as will be pointed out at the end of this section.

6.3.1. Determining the evolutionary fate of a population under a given adaptive regime

Accessible paths in the sense used in this thesis and studied in chapter 4 only play a role in the adaptation of a population if the adaptation takes place in the strong selection/weak mutation regime.
Greedy dynamics are restricted to situations where selection is strong and the product μN of mutation rate (per individual genome) and population size is such that all one-mutants are present at all times but essentially no two mutations are (i.e. $\mu N \gg 1$ and $\mu^2 N \ll 1$). Then the basin of attraction of the GM studied in chapter 5 becomes important for the adaptation of the populations.

If an organism's fitness landscape can be classified as HoC type, genome wide ($L \gg 1$) adaptation will

- In the SS/WM regime have on average one accessible path regardless of the sequence length L, while a typical configuration will only with very small probability have accessible paths at all and
- In the greedy regime have a basin of attraction growing linearly with sequence length.

If adaptation takes place on a landscape of the type with trend, it will

- In the SS/WM regime always (with probability close to 1) have accessible paths and on average very large number of them and
- In the greedy regime have on average a basin of attraction of the global optimum that includes a much larger portion of state space than one would obtain by evenly distributing all states on the expected number of optima.

6.3.2. Classification of empirical fitness landscapes

No current technology and none that will be available in the near future is able to assess the full fitness landscape for all possible mutations of any natural organism. However with the method of analyzing subgraphs as established in appendix C, empirical fitness landscapes consisting of very few mutations can systematically be classified as either HoC type or 'with trend' by observing the behavior of the expected number of accessible paths and the expected size of the global optimum basin of attraction. For the *A. niger* landscape introduced in section 2.2.1 and the *E. coli* landscape from [134] introduced in section 2.2.2, this has already been done in chapters 4 and 5, see figs 4.17 and 5.9 respectively. Here the same analysis was performed for two different *E. coli* landscapes, Weinreich's famous β-lactamase antibiotic resistance data [146] and a recently published *E. coli* landscape [77] by Khan *et al.* which consisted of direct competition measurements (like the fitness values of the *A. niger* landscape [51] introduced in section

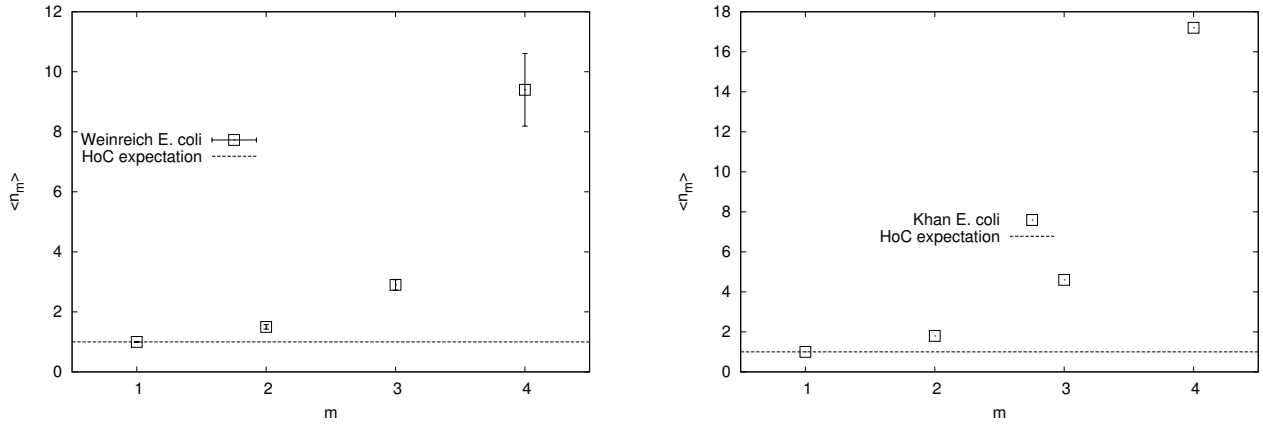


Figure 6.1.: Subgraph analysis of the empirical landscapes by Weinreich *et al.* (left) and Khan *et al.* (right). The expected number of accessible paths was observed on both landscapes to be incompatible with a HoC hypothesis. Error bars were obtained by the resampling method for the Weinreich landscape when estimates for a measurement error were available.

2.2.1 and in contrast to the other *E. coli* landscapes considered in this thesis, which consisted of antibiotic resistance measurements). It was found these two landscapes can be classified as 'with trend' with respect to both accessible paths and basin of attraction, see fig 6.1. In that figure, however, only the plots for the mean number of accessible paths are shown, as both landscapes only had one maximum (even under resampling for the Weinreich data) and thus the entire landscape was in the basin of attraction of the global optimum. Thus the basin of attraction grew like 2^m with m the subgraph size in both cases.

Together with the results in appendix C and the subgraph analysis presented in chapters 4 and 5 of the *A. niger* and *E. coli* landscapes introduced in chapter 2, this serves as proof of principle that empirical fitness landscapes can be classified in the ways presented in this thesis. The only requirement is sufficient size: While on data sets consisting of 5 mutations or more as used here, the analysis was possible, it proved inconclusive when applied to data sets of 4 mutations, like the malaria data set [93] or the *E. coli* competition data set [21].

Thus systematic subgraph analysis can be used as preliminary check on any fitness landscape, since it does not make any inherent assumptions about an adaptive regime.

6.3.3. Possible consequences for adaptation on arbitrary fitness landscapes

The HoC model was originally introduced not because it was believed to be particularly realistic but simply because in this framework, computations become significantly easier or even possible at all. It remains poplar because it is widely believed that results derived within this model will at least present a first-order approximation to naturally occurring fitness landscapes. With respect to the two objects considered in this thesis and quite possibly also with respect to other quantities of interest, this similarity was shown to hold only for small sequence lengths L. However, since natural genomes are on the order of millions to billions of base-pairs in length

and mutations with subsequent adaptation can essentially appear anywhere on these genomes, the case of relevance for natural adaptation is the limit of large L, where extrapolations from HoC based computations are questionable at best.
Intergenic epistasis was observed to be less pronounced than intragenic epistasis [101], which might imply that not all mutations have equal influence on fitness and is thus at odds with the HoC hypothesis. However, epistasis between mutations even on different chromosomes was observed to be far from vanishingly small [51], thus *a priori* no-one can legitimately argue that any fitness landscape should be in either class (HoC type or with trend). Nonetheless it was not altogether surprising to find that the empirical fitness landscapes considered here showed less ruggedness than the HoC type and were thus classified as 'with trend' in the sense defined above. If this were true in general, it would allow to resolve some apparent contradictions between predictions based on numerical simulations considering a limited number of mutations and observations on the outcome of genome-wide natural evolution.Two of these apparent contradictions are:

- Digital organisms of genome length $L = 100$ when given the freedom to evolve their own mutation rate were observed to have small mutation rates when adapting on a rugged fitness landscape [24]. They evaded deleterious mutations at the cost of long term evolvability. Only on very smooth landscapes could the digital organisms evolve to the mutation rate optimizing long-term adaptability.

- Simulations of population dynamics on the empirical *A. niger* landscape showed that recombination between different genotypes (sexual reproduction) can 'trap' the population at a local optimum, thus inhibiting adaptation towards the global optimum and giving populations without sexual reproduction a higher mean fitness [31]. This is at odds with the ubiquity of sexual reproduction.

As was pointed out by J. Arjan G. M. de Visser[3], the results on fitness landscapes with trend shown here can be interpreted that to genome wide evolution, any less-than-maximally rugged fitness landscape tends to look like a smooth landscape with, for example, many accessible paths leading to the global optimum and much of the landscape in the basin of attraction of the *global* optimum.
If this were true, it would mean that the two observations made above are finite size effects: Due to the technical limitations present in the studies, only fixed values of L were considered in both cases. In the limit of genome-wide adaptation, these problems would vanish for all fitness landscapes with trend such as those considered within this thesis[4].

6.4. Open problems

6.4.1. Technical problems

The most obvious open problems are technical in nature. These have proven very challenging.

[3] during a talk given at the Institute of Genetics at the university of Cologne in 2011
[4] For the *E. coli* landscape by Schenk *et al.*, this is only valid with the limitation that the HoC type could only be rejected with respect to $\langle n_L \rangle$ and not for $\langle b_{GM} \rangle$.

- Computing the probability that no accessible path is present in a given fitness landscape for models from the HoC and RMF models has thus far defied all efforts. Even in the auxiliary state space considered in chapter 4, only an implicit fix point equation could be set up but not solved. A more thorough treatment would be most interesting.

- The basin of attraction of the global optimum has so far only been treated numerically. Any analytical approach to the large L-behavior of this quantity would be most welcome.

- The underlying trend found to arise implicitly in LK landscapes where L/K was kept fix might only be a feature of the LK model. It would be very interesting to find out if any model such as p-spin models or other variants of the LK model fall into one of the two categories HoC or correlated with an *effective* linear trend.

- Explain the transition in behavior of the LK model at $K = 2$. Is this related to transitions in other spin glasses?

6.4.2. Biological problems

On the biological side, the open problems are harder to define and at the same time potentially of higher relevance.

- Are empirical fitness landscapes generally in the class with trend? While all those considered here are (as well as [101]) were, no arguments could be found why this should generally be the case. In all cases where a new empirical fitness landscape is presented, the subgraph method could be employed to check the class to which this landscape belongs.

- Does the classification of FLs as either 'HoC-type' or 'with trend' presented here play a role for the adaptation of populations? In particular, is it true that the contradictory results of adaptive processes vanish as the sequence length grows? To decide this question, one could simulate adaptive processes for populations with genomes of varying sequence length and observe if the contradictions with empirical observations mentioned above (suboptimal mutation rates for populations or disadvantage of sexual reproduction on rugged fitness landscapes) diminish as the sequence length increases. However, given that the properties discussed here do not always behave monotonically with sequence length (see e.g. the behavior of $\pi_L(0)$ in chapter 4), one might have to go to extremely large systems to see such a limiting behavior. This program might therefore meet severe computational challenges.

A. Tables and figures of the Fitness landcscapes

The table of fitness values of the *Aspergillus niger* data set is provided. Some instances $m = 4$ subgraphs of the full empirical landscape are shown as examples for the wide range of behavior they display.

A.1. Fitness values

For each combination of mutations, two replicate measurements of the mycelial growth rate were performed. The mean value is given, the variance around this mean was used to estimate error bars by the resampling method, see Appendix D. The effects of the mutations and the abbreviations are introduced in table 2.1. Zero fitness indicates that the corresponding state was not observed among the segregants and was thus assumed to be not viable, see the discussion in section 2.2.1.

fwn	arg	pyr	leu	phe	lys	oli	crn	mean	rep 1	rep 2
0	0	0	0	0	0	0	0	16.9	16.28	17.52
1	0	0	0	0	0	0	0	12.69	12.79	12.59
0	1	0	0	0	0	0	0	13.06	14.21	11.91
0	0	1	0	0	0	0	0	11.79	12.71	10.87
0	0	0	1	0	0	0	0	12.79	12.38	13.2
0	0	0	0	1	0	0	0	13.97	13.03	14.9
0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	1	0	10.11	9.03	11.19
0	0	0	0	0	0	0	1	10.62	10.46	10.78
1	1	0	0	0	0	0	0	14.04	13.89	14.18
1	0	1	0	0	0	0	0	12.93	12.75	13.11
1	0	0	1	0	0	0	0	14.03	14.03	14.03
1	0	0	0	1	0	0	0	16.45	16.85	16.04
1	0	0	0	0	1	0	0	12.62	12.27	12.97
1	0	0	0	0	0	1	0	12.13	11.9	12.35
1	0	0	0	0	0	0	1	15.15	14.34	15.95
0	1	1	0	0	0	0	0	12.64	13.02	12.25
0	1	0	1	0	0	0	0	12.34	13	11.68
0	1	0	0	1	0	0	0	14.42	15.09	13.74
0	1	0	0	0	1	0	0	0	0	0
0	1	0	0	0	0	1	0	10.11	10.65	9.56
0	1	0	0	0	0	0	1	12.87	13.11	12.63
0	0	1	1	0	0	0	0	11.25	10.85	11.65
0	0	1	0	1	0	0	0	12.28	12.64	11.91
0	0	1	0	0	1	0	0	12.21	12.7	11.71
0	0	1	0	0	0	1	0	8.65	8.46	8.84
0	0	1	0	0	0	0	1	12.15	12.17	12.13
0	0	0	1	1	0	0	0	12.35	12.45	12.25
0	0	0	1	0	1	0	0	12.27	12.27	12.27
0	0	0	1	0	0	1	0	10.22	9.94	10.49
0	0	0	1	0	0	0	1	11.36	11.6	11.12
0	0	0	0	1	1	0	0	10.98	10.22	11.73
0	0	0	0	1	0	1	0	10.41	10.77	10.05
0	0	0	0	1	0	0	1	0	0	0
0	0	0	0	0	1	1	0	0	0	0
0	0	0	0	0	1	0	1	14.01	14.47	13.55
0	0	0	0	0	0	1	1	15.97	16.72	15.22
1	1	1	0	0	0	0	0	13.23	13.45	13
1	1	0	1	0	0	0	0	0	0	0
1	1	0	0	1	0	0	0	12.68	12.96	12.39
1	1	0	0	0	1	0	0	12.24	11.99	12.48

fwn	arg	pyr	leu	phe	lys	oli	crn	mean	rep 1	rep 2
1	1	0	0	0	0	1	0	9.19	8.38	9.99
1	1	0	0	0	0	0	1	13.77	13.72	13.81
1	0	1	1	0	0	0	0	9.74	10.33	9.14
1	0	1	0	1	0	0	0	10.75	11.52	9.97
1	0	1	0	0	1	0	0	12.99	13.12	12.85
1	0	1	0	0	0	1	0	8.72	8.14	9.29
1	0	1	0	0	0	0	1	13.95	13.35	14.54
1	0	0	1	1	0	0	0	12.42	13.02	11.81
1	0	0	1	0	1	0	0	0	0	0
1	0	0	1	0	0	1	0	8.83	7.28	10.38
1	0	0	1	0	0	0	1	11.94	12.9	10.97
1	0	0	0	1	1	0	0	11.93	12.02	11.84
1	0	0	0	1	0	1	0	7.92	8.51	7.32
1	0	0	0	1	0	0	1	14.28	14.32	14.23
1	0	0	0	0	1	1	0	0	0	0
1	0	0	0	0	1	0	1	12.64	12.64	12.64
1	0	0	0	0	0	1	1	12.47	12.81	12.13
0	1	1	1	0	0	0	0	11.27	10.54	11.99
0	1	1	0	1	0	0	0	13.04	12.86	13.22
0	1	1	0	0	1	0	0	0	0	0
0	1	1	0	0	0	1	0	9.45	9.76	9.13
0	1	1	0	0	0	0	1	11.69	11.6	11.78
0	1	0	1	1	0	0	0	14.99	14.5	15.47
0	1	0	1	0	1	0	0	0	0	0
0	1	0	1	0	0	1	0	9.49	10.34	8.64
0	1	0	1	0	0	0	1	10.6	10.68	10.51
0	1	0	0	1	1	0	0	0	0	0
0	1	0	0	1	0	1	0	10.7	9.77	11.63
0	1	0	0	1	0	0	1	13.13	13.72	12.54
0	1	0	0	0	1	1	0	7.73	6.41	9.04
0	1	0	0	0	1	0	1	13.93	13.83	14.03
0	1	0	0	0	0	1	1	9.58	9.49	9.67
0	0	1	1	1	0	0	0	12.44	12.26	12.62
0	0	1	1	0	1	0	0	12.48	13.07	11.88
0	0	1	1	0	0	1	0	6.44	6.33	6.54
0	0	1	1	0	0	0	1	11.1	11.48	10.71
0	0	1	0	1	1	0	0	0	0	0
0	0	1	0	1	0	1	0	8.86	9.08	8.63
0	0	1	0	1	0	0	1	13.61	14.32	12.9
0	0	1	0	0	1	1	0	0	0	0
0	0	1	0	0	1	0	1	14.64	14.33	14.94

fwn	arg	pyr	leu	phe	lys	oli	crn	mean	rep 1	rep 2
0	0	1	0	0	0	1	1	7	7.18	6.81
0	0	0	1	1	1	0	0	0	0	0
0	0	0	1	1	0	1	0	9.39	9.89	8.88
0	0	0	1	1	0	0	1	11.07	10.95	11.19
0	0	0	1	0	1	1	0	0	0	0
0	0	0	1	0	1	0	1	13.44	12.71	14.16
0	0	0	1	0	0	1	1	7.62	7.51	7.72
0	0	0	0	1	1	1	0	0	0	0
0	0	0	0	1	1	0	1	15.69	16.34	15.04
0	0	0	0	1	0	1	1	0	0	0
0	0	0	0	0	1	1	1	10.23	10.87	9.59
1	1	1	1	0	0	0	0	10.52	11.04	9.99
1	1	1	0	1	0	0	0	12.42	12.42	12.42
1	1	1	0	0	1	0	0	0	0	0
1	1	1	0	0	0	1	0	8.85	8.93	8.76
1	1	1	0	0	0	0	1	13.24	13.85	12.63
1	1	0	1	1	0	0	0	14.37	14.88	13.86
1	1	0	1	0	1	0	0	0	0	0
1	1	0	1	0	0	1	0	8.31	8.88	7.73
1	1	0	1	0	0	0	1	11.84	11.73	11.94
1	1	0	0	1	1	0	0	13.01	14.52	11.49
1	1	0	0	1	0	1	0	0	0	0
1	1	0	0	1	0	0	1	14.15	13.99	14.31
1	1	0	0	0	1	1	0	0	0	0
1	1	0	0	0	1	0	1	14.22	13.55	14.88
1	1	0	0	0	0	1	1	12.21	11.72	12.69
1	0	1	1	1	0	0	0	0	0	0
1	0	1	1	0	1	0	0	0	0	0
1	0	1	1	0	0	1	0	4.63	4.84	4.42
1	0	1	1	0	0	0	1	11.15	11.41	10.88
1	0	1	0	1	1	0	0	11.4	11.75	11.05
1	0	1	0	1	0	1	0	9.11	9.24	8.98
1	0	1	0	1	0	0	1	13.81	14.03	13.59
1	0	1	0	0	1	1	0	0	0	0
1	0	1	0	0	1	0	1	12.56	12.73	12.38
1	0	1	0	0	0	1	1	10.81	11.61	10.01
1	0	0	1	1	1	0	0	0	0	0
1	0	0	1	1	0	1	0	10.88	9.22	12.53
1	0	0	1	1	0	0	1	12.23	11.94	12.51
1	0	0	1	0	1	1	0	0	0	0
1	0	0	1	0	1	0	1	13.48	13.33	13.63

fwn	arg	pyr	leu	phe	lys	oli	crn	mean	rep 1	rep 2
1	0	0	1	0	0	1	1	8.44	8.36	8.51
1	0	0	0	1	1	1	0	0	0	0
1	0	0	0	1	1	0	1	12.66	13.2	12.11
1	0	0	0	1	0	1	1	10.19	10	10.38
1	0	0	0	0	1	1	1	11	11.24	10.75
0	1	1	1	1	0	0	0	11.51	11.82	11.2
0	1	1	1	0	1	0	0	0	0	0
0	1	1	1	0	0	1	0	8.13	6.36	9.9
0	1	1	1	0	0	0	1	12.35	11.77	12.92
0	1	1	0	1	1	0	0	0	0	0
0	1	1	0	1	0	1	0	7.47	7.46	7.48
0	1	1	0	1	0	0	1	13	14	12
0	1	1	0	0	1	1	0	0	0	0
0	1	1	0	0	1	0	1	13.72	14.11	13.32
0	1	1	0	0	0	1	1	7.14	7.93	6.35
0	1	0	1	1	1	0	0	0	0	0
0	1	0	1	1	0	1	0	7.97	8.44	7.49
0	1	0	1	1	0	0	1	0	0	0
0	1	0	1	0	1	1	0	0	0	0
0	1	0	1	0	1	0	1	12.68	12.71	12.65
0	1	0	1	0	0	1	1	8.09	8.49	7.68
0	1	0	0	1	1	1	0	0	0	0
0	1	0	0	1	1	0	1	14.64	14.83	14.45
0	1	0	0	1	0	1	1	0	0	0
0	1	0	0	0	1	1	1	0	0	0
0	0	1	1	1	1	0	0	0	0	0
0	0	1	1	1	0	1	0	6.94	6.47	7.4
0	0	1	1	1	0	0	1	11.87	12.07	11.66
0	0	1	1	0	1	1	0	0	0	0
0	0	1	1	0	1	0	1	13.41	14.03	12.79
0	0	1	1	0	0	1	1	7	6.99	7
0	0	1	0	1	1	1	0	0	0	0
0	0	1	0	1	1	0	1	14.31	14.71	13.91
0	0	1	0	1	0	1	1	8.88	9.08	8.67
0	0	1	0	0	1	1	1	11.14	10.94	11.34
0	0	0	1	1	1	1	0	0	0	0
0	0	0	1	1	1	0	1	12.97	12.85	13.09
0	0	0	1	1	0	1	1	7.66	7.66	7.66
0	0	0	1	0	1	1	1	7.82	7.51	8.12
0	0	0	0	1	1	1	1	0	0	0
1	1	1	1	1	0	0	0	10.22	9.97	10.47

fwn	arg	pyr	leu	phe	lys	oli	crn	mean	rep 1	rep 2
1	1	1	1	0	1	0	0	0	0	0
1	1	1	1	0	0	1	0	7.27	7.27	7.27
1	1	1	1	0	0	0	1	11.59	12.17	11
1	1	1	0	1	1	0	0	10.8	10.43	11.16
1	1	1	0	1	0	1	0	9.47	8.66	10.28
1	1	1	0	1	0	0	1	11.94	11.48	12.4
1	1	1	0	0	1	1	0	0	0	0
1	1	1	0	0	1	0	1	13.43	13.74	13.11
1	1	1	0	0	0	1	1	10.76	10.76	10.76
1	1	0	1	1	1	0	0	12.28	12.21	12.35
1	1	0	1	1	0	1	0	10.21	10.44	9.98
1	1	0	1	1	0	0	1	12.21	12.04	12.38
1	1	0	1	0	1	1	0	0	0	0
1	1	0	1	0	1	0	1	10.33	10.02	10.63
1	1	0	1	0	0	1	1	10.43	10.77	10.08
1	1	0	0	1	1	1	0	0	0	0
1	1	0	0	1	1	0	1	13.92	13.77	14.07
1	1	0	0	1	0	1	1	10.79	10.69	10.88
1	1	0	0	0	1	1	1	0	0	0
1	0	1	1	1	1	0	0	0	0	0
1	0	1	1	1	0	1	0	5.7	5.87	5.53
1	0	1	1	1	0	0	1	10.93	11.8	10.06
1	0	1	1	0	1	1	0	0	0	0
1	0	1	1	0	1	0	1	11.84	12.56	11.12
1	0	1	1	0	0	1	1	6.91	8.29	5.53
1	0	1	0	1	1	1	0	0	0	0
1	0	1	0	1	1	0	1	15.69	16.04	15.34
1	0	1	0	1	0	1	1	7.54	8.84	6.24
1	0	1	0	0	1	1	1	10.46	11.28	9.63
1	0	0	1	1	1	1	0	0	0	0
1	0	0	1	1	1	0	1	12.16	11.76	12.56
1	0	0	1	1	0	1	1	9.07	8.23	9.91
1	0	0	1	0	1	1	1	0	0	0
1	0	0	0	1	1	1	1	10.39	10.34	10.44
0	1	1	1	1	1	0	0	0	0	0
0	1	1	1	1	0	1	0	6.54	6.57	6.51
0	1	1	1	1	0	0	1	11.31	11.84	10.77
0	1	1	1	0	1	1	0	0	0	0
0	1	1	1	0	1	0	1	13.17	12.89	13.45
0	1	1	1	0	0	1	1	7.04	7.44	6.63
0	1	1	0	1	1	1	0	0	0	0

fwn	arg	pyr	leu	phe	lys	oli	crn	mean	rep 1	rep 2
0	1	1	0	1	1	0	1	13.98	14.44	13.51
0	1	1	0	1	0	1	1	9.5	10.35	8.65
0	1	1	0	0	1	1	1	10.45	10.77	10.12
0	1	0	1	1	1	1	0	0	0	0
0	1	0	1	1	1	0	1	12.51	13.74	11.27
0	1	0	1	1	0	1	1	7.93	7.93	7.39
0	1	0	1	0	1	1	1	9.1	8.72	9.47
0	1	0	0	1	1	1	1	10.39	9.91	10.86
0	0	1	1	1	1	1	0	0	0	0
0	0	1	1	1	1	0	1	10.46	10.73	10.18
0	0	1	1	1	0	1	1	7.01	7.24	6.77
0	0	1	1	0	1	1	1	8.74	8.79	8.69
0	0	1	0	1	1	1	1	10.91	11.77	10.04
0	0	0	1	1	1	1	1	7.68	8.05	7.3
1	1	1	1	1	1	0	0	11.84	11.76	11.91
1	1	1	1	1	0	1	0	5.82	5.92	5.72
1	1	1	1	1	0	0	1	11.67	11.94	11.4
1	1	1	1	0	1	1	0	0	0	0
1	1	1	1	0	1	0	1	11.19	11.04	11.33
1	1	1	1	0	0	1	1	7.49	7.72	7.25
1	1	1	0	1	1	1	0	0	0	0
1	1	1	0	1	1	0	1	12.34	12.81	11.86
1	1	1	0	1	0	1	1	10.35	10.14	10.55
1	1	1	0	0	1	1	1	10.63	11.1	10.16
1	1	0	1	1	1	1	0	5.38	4.95	5.81
1	1	0	1	1	1	0	1	14.22	14.03	14.4
1	1	0	1	1	0	1	1	0	0	0
1	1	0	1	0	1	1	1	0	0	0
1	1	0	0	1	1	1	1	10.3	9.44	11.16
1	0	1	1	1	1	1	0	0	0	0
1	0	1	1	1	1	0	1	11.32	11.38	11.25
1	0	1	1	1	0	1	1	9.22	8.57	9.86
1	0	1	1	0	1	1	1	0	0	0
1	0	1	0	1	1	1	1	8.09	7.93	8.25
1	0	0	1	1	1	1	1	8.03	7.66	8.4
0	1	1	1	1	1	1	0	0	0	0
0	1	1	1	1	1	0	1	13.36	13.69	13.02
0	1	1	1	1	0	1	1	7.94	9.32	6.55
0	1	1	1	0	1	1	1	0	0	0
0	1	1	0	1	1	1	1	0	0	0
0	1	0	1	1	1	1	1	8.68	7.75	9.61

fwn	arg	pyr	leu	phe	lys	oli	crn	mean	rep 1	rep 2
0	0	1	1	1	1	1	1	0	0	0
0	1	1	1	1	1	1	1	0	0	0
1	0	1	1	1	1	1	1	0	0	0
1	1	0	1	1	1	1	1	7.7	7.41	7.99
1	1	1	0	1	1	1	1	10.87	9.96	11.78
1	1	1	1	0	1	1	1	7.55	8.11	6.98
1	1	1	1	1	0	1	1	7.57	7.87	7.27
1	1	1	1	1	1	0	1	0	0	0
1	1	1	1	1	1	1	0	0	0	0
1	1	1	1	1	1	1	1	0	0	0

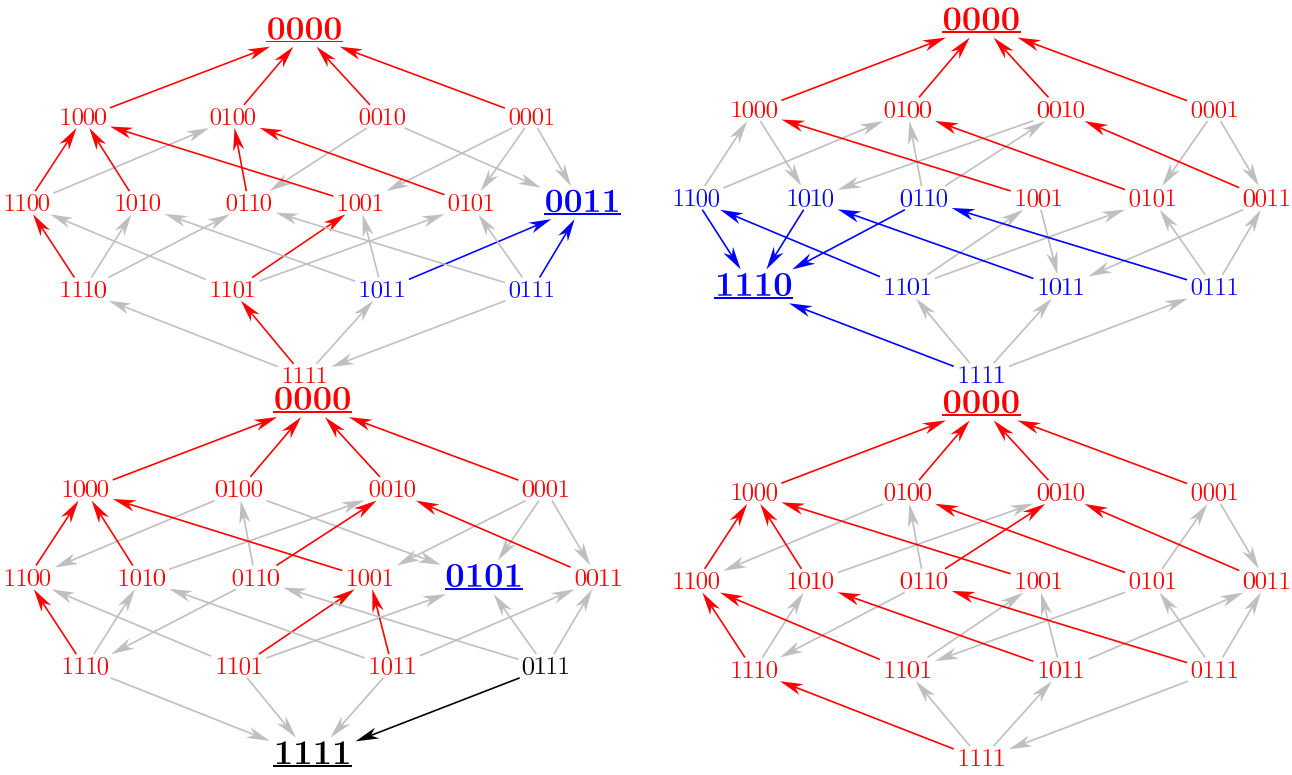


Figure A.1.: Subgraphs consisting of the mutations (from left to right in each row) arg, leu, oli, crn, arg, leu, phe, oli (top row), arg, pyr, leu, crn and arg, pyr, leu, oli (bottom row).

A.2. Examples of subgraphs

In this section, some examples of $m = 4$ subgraphs are presented. Again, the arrows point to states of higher fitness and color indicate that a state belongs to the basin of attraction of the corresponding maximum (underlined states).

The purpose of this presentation is to emphasize the wide range of behavior found in these landscapes, ranging from those with only one maximum (and consequently all states that maximum's BoA) and many accessible paths to those with several local maxima and no accessible paths.

Only complete subgraphs were considered, i.e. those containing no lethal states. Note that none of the subgraphs presented here contains the $lysD25$ mutation.

Missing arrows between two states at Hamming distance 1 indicate that those two states had equal fitness. This possibility had been out-ruled in analytical studies by considering only continuous distributions. However, since the empirical fitness values are only given to two decimal places, such a situation can arise.

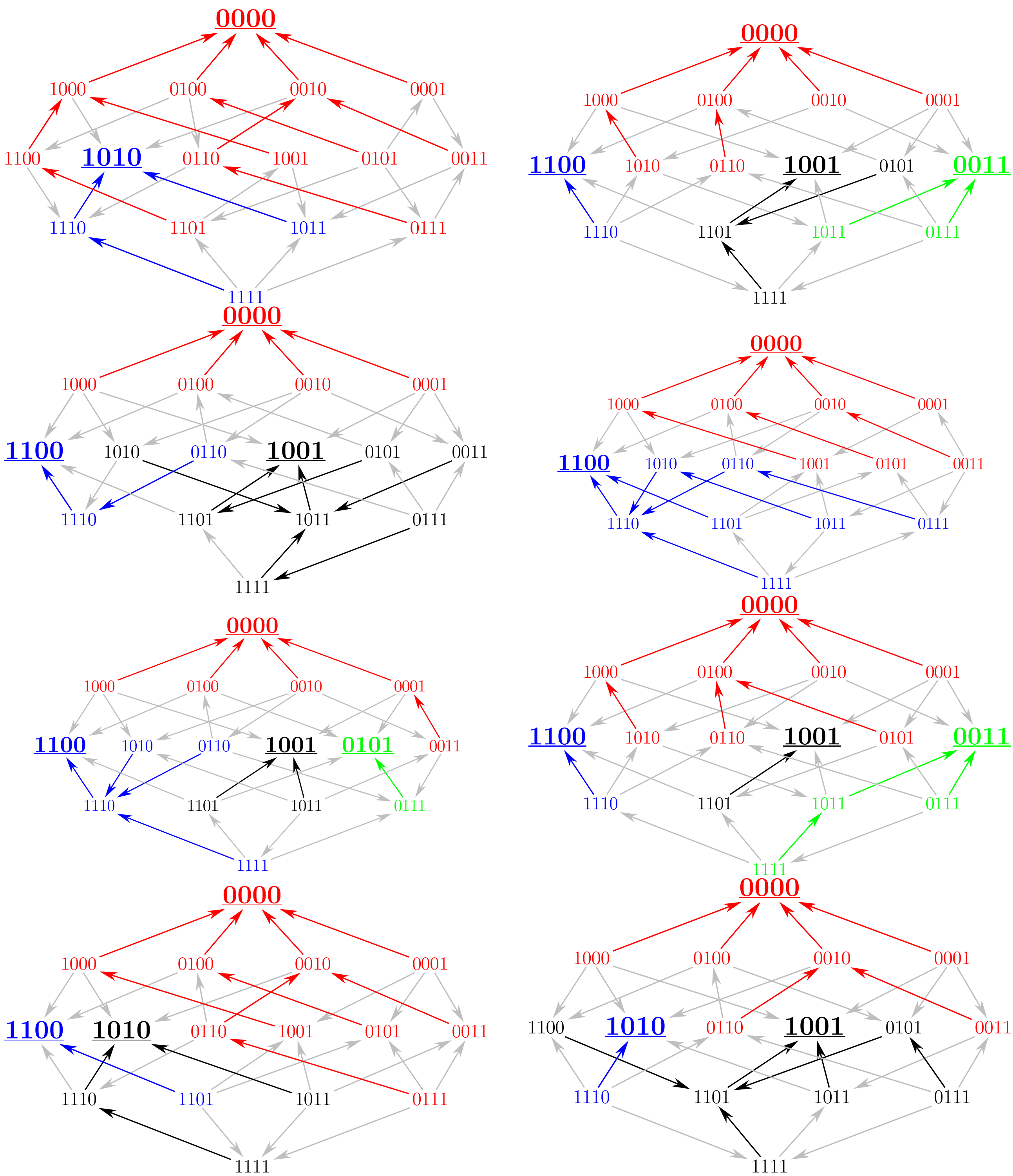


Figure A.2.: Subgraphs consisting of the mutations (from left to right in each row) arg, pyr, phe, oli, fwn, arg, oli, crn (first row), fwn, arg, pyr, crn, fwn, arg, pyr, oli (second row), fwn, arg, pyr, phe, fwn, leu, oli, crn (third row), fwn, leu, phe, oli and fwn, pyr, leu, crn (last row).

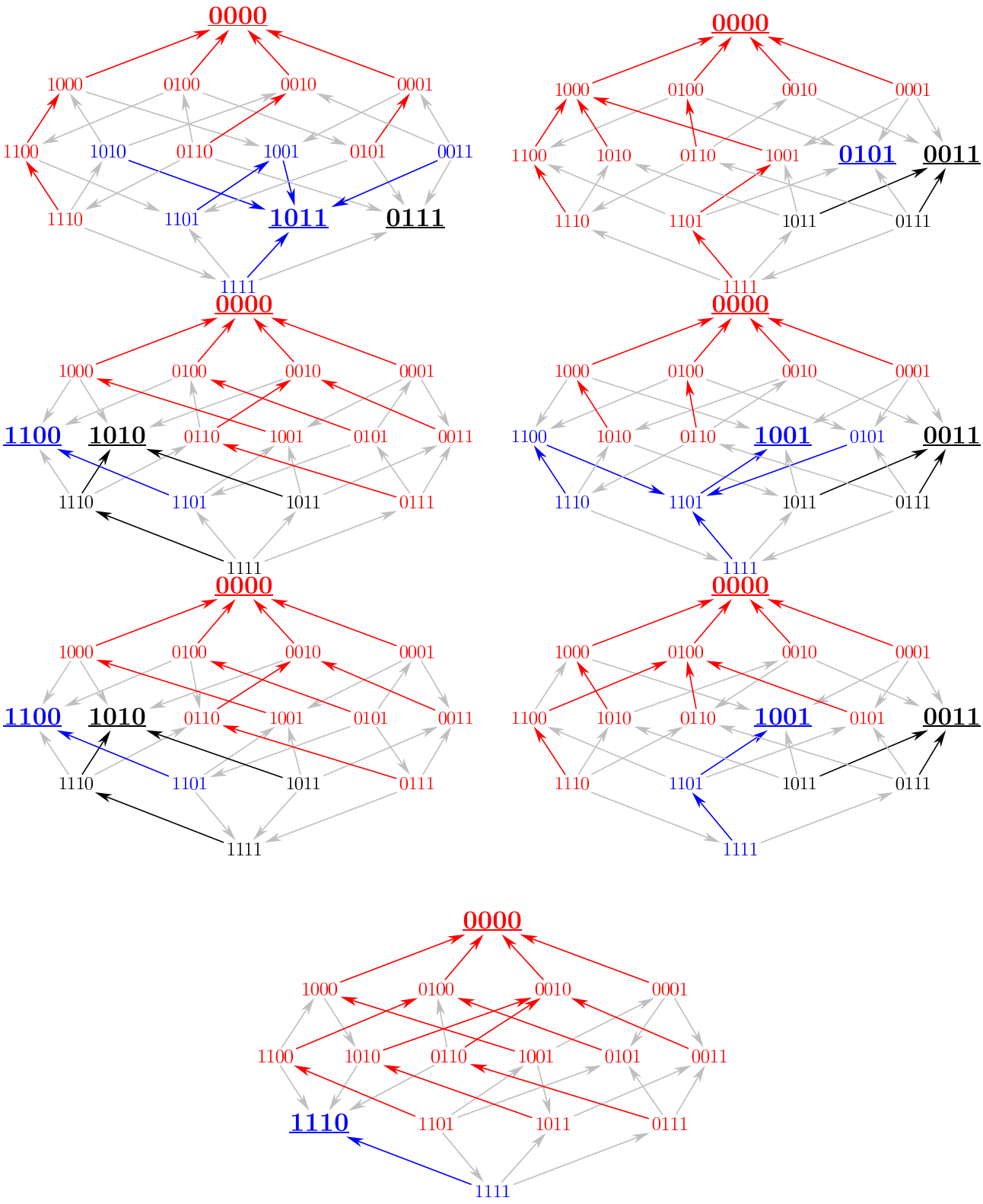


Figure A.3.: Subgraphs consisting of the mutations (from left to right in each row) arg, pyr, leu, phe, arg, pyr, oli, crn (first row) fwn, pyr, leu, oli, fwn, pyr, oli, crn (second row), fwn, pyr, phe, oli, pyr, leu, oli, crn (third row) and last but not least pyr, leu, phe, oli.

B. Record statistics

Computational steps omitted in studying the record statistics of random variables with a linear trend are presented.

B.1. Details on the expansion of $P_n(c)$

Details on the collapse of the sum of chains of integrals in eq (3.31) to the simple form leading to eq (3.32) will be given. The presentation here follows quite closely that given in the appendix III of [52].
The relation

$$F^l(x)\int^x \mathrm{d}y f^2(y) = l\int^x \mathrm{d}y f(y)F^{l-1}(y)\int^y \mathrm{d}x f^2(z) + \int^x \mathrm{d}f^2(z)F^l(z) \tag{B.1}$$

is valid for any probability density $f(\cdot)$ and the corresponding cumulative probability function $F(\cdot)$, as can be seen by applying integration by parts to the first term on the right hand side. With upper and lower bound of the support of $f(\cdot)$ denoted by s^+ and s_- respectively, one has the identities $F^l(s^+) = 1$ and $F^l(s_-) = 0$ for all l. Thus

$$\begin{aligned}\frac{1}{l!}\int \mathrm{d}f^2(x) &= \left[\frac{F^l(x)}{n!}\int^x \mathrm{d}x f^2(x)\right]\Bigg|_{x=s_-}^{x=s^+} \\ &= \frac{1}{l!}\int \mathrm{d}x f^2(x)F^l(x) + \frac{1}{(l-1)!}\int \mathrm{d}x f(x)F^{l-1}(x)\int^x \mathrm{d}y f^2(y).\end{aligned} \tag{B.2}$$

For $l \equiv n-2$, the first term on the right hand side of the equation above already matches with the first term on the right hand side of eq (3.31). Using eq (B.1) on $F^{l-1}(x)\int^x \mathrm{d}y f^2(y)$ inside the integral of the second summand on the right hand side above, one obtains

$$\begin{aligned}\int \mathrm{d}x f(x)F^{l-1}(x)\int^x \mathrm{d}y f^2(y) &= \int \mathrm{d}x f(x)\int^x \mathrm{d}y f^2(y)F^{l-1}(y) \\ &= (l-1)\int \mathrm{d}x f(x)\int^x \mathrm{d}y f(y)F^{n-2}(y)\int^y \mathrm{d}z f^2(z).\end{aligned} \tag{B.3}$$

Substituting the right hand side of the equation above for the integral in the second term of the

right hand side of eq (B.2), one sees that now the first *two* terms match eq (3.31). Since one can iterate this procedure with those terms that do not yet match, one has, with the identity $F^0(x) = 1$, finally expanded $\frac{1}{l!}\int \mathrm{d}x f^2(x)$ into the right hand side of eq (3.31) and thus arrives at the expression given in eq (3.32).

B.2. Examples of $\epsilon(c)$

The computations of the example expressions for $\epsilon(c)$ quoted in eqs (3.36) and (3.35) are provided.
Recall the definition given in eq (3.34),

$$\epsilon(c) = \int_c \mathrm{d}y f^{*2}(y) = \int_c \mathrm{d}y \int \mathrm{d}x f(x) f(y-x) \tag{B.4}$$

where the convolution $f^{*2}(\cdot)$ is an expression for the probability density of two RVs with with common density $f(\cdot)$ [46].
For $f(x) = \exp(-|x|)/2$, noting that since $c > 0$,only $y > 0$ need to be considered in computing the integral in eq (3.36), the two-fold convolution can be written as

$$\begin{aligned} f^{*2}(y) &= \frac{1}{4}\int^0 \mathrm{d}x e^x e^{x-y} + \frac{1}{4}\int_0^y \mathrm{d}x e^{-x} e^{x-y} + \frac{1}{4}\int_y e^{-x} e^{y-x} \\ &= \frac{1}{4}e^{-y} + \frac{1}{4}y e^{-y}. \end{aligned} \tag{B.5}$$

Thus

$$\int_c \mathrm{d}y f^{*2}(y) = \frac{1}{4}e^{-c} + \int_c \mathrm{d}y y e^{-y} = \left(\frac{1}{2} + \frac{c}{4}\right) e^{-c} \tag{B.6}$$

as stated in eq (3.36).
To obtain the expression eq (3.37) for the Lévy-stable distribution, first note that $f_\mu(x)$ is defined as inverse Fourier transformation of the corresponding $\hat{f}(k) = \exp(-|k|^\mu)$ and that thus the convolution $f^{*2}(x)$ is the inverse Fourier transform of $\hat{f}(k)^2$. Then

$$\begin{aligned} \epsilon(c) &= \frac{1}{2\pi}\int_c \mathrm{d}x \int \mathrm{d}k \exp\left(-ikx - 2|k|^\mu\right) \\ &= \frac{1}{2\pi}\int_c \mathrm{d}x \int \mathrm{d}k \exp\left(-ikx - \left|2^{1/\mu}k\right|^\mu\right). \end{aligned} \tag{B.7}$$

With the substitution $y \equiv 2^{1/\mu}k$ and subsequently $z \equiv 2^{-1/\mu}x$, this becomes

$$\begin{aligned} \epsilon(c) &= \frac{1}{2^{1+1/\mu}\pi}\int_c \mathrm{d}x \int \mathrm{d}y \exp\left(-iy2^{-1/\mu}x - |y|^\mu\right) \\ &= \frac{1}{2\pi}\int_{2^{1/\mu}c} \mathrm{d}z \int \mathrm{d}y \exp\left(-iyz - |y|^\mu\right). \end{aligned} \tag{B.8}$$

Since $y > c$ and c is assumed to be large for the expansion at hand, the inner integral (the inverse Fourier transformation) of the last line of the equation above can be approximated by the asymptotic expression for large y as given by [16] (see eq B.20 of Appendix B and references given there),

$$\int \mathrm{d}y \exp\left(-iyz - |y|^{\mu}\right) \approx 2z^{-\mu-1}\Gamma(1+\mu)\sin(\pi\mu/2). \tag{B.9}$$

With this expression, the remaining integration in z can be performed to finally yield the expression for $\epsilon(c)$ stated in eq (3.37).

B.3. Leading order coefficients

The record rate $p_n(c)$, the joint probability $p_{n,n-1}(c)$ of two consecutive record events and the record correlator $l_{n,n-1}(c)$ were studied by series expansion in c around $c = 0$. The leading order coefficients given by eqs (3.45), (3.47) and (3.48) will be computed here.
Starting with the explicit expression of the record rate in eq (3.14), one has

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}c}p_n(c) &= \frac{\mathrm{d}}{\mathrm{d}c}\left[\int \mathrm{d}x f(x) \int^{x+c} \mathrm{d}y f(y) \prod_{i=1}^{n-1} F(y+ic)\right] \\ &= \int \mathrm{d}x f(x) \sum_{i=1}^{n-1} i f(x+ic) \prod_{j=1, j\neq i}^{n-1} F(y+jc).\end{aligned} \tag{B.10}$$

Thus

$$\left.\frac{\mathrm{d}}{\mathrm{d}c}p_n(c)\right|_{c=0} = \frac{n(n-1)}{2}\int \mathrm{d}f^2(x)F^{n-2}(x), \tag{B.11}$$

as stated in eq (3.45). In the last step, the Gaussian sum formula was used.
For $p_{n,n-1}(c)$, with the explicit expression given by eq (3.42) one has

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}}p_{n,n-1}(c) =& \frac{\mathrm{d}}{\mathrm{d}c}\int \mathrm{d}x f(x) \int^{x+c} \mathrm{d}y f(y) \prod_{i=1}^{n-2} F(y+ic) \\ =& \int \mathrm{d}x f(x) f(x+c) \prod_{i=1}^{n-2} F(x+c+ic) \\ &+ \int \mathrm{d}x f(x) \int^{x+c} \mathrm{d}y f(y) \sum_{i=1}^{n-2} i f(x+ic) \prod_{j=1, j\neq i}^{n-2} F(y+jc)\end{aligned} \tag{B.12}$$

and therefor

$$\left.\frac{\mathrm{d}}{\mathrm{d}c}p_{n,n-1}(c)\right|_{c=0} = \int \mathrm{d}x f^2(x)F^{n-2}(x) + \frac{(n-1)(n-2)}{2}\int \mathrm{d}x f(x)\int^x \mathrm{d}y f^2(y)F^{n-3}(y) \tag{B.13}$$

as stated in eq (3.47) (again with use of the Gaussian sum formula).
With the definition of $l_{n,n-1}(c)$ given by eq (3.40) for $k = 1$, one has

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}c}l_{n,n-1}(c) =& \frac{\mathrm{d}}{\mathrm{d}c}\frac{p_{n,n-1}(c)}{p_n(c)p_{n-1}(c)} \\ =& \frac{1}{p_n(c)p_{n-1}(c)}\frac{\mathrm{d}}{\mathrm{d}c}p_{n,n-1} \\ & - \frac{p_{n,n-1}(c)}{p_n^2(c)p_{n-1}(c)}\frac{\mathrm{d}}{\mathrm{d}c}p_n(c) - \frac{p_{n,n-1}(c)}{p_n(c)p_{n-1}^2(c)}\frac{\mathrm{d}}{\mathrm{d}c}p_{n-1}(c).\end{aligned} \tag{B.14}$$

Setting $c \equiv 0$, one obtains

$$\begin{aligned}\left.\frac{\mathrm{d}}{\mathrm{d}c}l_{n,n-1}(c)\right|_{c=0} =& n(n-1)\left.\frac{\mathrm{d}}{\mathrm{d}c}p_{n,n-1}(c)\right|_{c=0} \\ & - n\left.\frac{\mathrm{d}}{\mathrm{d}c}p_n(c)\right|_{c=0} - (n-1)\left.\frac{\mathrm{d}}{\mathrm{d}c}p_{n-1}(c)\right|_{c=0}.\end{aligned} \tag{B.15}$$

Substituting the expressions given in eqs (B.11) and (B.13) finally yields the leading order coefficient of eq (3.49).

B.4. Records from processes with waiting times

A RV x_n is a record if it exceeds all *previous* RVs, i.e. if $x_n > \max_{j=1,\ldots,n-1}\{x_j\}$. Thus the family $\{x_n\}$ of RVs can be called a time series and the size n of this family can be interpreted as the number of 'time steps' the process has taken. If the process is such that the RVs are sampled at fixed intervals (each day at noon when $\{x_n\}$ is a time series of daily temperatures, for example), the index n of a given RV allows to locate it in continuous time (at least relative to the first entry x_1). It is, however, straightforward to extend these considerations to a process in which the entries of the times series are sampled at time points $t_1, \ldots, t_n$ separated by iid waiting times $t_n - t_{n-1} = \tau_n$, where the τ_i have a common probability density $\psi(\tau)$, called 'waiting time density (WTD)'. Of particular interest is the case where

$$\psi(\tau) \to \tau^{-\alpha} \tag{B.16}$$

for large τ. For this type of waiting time density, asymptotic expressions of the Laplace transform are readily available [43],

$$\tilde{\psi}(u) \to \begin{cases} 1 - \langle\tau\rangle u & \alpha > 1 \\ 1 - A|\Gamma(-\alpha)|u^{\alpha} & \alpha \leq 1 \end{cases} \tag{B.17}$$

The case $\alpha \leq 1$ is of special interest, because then the mean waiting time

$$\langle\tau\rangle = \int_0^{\infty} \mathrm{d}\tau\tau\psi(\tau) \tag{B.18}$$

diverges and the scaling behavior changes strongly (note that in the context of waiting times, the lower bound of the support of $\psi(\tau)$ is zero). The methods used to extend discrete time results like $p_n(c)$ to continuous time are in principle known [102] and use standard method of renewal theory (see e.g. [70] and references therein) but have to the authors knowledge not been applied to extract the results presented here yet.

General considerations Consider a quantity a_n that depends on discrete time, for example the expected value of the maximum attained during the random walk of length n [25,95] or the probability that the n^{th} entry in a time series of independent RVs is a record. Then for any real $z \in [0,1)$, the generating function (or z-transform) [140] of a_n is defined by

$$\hat{a}(z) = \sum_{n=0}^{\infty} z^n a_n. \tag{B.19}$$

Typically, at least for properties of a random walk, under this transform, the recursion relation for a_n, which usually is quite simple to set up but hard to solve, becomes an algebraic equation in z which is simpler to solve. One is however confronted with the problem of inverting this transform to obtain the desired object a_n, which is in general a difficult problem but will largely be circumvented in this presentation.
In order to embed the object a_n into real time, i.e. obtaining the corresponding object as a function of time t rather than the number n of steps taken, one has to compute the sum

$$a(t) = \sum_{n=0}^{\infty} a_n \times \text{Prob}[n \text{ steps taken in time } t]. \tag{B.20}$$

An expression for the probability of having exactly n events in time t is an elementary result from renewal theory (see e.g. [70]): The probability that the nth event has happened at time $t_1 < t$ is simply $\psi^{*n}(t_1)$, the n-fold convolution of the waiting time density. The probability of not having any further event in the remaining time $t-t_1$ is given by $\phi(t-t_1) \equiv 1-\int_0^{t-t_1} \mathrm{d}\tau\psi(\tau)$ and since the instant t_1 can be anywhere in $[0,t]$, one has

$$\text{Prob}[n \text{ steps taken in time } t] = \int_0^t \mathrm{d}t_1\psi^{*n}(t_1)\phi(t-t_1) = \left\{\psi^{*n} * \phi\right\}(t) \tag{B.21}$$

by the definition of the convolution.
Taking the Laplace transform of (B.20) and noting that the Laplace transform of convolutions is the product of the Laplace transformed functions, one gets

$$\int_0^\infty \mathrm{d}t a(t) e^{-ut} = \tilde{a}(u) = \sum_{n=0}^{\infty} a_n \tilde{\psi}(u)^n \tilde{\phi}(u) \tag{B.22}$$

where $\tilde{k}(u)$ denotes the Laplace transform of a function $k(t)$. The Laplace transform of $\phi(t)$ is given by

$$\tilde{\phi}(u) = \int_0^\infty \mathrm{d}t \phi(t) e^{-ut} = \int_0^\infty \mathrm{d}t \left(1 - \int_0^t \mathrm{d}\tau \psi(\tau)\right) e^{-ut} = \frac{1-\tilde{\psi}(u)}{u} \tag{B.23}$$

and hence the Laplace transform of the object $a(t)$ under consideration is given by

$$\tilde{a}(u) = \frac{1-\tilde{\psi}(u)}{u} \sum_{n=0}^{\infty} \tilde{\psi}^n(u) a_n. \tag{B.24}$$

It is quite simple to check that $\tilde{\psi}(u) \in [0,1]$ and hence the sum on the right hand side of (B.24) can be interpreted as the generating function $\hat{a}(z)$ with $z \equiv \tilde{\psi}(u)$,

$$\tilde{a}(u) = \frac{1-\tilde{\psi}(u)}{u} \hat{a}\left(\tilde{\psi}(u)\right). \tag{B.25}$$

Equation (B.25) is an exact relation between the generating function of a functional of a discrete time variable and the Laplace transform of the corresponding functional in continuous time. Hence whenever one has an expression for the generating function $\hat{a}(z)$ of an quantity that depends on a random walk (or more generally on any process in discrete time), by (B.25) one also has directly an expression for the Laplace transform of the continuous time quantity $a(t)$. Inverting this Laplace transform is, of course, a nontrivial task that is yet to be done, but as will be seen here, it usually suffices to insert a waiting time density with a sufficiently heavy tail to get an estimate of the long-time behavior of $a(t)$.

Eq (B.25) can be used to extract the scaling of record rates as follows: Consider a sequence $\{Y_i\}_{i\in\{1,\ldots,n\}}$ of RVs such that

$$Y_i = ci + X_i, \tag{B.26}$$

where $c \geq 0$ is a constant and the X_i are iid RVs with common density $f(x)$. Thus for $c \equiv 0$, one simply has a family of iid RVs, whereas for positive c the Y_i are independent RVs with a linear trend c, see [10,11,52], where the record statistics of this model is considered (see [151] for applications to weather data in the context of global warming). Note that one can equivalently consider independent RVs drawn from a family of probability densities $\{f_i\}$ which have the same shape but non-zero mean value ci for the i^{th} of these probability densities.
Here, we will consider the generalization of this model to continuous time, assuming that while the i^{th} RV Y_i, once drawn, will obey equation (B.26), the time t_i at which it is drawn is again random. One can thus introduce a waiting time density $\psi(t)$, from which the times *between* Y_i and Y_{i+1} are independently drawn. For definiteness again consider a WTD with heavy tails, $\psi(\tau) \to A\tau^{-1-\alpha}, \alpha > 0$ with asymptotic Laplace transform given by (B.17). Then the expected number of record events up to time t, denoted $N(t)$, can be computed from known results on the expected number of record events out of n RVs.

Linear drift model For strictly positive linear trend $c > 0$, it was shown in [10, 11] that the probability $p_n(c)$ that the n^{th} RV is a record tends for $n \gg 1$ to a finite value $p(c) > 0$, see also the discussion at the end of section 3.3. Thus the expected number of records, given that n RVs were drawn is for large n linear in n,

$$N_n = \sum_{i=1}^{n} p_n(c) \approx np(c) \text{ for } n \gg 1. \tag{B.27}$$

Actually, $N_n \approx (n - \tilde{n})p(c) + \sum_{i=1}^{\tilde{n}} p_i(c)$, where $\tilde{n}$ is chosen such that $p_n(c) \approx p(c)$, that is to say that the limit regime of large n is to a good approximation attained. The value of $\tilde{n}$ depends strongly on the underlying distribution $f(\cdot)$ of the iid part X_i of the Y_i and on the value of c, but for $n \gg \tilde{n}$, equation (B.27) provides a good estimate of the expected number of records. For the generating function $N(z)$ of N_n follows

$$\hat{N}(z) = \sum_{n=1}^{\infty} z^n N_n \approx \sum_{n=1}^{\infty} z^n np(c) = p(c) z \frac{\mathrm{d}}{\mathrm{d}z} \sum_{n=0}^{\infty} z^n = p(c) \frac{z}{(1-z)^2}. \tag{B.28}$$

Using (B.25) and inverting the resulting Laplace transform by Tauberian theorem [70] yields

$$N(t) \to \begin{cases} p(c)t/\langle\tau\rangle & \alpha > 1 \\ t^\alpha p(c)/\left[\kappa_\alpha \Gamma(\alpha+1)\right] & \alpha \leq 1 \end{cases} \tag{B.29}$$

iid RVs If the drift is switched off, i.e. c is set to 0, all RVs are independent and identically distributed. Thus the probability p_n that of the n RVs drawn up to a given point, the last one happens to be the largest (and hence a record) is simply given by

$$p_n = \frac{1}{n}, \tag{B.30}$$

which goes to zero as $n \to \infty$ in contrast the case for strictly positive c.
The expected number of records that occurred up to time n is given by the sum over these record rates,

$$N_n = \sum_{j=1}^{n} p_j = \sum_{j=1}^{n} \frac{1}{j} \approx \ln(n) + \gamma + \mathcal{O}(1/n), \tag{B.31}$$

where $\gamma = 0.577216\ldots$ is the Euler-Mascheroni constant (see e.g. [86]). To extend this to continuous time, first one has to compute the generating function $\hat{N}(z)$, before (B.25) can be applied. By use of the known series expansion of the logarithm $\ln(1-z) = -\sum_{n=1}^{\infty} z^n/n$, one obtains

$$\begin{aligned} \hat{N}(z) &= \textstyle\sum_{n=1}^{\infty} z^n N_n = \sum_{n=1}^{\infty} z^n \sum_{j=1}^{n} \frac{1}{j} = \sum_{n=1}^{\infty} z \left(z^{n-1} \sum_{j=1}^{n-1} \frac{1}{j} + \frac{z^{n-1}}{n} \right) \\ &= z \textstyle\sum_{n=1}^{\infty} z^{n-1} \sum_{j=1}^{n-1} \frac{1}{j} + \sum_{n=1}^{\infty} \frac{z^n}{n} = z\hat{N}(z) - \ln(1-z) \end{aligned} \tag{B.32}$$

where in the last step the fact that when moving the index from $n-1$ to n in the first sum, the inner sum of the $n = 0$-term is empty was used. Bringing $z\hat{N}(z)$ to the other side and dividing

by $1/(z-1)$ yields $\hat{N}(z) = \ln(1-z)/(z-1)$ and thus by (B.25)

$$\tilde{N}(u) = -\frac{\ln\left(1-\tilde{\psi}(u)\right)}{u}. \tag{B.33}$$

Substituting again the two different forms for $\tilde{\psi}(u)$ from (B.17) followed by some elementary manipulations yields

$$\tilde{N}(u) = \begin{cases} [\ln(1/u) - \ln(\langle\tau\rangle)]u^{-1} & \alpha > 1 \\ [\alpha \ln(1/u) - \ln(\kappa_\alpha)]\, u^{-1} & \alpha \leq 1 \end{cases} \tag{B.34}$$

which can again be asymptotically inverted using the Tauberian theorems to yield

$$N(t) \to \begin{cases} \ln(t/\langle\tau\rangle) & \alpha > 1 \\ [\alpha \ln(t) - \ln(\kappa_\alpha)] & \alpha \leq 1 \end{cases} \tag{B.35}$$

Note that in comparison to (B.31), the case of $\alpha > 1$ misses the Euler-Mascheroni constant γ. This constant can be recovered by use of the inverse Laplace transform $\ln(t) + \gamma$ for the Laplace transform $-\ln(u)/u$ from [42]. Using this inverse rather than the asymptotics provided by the Tauberian theorem recovers also the sub-leading order of $N(t)$, matching (B.31).
Even though by use of the method presented here it was straightforward to extract asymptotic expressions for the expected number of records from random variables drawn at random intervals, apparently only special cases of the waiting time density have been considered so far (see e.g. [114], where drawing of random variables was modeled as a Poisson process). Specifically the result for diverging mean waiting times between events seems to be new.

C. Subgraphs

The method of subgraphs used subsequently to compare the findings on model studies to the empirical data ist presented here. The validity of this method is discussed.

C.1. Introduction

Any empirical fitness landscape necessarily considers a fixed number L (typically $L \sim \mathcal{O}(10)$) of mutations, while biological evolution acts genome wide and the number of possible mutations available to it is far beyond anything that can be considered in a laboratory. In order to extrapolate to large L, model studies are used. To check which of these models (if any) show behavior similar to that of empirical FLs, subgraphs of the full data set as introduced in section 1.5.3 can be used. As was mentioned there, any two subgraphs of size $m < L$ share a certain number of fitness values, which implies that they cannot be treated as independent data sets. However, as will be shown in this section, as far as the expected number of accessible paths and the average size of the global optimum's basin of attraction under greedy dynamics are concerned, correlations between subgraphs are sufficiently weak to allow the use of subgraphs instead of truly independent data sets. The general considerations about subgraphs presented in this section as well as the results on correlations of the expected number of accessible paths $\langle n_L \rangle$ for the 'House of Cards' and 'Rough Mt. Fuji' models presented in the next section were originally obtained in the Bachelor thesis of Tim Dressler [37] while the results on the LK model as well as those on the correlations of the basin of attraction presented in section C.3 have not been published so far.

C.1.1. Subgraph distance

To create a subgraph of size m, choose m of the $L > m$ sites of the whole sequence. When the state of the $L - m$ sites not present in the subgraphs is fixed, for example by the condition that the global maximum must be present in all subgraphs, there are a total of $\binom{L}{m}$ subgraphs possible (without such a condition there are $\binom{L}{m}$ different subgraphs for each of the 2^{L-m} states of the remaining sites, see fig C.1).

The SG then consists of the 2^m configurations obtained by creating all possible states of the sub-sequence consisting of those m sites chosen to be in the SG while keeping those $L - m$ states *not* involved in the SG in the state they have on the GM. Together with the state of the GM on the remaining sites, these m sites completely define the SG. Hence a SG g can be denoted as $g = \{j_1, \ldots, j_m\}$, where it is understood that the sequence of the GM is yet to be specified. In simulations presented in this section, the GM was always at $000\ldots0$.
The distance between two SGs $g_1 = \{j_1, \ldots, j_m\}$ and $g_2 = \{j'_1, \ldots, j'_m\}$ of equal size m is given by counting the number of sites which are present in one but not in the other subgraph. For

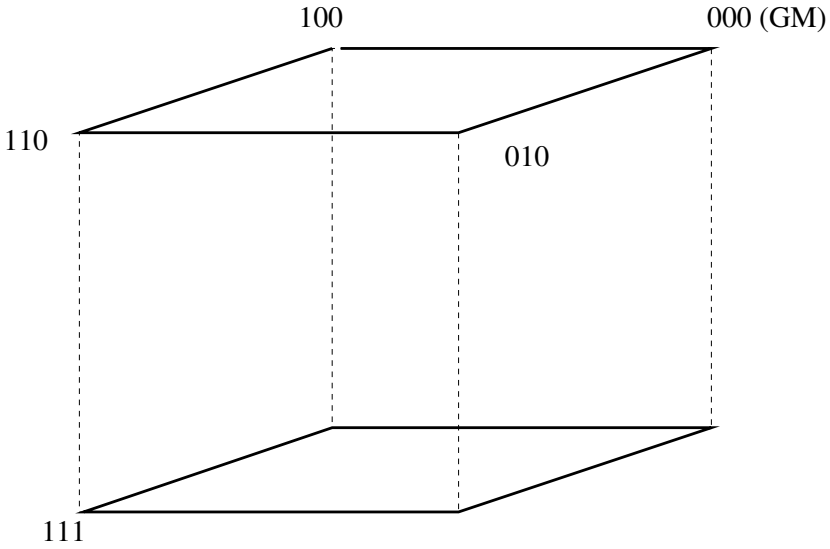


Figure C.1.: In the $L = 3$ Boolean hypercube shown here, the GM is at position 000. For the subgraphs consisting of the first and second site (the top and bottom face of the cube, marked by darker lines), the condition that the GM be included in the subgraphs implies that the sites not present in the SG must be set to 0. This singles out the top face as the SG to be considered here.

example the SGs made up of mutations $\{1, 2, 3\}$ and $\{1, 2, 4\}$ have SG-distance 1.
Clearly the maximum SG-distance is

$$d_{SG}(g_1, g_2) \leq \begin{cases} m & \text{if } m \leq \frac{L}{2} \\ L - m & \text{if } m > \frac{L}{2}. \end{cases} \tag{C.1}$$

The upper bound on SG-distance arises because once the two SGs are completely disjoint sets (i.e. they do not share any sites), the distance between them cannot be increased any further. This situation can only arise if $m \leq L/2$.
There are

$$\mathcal{N}_{SG}(L, m, \delta) = \frac{1}{2}\binom{L}{m}\binom{m}{\delta}\binom{L-m}{\delta}, \tag{C.2}$$

pairs of SGs at SG-distance δ since for a given sequence length L there are $\binom{L}{m}$ SGs of size m, and any δ of these m sites can be replaced by any of the $L - m$ remaining sites, where the factor $1/2$ compensates for the over-counting that occurs because each pair of SGs at a given distance is otherwise counted twice. For $m \sim L/2$, this number can be quite large, allowing for reasonable statistics even on one single FL.

C.1.2. Correlator

To quantify correlations of a random variable x (e.g. the number of accessible paths n_m) across SGs as function of SG-distance δ, one can use the correlator defined as follows: For a given FL, for each pair of SGs $\{g_1, g_2\}$ at given distance $d_{sg}(g_1, g_2) = \delta$, first evaluated x_1 and x_2 on g_1 and g_2 respectively, then form their product $x_1 x_2$. If this is done for all $\mathcal{N}_{SG}(L, m, \delta)$

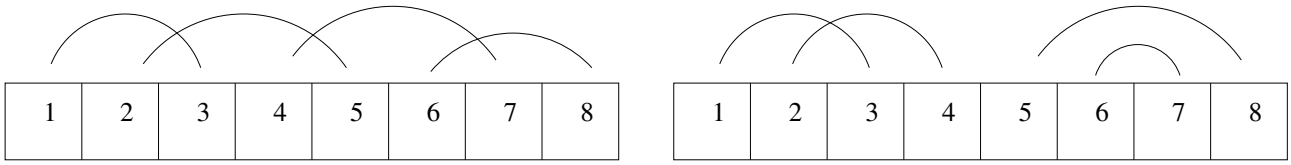


Figure C.2.: The left panel shows a possible way in which spin glass interactions might induce correlations between subgraphs, even when these subgraphs are at maximum distance m. The right panel shows a very special instance of a spin glass where the interactions are such that a pair of subgraphs at maximum distance m can be chosen such that there are no correlations. If these interactions are uniformly distributed across the sequence L, such realizations are clearly a minority. On the other hand, if interactions are only present between neighboring sites, such a scenario is much more likely.

SGs at distance δ, one can form the average $\langle x_1 \rangle_{SG}$, $\langle x_2 \rangle_{SG}$ and $\langle x_1 x_2 \rangle_{SG}$, where the subscript SG on $\langle \cdot \rangle_{SG}$ indicates the arithmetic mean with respect to the number $\mathcal{N}_{SG}(L, m, \delta)$ rather than an arithmetic mean with respect to a large number of independent (numerically obtained) realizations as used this far in this thesis. Then, in analogy to the covariance, the subgraphs correlator of the RV x can be defined as

$$C_{L,m}(\delta) = \langle x_1 x_2 \rangle_{SG} - \langle x_1 \rangle_{SG} \langle x_2 \rangle_{SG} . \tag{C.3}$$

The correlator for SG-distance $\delta = 0$ simply measures the standard deviation $\left\langle x^2 \right\rangle_{SG} - \langle x \rangle_{SG}^2$ of the RV x. To study the *typical* behavior of the correlator, it was averaged over a large number (typically 10^4) of independent landscapes. This average is denoted by regular brackets $\langle \cdot \rangle$.

C.1.3. General behavior

The properties to be considered here are defined globally in the sense that they concern a large part of the FL (if a given state is a local optimum, to give an example of a *locally* defined object, can be decided by looking at the L immediate neighbors). *A priori* one expects the correlations across subgraphs to be more important for global objects than for local ones, making them more interesting.

Clearly, if the RVs x_1 and x_2 are uncorrelated, the correlator $\left\langle C_{L,m}(\delta) \right\rangle$ averaged over many independent landscapes must vanish. This is for example the case when the FL is drawn from the HoC or RMF model and the SGs have maximum distance $\delta = m$. For FLs drawn from the LK or other spin glass like models, this is not necessarily the case, as there may be interdependencies due to interactions across SGs, see e.g. fig C.2. This figure also shows that among those spin glass like models, those with local (e.g. next-neighbor) interactions arguably have lower correlations than those models where the interaction partners are scattered across the whole sequence, see the discussion on the choice of neighborhood in the LK model in section 2.1.2. Thus the results on the correlations within the LK model with random neighborhoods can heuristically be argued to present an upper bound to the correlations of the adjacent neighborhood LK model.

On a totally smooth landscape such as the RMF model in the $c \to \infty$ limit, two given subgraphs

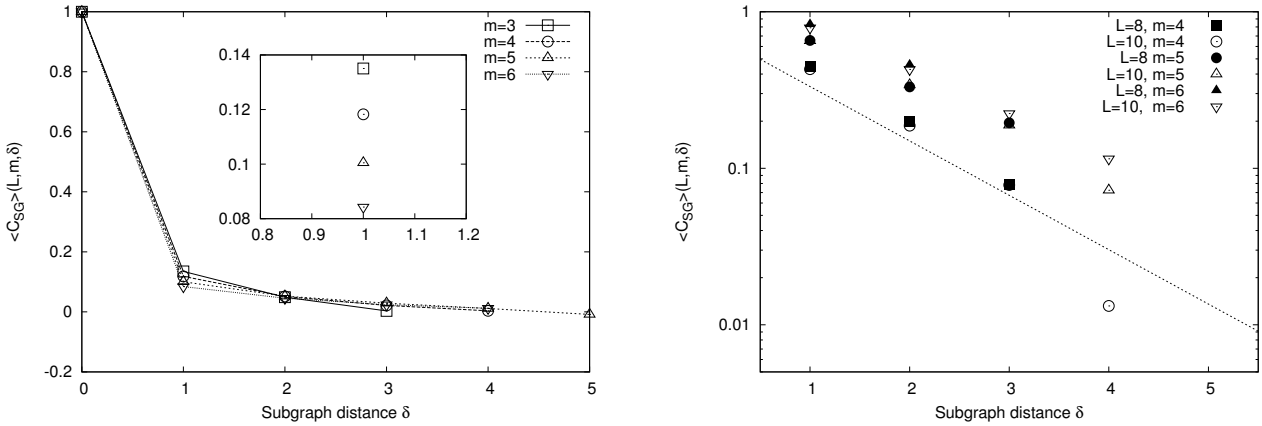


Figure C.3.: The average correlator $\langle C_{SG}(L,m,\delta)\rangle$ of n_m for the HoC model shown as function of subgraph distance δ for different values of m and $L = 10$ (and on the right also for $L = 8$). The correlator is normalized by the value at $\delta = 0$ to allow for comparison of different values of m. The inset in the left figure shows a detail of the main plot. The right figure is a semi-log-plot to show exponential decay indicated by the straight line. Note that the curves are independent of L. If $\delta = m$, the average correlator vanishes, thus the corresponding points are zero up to numerical precision.

will be identical and thus highly correlated. The fluctuations on the other hand (the correlator at $\delta = 0$) will vanish. Thus one expects that the correlator will become trivially independent of δ as the landscape becomes smoother. To avoid this trivial regime, the simulations presented in the next two sections for the correlator $C_{L,m}(\delta)$ defined in eq (C.3) will only treat the regime of high ruggedness (small c or large K in the RMF or LK model respectively).

C.2. Number of accessible paths

This section presents numerical results on the correlations of the number of accessible paths across subgraphs. For the numerical studies, the correlator $C_{L,m}(\delta)$ was, after averaging over all possible subgraphs, also averaged over a large number of independent FLs. Most results presented here were originally obtained in the Bachelor's thesis of Tim Dressler [37].

C.2.1. HoC and RMF models

As was mentioned above, in these two models the only interdependence between SGs arises from shared fitness values. Hence, if $m \leq L/2$, at SG-distance $d_{SG} = m$ the averaged correlator $\langle C_{SG}(L,m,\delta)\rangle$ vanishes for all values of c. This can be seen on the left of fig C.3, where the averaged correlations as function of subgraphs distance are shown for the HoC model. Clearly (under the condition that $m < L$ of course), for these models the total sequence length L does not play any role, see fig C.3. This figure shows that for $\delta \geq 1$, $\langle C_{SG}(L,m,\delta)\rangle$ decays exponentially, see right panel of fig C.3. Extrapolating this exponential decay to $\delta = 0$ does not yield the observed fluctuations, which means that relative to the fluctuations at $\delta = 0$, correlations drop fast going from $\delta = 0$ to $\delta = 1$. This initial drop increases as the subgraph

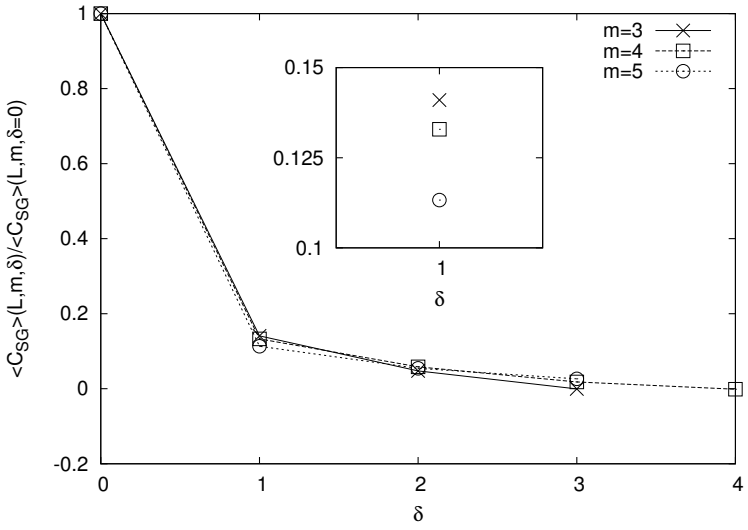


Figure C.4.: The average correlator of n_m for the RMF model with $c = 0.1$ on landscapes with $L = 8$. One observes the same behavior as for the HoC case: Initially a super-exponentially sharp drop followed by an exponential decay in δ. As shown in the inset, the initial drop in the correlator becomes steeper as m grows.

size m grows, see inset of the left panel of fig C.3.
For the RMF model, the situation is slightly more complicated, as can be seen in fig C.4. Note the similarity to the behavior observed in the HoC model. For moderate values of c, the behavior of the curves is essentially the same as in the HoC case, while for larger c, it changes slightly, see fig C.5.

Fig C.5 shows the effect of increasing the underlying drift velocity c. For sufficiently large c, the correlations do not drop monotonically with SG-size m but rather seem to increase. However, as m is increased further, eventually the correlations drop again in m, which leads to suppose that for any value of c, correlations between SGs eventually decay.

Thus under both HoC and RMF model, the correlations relative to the standard deviation (at $\delta = 0$) are found to be very small even at $\delta = 1$ and decline exponentially from there on.

C.2.2. The LK model

The LK model as introduced in section 2.1.2 explicitly takes interactions between mutations into account and models them in the spirit of a diluted spin glass. As stated in the introduction of this section, here only the version with interactions ranging across the entire sequence will be discussed.
As for the RMF and HoC models discussed above, one sees an exponential decline with SG-distance. However the $\delta = 0$ point fits in with the points for $\delta \geq 1$ without the jump observed for the models discussed above, see fig C.6.

Thus also in this case, correlations decay exponentially as function of SG-distance. Even though the marked jump going from $\delta = 0$ to $\delta = 1$ is not present here, correlations decay quite strongly.

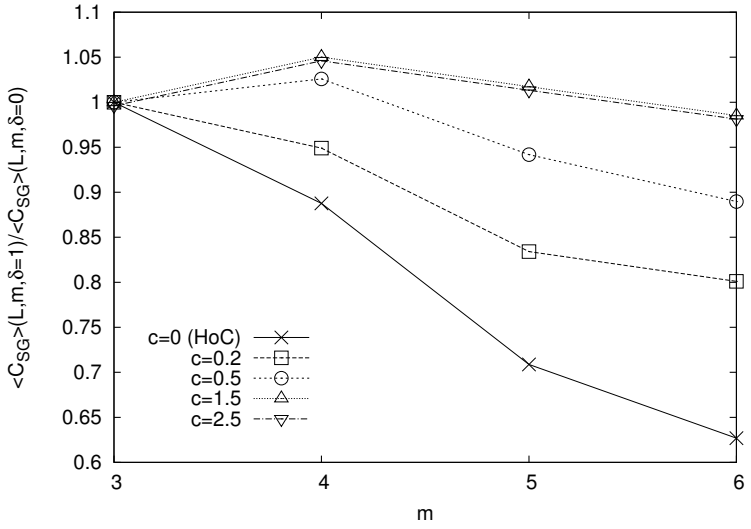


Figure C.5.: Simulations showing the c-behavior of the value of $\langle C_{SG}\rangle\,(L, m, \delta = 1)$ relative to the fluctuations ($\delta = 0$) normalized by $\langle C\rangle\,(L, m = 3, \delta = 1)$ for n_m.

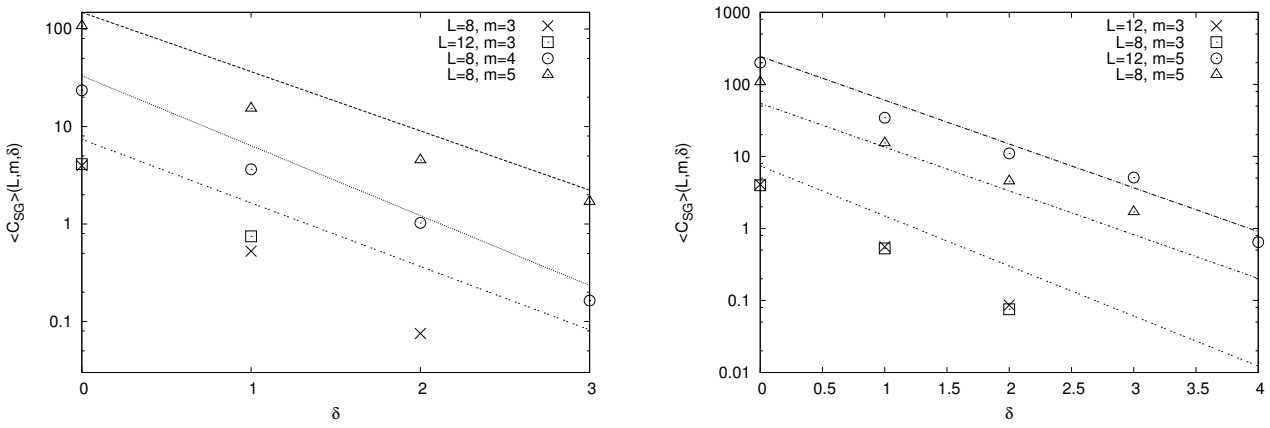


Figure C.6.: Correlations of n_m as derived from the LK model for different values of m as function of SG-distance δ. The left panel shows simulations for fixed value of $K = 4$, the right for fixed $K/L = 1/2$. In this semi logarithmic plot, the decay follows approximately the exponential indicated by the lines with slopes (from top to bottom) -1.4, -1.65 and -1.5 in the left and -1.4, -1.4 and -1.6 in the right panel.

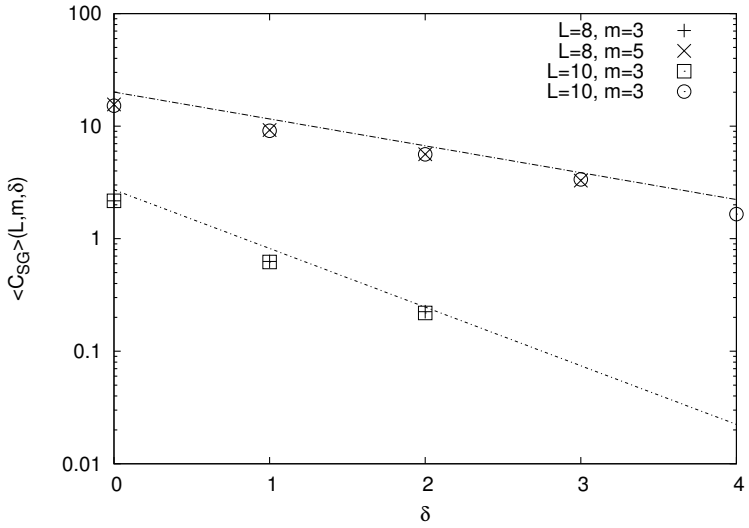


Figure C.7.: Correlator for the basin of attraction as function of subgraph distance δ for the HoC model and different values of L and m. Clearly, the curves do not depend on L and show an exponential decrease. The linear decline in the semi-logarithmic plot with slopes of -0.55 (top) and -1.2 (bottom) indicates exponential decay.

C.3. Basin of attraction of the global maximum

If one considers the expected size of the basin of attraction of the global optimum instead of the mean number of accessible paths, the correlator defined in eq (C.3) can be used in the same way as above to determine the degree to which subgraphs are correlated with respect to the mean size of the basin of attraction. Again numerical simulations will be used for the three models also treated above.

C.3.1. HoC and RMF

For the HoC model, numerical simulations are presented in fig C.7. Even though the decline is not as steep as for the number of accessible paths, nonetheless it seems exponential in subgraph distance δ. Furthermore, it is clearly seen that the correlations do not depend on sequence length but only on subgraph size.

For the RMF model, i.e. for $c > 0$, the exponential decay is still present, as can be seen in fig C.8. There, numerical simulations for two different values of c are presented. Just like for the HoC model, the exponential decay in the correlations is not as marked as for the expected number of accessible paths, yet shows in the simulations.

One observes that for increasing value of c, the slope in the semi-log plots in fig C.8, in particular in the right panel, decreases with subgraph size m. This slope, however, does not seem do depend on c but only on subgraph size m. Whether this is just a coincidence or can be explained remains an open question up to this point.

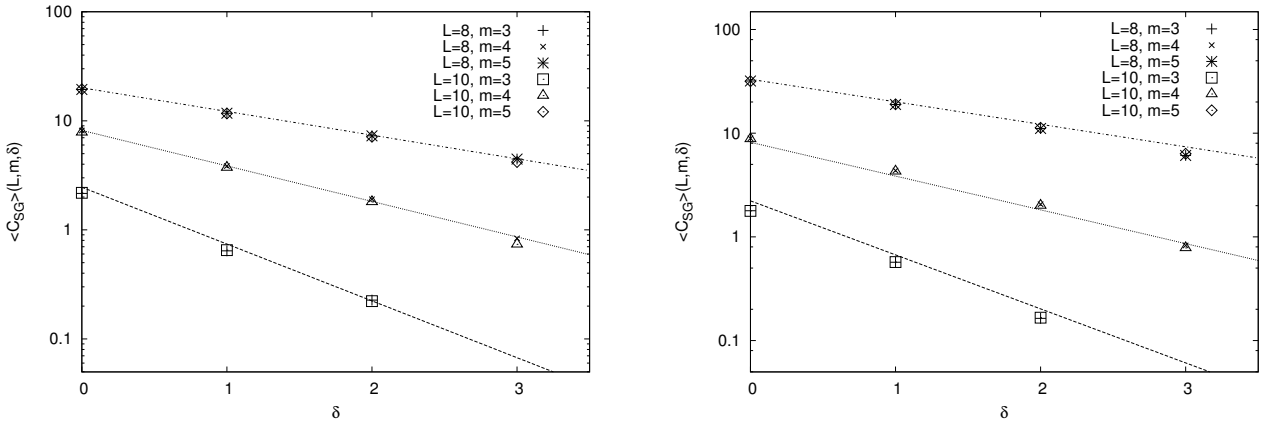


Figure C.8.: Correlations between the sizes of the basin of attraction as function of subgraph distance for the RMF model with $c = 0.1$ (left panel) and $c = 0.5$ (right panel). The lines are shown to emphasize the linear decline in the semi-log plot. They have slopes of $-0.5, -0.75$ and -1.2 (from top to bottom) in both panels.

C.3.2. LK model

In the LK model, exponential decay of correlations was observed both for fixed number K of interacting sites and fixed ration K/L. Examples are shown in fig C.9.

C.4. Conclusions

The simulations presented in this chapter show that the correlations of both the number of accessible paths n_m and the size of the basin of attraction of the global maximum b_m across subgraphs decay *at least* exponentially with subgraph distance δ. In models where correlations between different subgraphs are only induced by shared states (HoC and RMF) this is not surprising, while for the LK model this was not *a priori* obvious. It was also not obvious that in the LK model, correlations should only depend on L as weakly as observed in fig C.9, while in the other models, this was expected.
The exponential decay observed throughout all models shows that analyzing the m-behavior of either object (n_m and b_m) on a given empirical fitness landscape of fixed sequence length L allows to estimate the L-behavior of the corresponding object on the FL. This will prove to be an important tool in comparing empirical fitness landscapes to theoretical results in chapters 4 and 5.

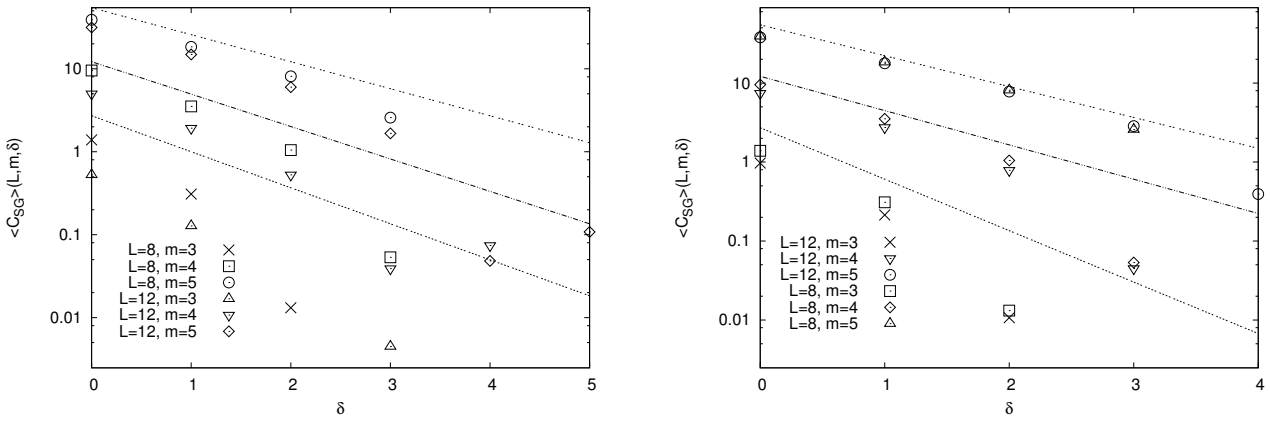


Figure C.9.: Correlations of the basins of attraction for the LK model. The left panel shows simulations for a fixed value of $K = 4$ while on the right, $L/K = 1/2$ was kept fixed. Note that in the right panel, the correlations seem to be independent of L, while on the left, they at least do not seem to depend too strongly on L. The lines from top to bottom have slopes of $-0.75, -0.9$ and -1 on the left and $-0.9, -1$ and -1.5 on the right.

D. Notes on the numerical procedures

Details on the numerical procedures employed throughout this thesis are presented. This includes the resampling procedure by which the error bars on empirical data points were obtained, search routine employed to count the number of accessible paths on a given FL and the routine to compute the HTI.

D.1. The resampling procedure

Whenever error bars to empirical data points were presented in this thesis, they were obtained using the resampling method [31]. This method was originally intended to test the robustness of certain features of the fitness landscapes w.r.t. the measurement error. Whenever a data set has m replicate measurements $f_1[n], f_2[n], \dots f_m[n]$ of the fitness for each state[1] n, one has a mean fitness

$$\mu[n] = \frac{1}{m} \sum_{j=1}^{m} f_j[n] \tag{D.1}$$

and a standard deviation

$$\sigma[n] = \sqrt{\frac{1}{m} \sum_{j=1}^{m} \left(f_j[n] - \mu[n]\right)^2} \tag{D.2}$$

fr each state.
Then, assuming that the measurement error is Gaussian, one can created (or 'resample') a large number S of artificial data sets according to

$$f[n] = \mu[n] + \eta_n, \qquad \eta_n \sim \mathcal{N}(0, \sigma[n]) \tag{D.3}$$

with $\mathcal{N}(0, x)$ the normal distribution with standard deviation x. On each of these S resampled FLs, one can the compute the property of interest, e.g. count the number of accessible paths and record mean and standard deviation of this object. The standard deviation then gives an estimate for the error bars.

It is not *a priori* obvious that because the error on each fitness value is small compared to the mean fitness values, all topological features should be robust against this errors. This can occur for example if in some states, adjacent fitness (or resistance levels) are close enough for the measurement errors to become important when determining the number of accessible paths while all states (except for the GM) have at least on neighbor with a fitness so great that under

[1] labeling these states by $n = 1$ through 2^L

resampling, none of them became a local maximum (and thus the entire FL was in the BoA of the GM).

D.2. Details on the search routines used

A large part of the numerical work presented in this thesis consisted of searching entire state spaces, e.g. the L-dimensional Boolean hypercube $\{-0,1\}^L$, for the property of interest. Since one aim was to reach high dimension L, depth-first routines were used throughout this thesis. Such routines follow for example one particular path[2] until they determine whether this path leads to the global maximum or is broken because it encounters a fitness valley. Since many of the $L!$ paths share edges of the graph, such a procedure considers the same edge several times. This redundancy however was necessary because the alternative, i.e. considering all possible paths at the same time, would have been too demanding on memory.
For the accessible paths, the routine used can be sketched as follows:

```
find-path(int n, double fit[], int GM)
```

- `start in state n`
- `if(n==GM) return('found the global maximum')`
- `else:  for(all neighbors i=1 through L): if(fit[i]>fit[n])`
  - `find-path(i, fit[])`
- `end`

If this routine is called for some starting state s, it finds all paths in connecting s to GM. If in the third line only such states are considered that are *closer* to the GM that the starting state n, only *shortest* paths are considered.
To compute $\pi_L(0)$, the routine aborts the search as soon as the first state is found because then the realization considered contains at least one state.

D.3. Computing the HTI for a given data set

To compute the HTI, assume that a data set is given as entries $x_1, x_2, \ldots x_N$. Then for given drift velocity c (chosen to match the mean spacing between entries of the data set) and some value $n < N$ the HTI can be computed as follows

[2]For definiteness, only the example of counting the number of accessible paths on a given FL is discussed here. The procedures for the other objects considered here are straightforward adaptations.

```
for(S iterations)
```

- `for(j<n)`
  - `signal-n=0`
  - `signal-(n-1)=0`
  - `pick one entry x[j] of the data set`
  - `y[j]=x[j]+cj`
  - `if(j==(n-1) and y[j] record):signal-(n-1)=1`
  - `if(j==(n) and y[j] record):signal-n=1`
- `p-(n, n-1)+=signal-n*signal-(n-1)/S`
- `p-(n-1)+=signal-(n-1)/S`
- `p-n+=signal-n/S`

At the end of this procedure, one has an estimate for the probabilities $\hat{p}_n(c)$, $\hat{p}_{n-1}(c)$ and $\hat{p}_{n,n-1}(c)$ on that data set and can compute the fraction

$$\hat{l}_{n,n-1}(c) = \frac{\hat{p}_{n,n-1}(c)}{\hat{p}_n(c)\hat{p}_{n-1}(c)}. \tag{D.4}$$

Bibliography

[1] *Guinness World Records 2011*, Guinness World Records, 14th ed., 2011.

[2] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions*, Dover, New York, 1965.

[3] 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 endeopeptidase and thermolysin*, Biopolymers, (2000), pp. 64–79.

[4] K. Aktories, U. Förstermann, F. B. Hofman, and K. von Starke, *Allgemeine und Spezielle Pharmakologie und Toxikologie*, Elsevier, 2004.

[5] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Molecular Biology of The Cell*, Garland Science, 2008.

[6] L. Altenberg, *NK Fitness Landscapes*, in The Handbook of Evolutionary Computation, T. Back, D. Fogel, and Z. Michalewicz, eds., Oxford University Press, 1997, ch. B2.7.2.

[7] B. Arnold, H. N. Nagaraja, and N. Balakrishnan, *Records*, John Wiley & Sons, 1998.

[8] M. Bachelier, *Théorie de la spéculation*, Annales Scientifiques de l'E.N.S., 17 (1900), pp. 21–86.

[9] P. Bak and K. Sneppen, *Punctuated equilibrium and criticality in a simple model of evolution*, Phys. Rev. Lett., 71 (1993), pp. 4083–4086.

[10] R. Ballerini and S. Resnick, *Records from improving populations*, Journal of Applied Probability, 22 (1985), pp. 487–502.

[11] ——, *Records in the presence of a linear trend*, Advances in Applied Probability, 19 (1987), pp. 801–828.

[12] A.-L. Barabasi and R. Albert, *Emergence of scaling in random networks*, Science, 286 (1999), pp. 509–512.

[13] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics*, Academic Press, London, 1982.

[14] E. Bertin and G. Gyögyi, *Renormalization flow in extreme value statistics*, J. Stat. Mech.: Theor. Exp., P08022 (2010).

[15] H. A. Bethe, *Statistical theory of superlattices*, Proc. Roy. Soc. (London) A, 150 (1935), pp. 552–575.

[16] J.-P. Bouchaud and A. Georges, *Anomalous diffusion in disordered media: Statistical mechanisms, models and physical applications*, Physics Reports, 195 (1990), pp. 127–293.

[17] A. Bovier, *Statistical Mechanics of Disordered Systems: A Mathematical Perspective*, Cambridge University Press, Cambridge, 2006.

[18] S. Carmi, P. Krapivsky, and D. ben Avram, *Partition of networks into basins of attraction*, Phys. Rev. E, 78, 066111 (2008).

[19] M. Carneiro and D. L. Harlt, *Adaptive landscapes and protein evolution*, Proc. Natl. Acad. Sci., 107 (2009), pp. 1747–1751.

[20] J. V. Chamary, J. L. Parmley, and L. D. Hurst, *Hearing silence: non-neutral evolution at synonymous sites in mammals*, Nature Reviews Genetics, 7 (2006), pp. 98–108.

[21] H.-H. Chou, H.-C. Chiu, N. F. Delaney, D. Segrè, and C. J. Marx, *Diminishing returns epistasis among beneficial mutations decelerates adaptation*, Science, 332 (2011), pp. 1190–1192.

[22] K. Christensen and N. R. Moloney, *Complexity and Criticality*, Imperial College Press, London, 2005.

[23] A. Clauset, C. R. Shalizi, and M. E. J. Newman, *Power-law distributions in empirical data*, SIAM Review, 51 (2009), pp. 661–703.

[24] J. Clune, D. Misevic, C. Ofria, R. E. Lenski, S. F. Elena, and R. Sanjuan, *Natural selection fails to optimize mutation rates for long-term adaptation on rugged fitness landscapes*, PLoS Comput. Biol., 4:e1000187 (2008).

[25] A. Comtet and S. N. Majumdar, *Precise asymptotics for a random walker's maximum*, J. Stat. Mech. Theor.Exp., P06013 (2005).

[26] A. Comtet, S. N. Majumdar, and S. Ouvry, *Integer partitions and exclusion statistics*, J. Phys. A: Math. Theor., 40 (2007), pp. 11255–11269.

[27] A. Comtet, S. N. Majumdar, S. Ouvry, and S. Sabhapandit, *Integer partitions and exclusion statistics: limit shapes and the largest parts of young diagrams*, J. Stat. Mech: Theor. Exp., P10001 (2007).

[28] C. Darwin, *On the Origin of Species By Means of Natural Selection*, John Murray, London, 1st ed., 1859.

[29] L. de Haan and A. Ferreira, *Extreme Value Theory: An Introduction*, Springer, 2006.

[30] J. de Visser, R. F. Hoekstra, and H. van den Ende, *Test of interaction between genetic markers that affect fitness in aspergillus niger*, Evolution, 51 (1997), pp. 1499–1505.

[31] J. de Visser, S.-C. Park, and J. Krug, *Exploring the effect of sex on an empirical fitness landscape*, Am. Nat., 174 (2009), p. S15.

[32] J. A. G. M. de Visser. private communication.

[33] J. A. G. M. de Visser, T. F. Cooper, and S. F. Elena, *The causes of epistasis*, Proc. Roy. Soc. B, 278 (2011), pp. 3617–3624.

[34] B. Derrida, *Random-energy model: Limit of a family of disordered models*, Phys. Rev. Lett, 45 (1980), pp. 79–82.

[35] B. Derrida, *Random-energy model: An exactly solvable model of disordered systems*, Physical Review B, 24 (1981), pp. 2613–2626.

[36] D. Dhar, *Theoretical studies of self-organized criticality*, Physica A, 369 (2005), pp. 29–70.

[37] T. Dressler, *Über die Korrelation von Subgraphen aus Fitnesslandschaften*. Bachelor's thesis, Universität zu Köln, 2011.

[38] R. Durrett and V. Limic, *Rigorous Results for the NK Model*, Annals of Probability, 31 (2003), pp. 1713–1753.

[39] A. M. Edwards, *Using likelihood to test for Lévy flight search patterns and for general power-law distributions in nature*, J. Animal Ecology, 77 (2008), pp. 1212–1222.

[40] A. M. Edwards and *et al.*, *Revisiting Lévy flight search patterns of wandering albatrosses, bumblebees and deer*, Nature, 447 (2007), pp. 1044–1048.

[41] A. Einstein, *Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen*, Ann. Phys., 17 (1905), pp. 549–560.

[42] A. Erdély, ed., *Bateman Manuscript Project*, McGraw-Hill, New York, 1953.

[43] M. R. Evans, S. N. Majumdar, and R. K. P. Zia, *Canonical analysis of condensation in factorised steady states*, Journal of Statistical Physics, 123 (2006), pp. 357–390.

[44] S. N. Evans and D. Steinsaltz, *Estimating some features of* NK *fitness landscapes*, The Annals of Probability, 12 (2002), pp. 1299–1321.

[45] W. J. Ewens, *Mathematical Population Genetics I. Theoretical Introduction*, Springer, New York, second edition ed., 2004.

[46] W. Feller, *An Introduction to Probability Theory and It's Applications, Vol. I and II*, Wiley, New York, 1968.

[47] R. A. Fisher, *On the dominance ration*, Proceeding of the Royal Society of Edinburgh, 42 (1922), pp. 321–341.

[48] R. A. Fisher, *The distribution of gene rations for rare mutations*, Proceeding of the Royal Society of Edinburgh, 50 (1930), pp. 205–220.

[49] P. J. Flory, *Molecular size distribution in three dimensional polymers*, J. Am. Chem. Cos, 63 (1941), pp. 3083–3090.

[50] F. G. Foster and A. Stuart, *Distribution-free test in time-series based on the breaking of records*, Journal of the Royal Statistical Society B, XVI (1954), pp. 1–22.

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

[52] J. Franke, G. Wergen, and J. Krug, *Records and sequences of records from random variables with a linear trend*, J. Stat. Mech.: Theor. Exp., P10013 (2010).

[53] ——, *Correlations of record events as a test for heavy-tailed distributions*, arXiv:1105.3915, (2011).

[54] N. Franke, *Versuche zur Täteridentifizierung an Minimalspuren mittels DNA-Typisierung*. MD thesis, Universitätsklinikum Hamburg-Eppendorf, 2010.

[55] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics*, Wiley & sons, New York, 1978.

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

[57] S. Gavrilets and J. Gravner, *Percolation on the fitness hypercube and the evolution of reproductive isolation*, J. Theor. Biol., 184 (1997), pp. 51–64.

[58] J. H. Gillespie, *Population Genetics, A Concise Guide*, The John Hopkins University Press, Baltimore, second edition ed., 2004.

[59] N. Glick, *Breaking records and breaking bords*, Am. Math. Monthly, 85 (1978), pp. 2–26.

[60] I. Gradshteyn and I. Ryzhik, *Table of Integrals, Series and Products*, Academic Press, San Diego, 6ht ed., 2000.

[61] G. Grimmett, *The Random Cluster Model*, Springer, New York, 2006.

[62] S. Gulati and W. J. Padgett, *Parametric and Nonparametric Inference from Record-Breaking Data*, Springer, New York, 2003.

[63] E. J. Gumbel, *Statistical theory of extreme values and some practical applications: A series of lectures*, U.S. National Bureau of Standards, Applied Mathematics Series, 33 (1954).

[64] J. B. S. Haldane, *A mathematical theory of natural selection*, Proc.eeings of the Cambridge Philosohpical Society, 27 (1931), pp. 137–142.

[65] A. Hall and J. R. Knowles, *Directed selective pressure on a β-lactamase to analyse molecular changes involved in development of enzyme function*, Nature, 264 (1976), pp. 803–804.

[66] G. H. Hardy, *Mendelian proportions in a mixed population*, Science, 28 (1908), pp. 49–50.

[67] D. L. Hartl and A. G. Clark, *Principles of Population Genetics*, Palgrave Macmillan, 4 ed., 2007.

[68] A. K. Hartmann and H. Rieger, *Optimization Algorithms in Physics*, Wiley-VCH, 2001.

[69] E. J. Hayden, E. Ferrada, and A. Wagner, *Cryptic genetic variation promotes rapid evolutionary adaptation in an rna enzyme*, Nature, 474 (2011), pp. 92–95.

[70] B. D. Hughes, *Random Walks and Random Environments Volume 1: Random Walks*, Oxford Sciences Publications, Clarendon Press, Oxford, 1995.

[71] L. D. Hurst, *Molecular genetics: The sound of silence*, Nature, 471 (2011), pp. 582–583.

[72] J. S. Huxley, *Evolution: The Modern Synthesis*, MIT Press, 2010.

[73] K. Jain, *Number of adaptive steps to a local fitness peak*, arXiv:1109.1235, (2011).

[74] K. Jain, J. Krug, and S.-C. Park, *Evolutionary advantage of small populations on complex fitness landscapes*, Evolution, 65 (2011), pp. 1945–1955.

[75] P. Joyce, D. R. Rokyta, C. J. 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 (2011), pp. 1627–1643.

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

[77] A. I. Khan, D. M. Dinh, D. Schneider, R. E. Lenski, and T. F. Cooper, *Negative epistasis between beneficial mutations in an evolving bacterial population*, Science, 332 (2011), pp. 1193–1196.

[78] M. Kimura, *Process leading to quasi-fixation of genes in natural populations dues to random fluctuation of selection intensities*, Genetics, 39 (1954), pp. 280–295.

[79] ——, *On the probability of fixation of mutant genes in a population*, Genetics, 47 (1962), pp. 713–719.

[80] ——, *Diffusion models in population genetics*, Journal of Applied Probability, 1 (1964), pp. 177–232.

[81] M. Kimura, *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, 1983.

[82] M. Kimura and T. Ohta, *The average number of generations until fixation of a mutant gene in a finite population*, Genetics, 61 (1969), pp. 736–771.

[83] J. Kingman, *A simple model for the balance between selection and mutation*, J. Appl. Probab., 15 (1978), pp. 1–12.

[84] A. Kloezer, *Diplomarbeit: NK Fitness Landscapes*, Universität zu Köln, 2008.

[85] E. V. Koonin, *Are there laws of genome evolution?*, PLoS Comput. Biol., 7:e1002173 (2011).

[86] J. Krug, *Records in a changing world*, J. Stat. Mech.: Theo. Exp., P07001 (2007).

[87] J. Krug and K. Jain, *Breaking records in the evolutionary race*, Physica A, 358 (2005), pp. 1–9.

[88] J. Krug and C. Karl, *Punctuated evolution for the quasispecies model*, Physica A: Stat. Mech. Appl., 318 (2003), pp. 137–143.

[89] P. Le Doussal and K. J. Wiese, *Driven particle in a random landscape: disorder correlator, avalanche distribution and extreme value statistics of records*, Phys. Rev. E, 79, 051150 (2009).

[90] V. Limic and R. Pemantle, *More rigorous results on tthe kauffman-levin model of evolution*, Annals of Probability, 32 (2004), pp. 2149–2178.

[91] G. Löffler, *Basiswissen Biochemie*, Springer, Berlin, 4th ed., 2003.

[92] M. A. Lomholt, K. Tal, R. Metzler, and J. Klafter, *Lévy search strategies in intermittend search processes are advantageous*, Proc. Natl. Acad. Sci. USA, 105 (2008), pp. 11055–11059.

[93] E. R. Lozovsky, T. Chookajorn, K. M. Brown, M. Imwong, P. J. Shaw, S. Kamchonwongpaisan, D. Neafsey, D. M. Weinreich, and D. L. Hartl, *Stepwise acquisition of pyrimethamine resistance in the malaria parasite*, Proc. Natl. Acad. Sci. USA, 106 (2009), pp. 12025–12030.

[94] M. Lunzer, S. P. Miller, R. Felsheim, and A. M. Dean, *The biochemical architecture of an ancient adaptive landscape*, Science, 310 (2005), pp. 499–501.

[95] S. N. Majumdar, A. Comtet, and R. M. Ziff, *Unified solution of the expected maximum of a discrete time random walk and the discrete flux to a spherical trap*, Journal of Statistical Physics, 122 (2006), pp. 833–856.

[96] S. N. Majumdar and R. M. Ziff, *Universal record statistics of random walks and lévy flights*, Phys. Rev. Lett., 101, 050601 (2008).

[97] G. Martin and T. Lenormand, *The distribution of beneficial and fixed mutation fitness effects close to an optimum*, Genetics, 179 (2008), pp. 907–916.

[98] G. Mendel, *Versuche über Pflanzen-Hybriden*, Verhandlungen des Naturforschenden Vereins zu Brünn, 4 (1866), pp. 3–47.

[99] M. Mezard and A. Montanari, *Information, Physics and Computation*, Oxford University Press, 2009.

[100] M. Mezard, G. Parisi, and M. A. Virasoro, *Spinglass Theory and Beyond*, World Scientific, Singapore, 1987.

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

[102] E. W. Montroll and H. Scher, *Random walks on lattives. iv. continuous-time walks and influence of absorbing boundaries*, J. Stat. Phys., 9 (1973), pp. 101–135.

[103] J. Neidhart, *Adaptive Walks in Random and Correlated Fitness Landscapes*, Master's thesis, Universität zu Köln, 2011.

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

[105] V. B. Nevzorov, *Records: Mathematical Theory*, American Mathematical Institute, 2000.

[106] M. Newman and G. J. Eble, *Power spectra of extinction in the fossil record*, Proc. R. Soc. London B, 226 (1999), pp. 1267–1270.

[107] M. E. J. Newman, *The structure and function of complex networks*, SIAM Review, 45 (2003), pp. 167–256.

[108] M. A. Nowak, *Evolutionary Dynamics*, The Belknap Press of Harvard University Press, Cambridge, Massachusets, 2006.

[109] L. P. Oliveira, H. J. Jensen, M. Nicodemi, and P. Sibani, *Record dynamics and the observed temperature plateau in the magnetic creep-rate of type-II superconductors*, Phys. Rev. B, 71, 104526 (2005).

[110] H. A. Orr, *The population genetics of adaptation: the distribution of factors fixed during adaptive evolution*, Evolution, 52 (1998), pp. 935–949.

[111] H. A. Orr, *The distribution of fitness effects among beneficial mutations*, Genetics, 163 (2003), pp. 1519–1526.

[112] H. A. Orr, *The distribution of fitness effects among beneficial mutations in fisher's geometric model of adaptaion*, J. Theo. Biol., 238 (2006), pp. 279–285.

[113] B. Ostman, A. Hintze, and C. Adami, *Impact of epistasis and pleiotropy on evolutionary adaptation*, Proc. R. Soc. B, (2011). published online.

[114] S.-C. Park and J. Krug, *Clonal interference in large populations*, Proc. Natl. Acad. Sci. USA, 104 (2007), pp. 18135–18140.

[115] ——, *Evolution in random fitness landscapes: The infinite sites model*, J. Stat. Mech.: Theor. Exp., P04014 (2008).

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

[117] J. L. Parmley and L. D. Hurst, *How do synonymous mutations affect fitness*, BioEssays, 29 (2007), pp. 515–519.

[118] J. Pickands III, *Statistical inference using extreme order statistics*, The Annals of Statistics, 3 (1975), pp. 119–131.

[119] M. J. Plank and E. A. Codling, *Sampling rate and misidentification of Lévy and non-Lévy movement paths*, Ecology, 90 (2009), pp. 3546–3553.

[120] F. J. Poelwijk, D. J. Kiviet, D. M. Weinreich, and S. J. Tans, *Empirical fitness landscapes reveal accessible evolutionary paths*, Nature, 445 (2007), pp. 383–386.

[121] E. P. Raposo, S. V. Buldyrev, M. G. E. da Luz, M. C. Santos, H. E. Stanley, and G. M. Viswanathan, *Dynamical robustness of Lévy search strategies*, Phys. Rev. Lett., 91 (2003).

[122] S. Redner. http://physics.bu.edu/∼redner/projects/citation/index.html.

[123] S. Redner, *How popular is your paper? an empirical study of the citation distribution*, EPJ B, 4 (1998), pp. 131–134.

[124] A. Reynolds, *How many animals really do the Lévy walk? A comment*, Ecology, 89 (2008), pp. 2347–2351.

[125] A. M. Reynolds and C. J. Rhodes, *The Lévy fligh paradigm: random search patterns and mechanisms*, Ecology, 90 (2009), pp. 877–887.

[126] D. R. Rokyta, C. J. Beisel, P. Joyce, M. T. Ferris, C. L. Burch, and H. A. Wichman, *Beneficial fitness effects are not exponential for two viruses*, J. Mol. Evo., 67 (2008), pp. 368–376.

[127] P. A. Romero and F. H. Arnold, *Exploring protein fitness landscapes by directed evolution*, Nature Reviews, 10 (2009), pp. 866–876.

[128] S. M. Ross, *Stochastic Processes*, John Wiley & Sons, 1983.

[129] S. Sabhapandit, *Record statistics of continuous time random walk*, Europhys. Lett., 94, 20003 (2011).

[130] S. Sabhapandit and S. N. Majumdar, *Density of near-extreme events*, Phys. Rev. Lett., 98, 140201 (2007).

[131] M. L. M. Salverda, *On the natural and laboratory evolution of an antibiotic resistance gene*, PhD thesis, Wageningen University, Wageningen, The Netherlands, 2008.

[132] M. L. M. Salverda, E. Dellus, F. A. Gorter, A. J. M. Debets, J. van der Oost, R. F. Hoekstra, D. S. Tawfik, and J. A. G. M. de Visser, *Initial mutations direct alternatics pathways of protein evolution*, PLoS Genetics, 7:e1001321 (2011).

[133] Z. E. Sauna and C. Kimchi-Sarfaty, *Understanding the contribution of synonymous mutations to human diseases*, Nature Reviews Genetics, 12 (2011), pp. 683–691.

[134] M. Schenk and *et al.*, *Title: tba*, Journal: tba, (2011). in preparation.

[135] P. Sibani, M. Brandt, and P. Alstrom, *Evolution and extinction dynamics in rugged fitness landscapes*, Int. J. Mod. Phys. B, 12 (1998), pp. 361–391.

[136] P. Sibani, G. F. Rodriguez, and G. G. Kenning, *Intermittent quakes and record dynamics in the thermoremanent magnetization of a spin-glass*, Phys. Rev. B, 74,224407 (2006).

[137] D. Sornette, *Critical Phenomena in Natural Sciences*, Springer, Berlin, 2004.

[138] H. Spencer, *Principles of Biology*, vol. 1, 1880.

[139] D. Stauffer and A. Aharony, *Introduction to Percolation Theory*, Taylor and Francis, 1991.

[140] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry*, Elsevier, 4th, reprinted ed., 2006.

[141] G. M. Viswanathan and *et al.*, *Lévy flight search patterns of wandering albatrosses*, Nature, 381 (1996), pp. 413–415.

[142] W. Weinberg, *Über den Nachweis der Vererbung beim Menschen*, Jahrhefte des Vereins für Vater. Naturkunde in Württemberg, (1908).

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

[144] E. D. Weinberger, *Local properties of Kauffman's $N-K$ model: A tunably rugged energy landscape*, Physical Review A, 44 (1991), pp. 6399–6413.

[145] ——, *$\mathcal{NP}$ Completeness of Kauffman's NK Model, a Tunably Rugged Fitness Landscape* , Santa Fé Institute Working papers, (1996).

[146] D. M. Weinreich, N. F. Delaney, M. A. DePristo, and D. L. Hartl, *Darwinian evolution can follow only very few mutational paths to fitter proteins*, Science, 312 (2006), pp. 111–114.

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

[148] E. Weisstein, *q-Pochhammer symbol*. From *MathWorld*-A Wolfram Web Resource. http://mathworld.wolfram.com/q-PochhammerSymbol.html.

[149] G. Wergen, M. Bogner, and J. Krug, *Record statistics for biased random walks, with an application to financial data*, Phys. Rev. E, 83, 051109 (2011).

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

[151] G. Wergen and J. Krug, *Record-breaking temperatures reveal a warming climate*, Europhys. Lett., 92, 30008 (2010).

[152] S. Wright, *Evolution in Mendelian populations*, Genetics, 16 (1931), pp. 97–159.

[153] S. Wright, *The roles of mutation, inbreeding, crossbreeding and selection in evolution*, in Proceedings of the Sixth International Conference of Genetics, 1932, pp. 356–366.

Frequently used abbreviations

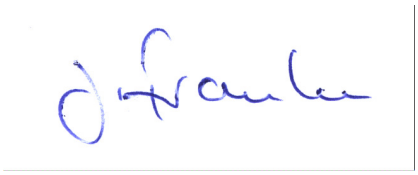
Abbreviation	Meaning
LDM	Linear drift model
RMF	Rough Mt. Fuji
HoC	House of cards
BoA	Basin of attraction
AS	Antipodal sequence
GM	Global maximum
SDM	Site directed mutagenesis
DNA	Desoxyribonucleic acid
RNA	Ribonucleic acid
PCR	Polymerase chain reaction
SG	Sub-graph
EVT	Extreme value theory
iid	Independently, identically distributed
RV	Random variable
WT	Wild type
HTI	Heavy tail indicator
SS/WM	Strong selection /weak mutation
WTD	Waiting time density

Danksagungen

Zunächst möchte ich mich sehr herzlich bei Herrn Prof. Dr. Joachim Krug für die geduldige Betreuung dieser Doktorarbeit bedanken und dafür, daß er mir die Arbeit auf diesem Gebiet ermöglicht hat. Von seiner Intuition bezüglich bezüglich interessanter Fragen habe ich sehr profitiert.
Desweiteren gilt mein Dank J. Arjan G. M. de Visser und Martijn Schenk für die sehr interessante Zusammenarbeit und die Betreuung und Anleitung meines Laborpraktikums in Wageningen.
Martijn Schenk danke ich darüberhinaus dafür, daß er mir gestattet hat, seine bisher unveröffentlichten Datensätze in dieser Doktorarbeit zu verwenden.
Bei Gregor Wergen bedanke ich mich für sehr produktive Zusammenarbeit an den Fragestellungen zur Rekordstatistik und für das Korrekturlesen dieser Doktorarbeit.
Für das Korrekturlesen möchte ich mich auch bei Johannes Neidhart und Ivan Szendro bedanken. Allen drei Korrekturlesern danke ich nicht nur für wertvolle Hinweise und Anmerkungen während der Redaktion dieser Arbeit, sondern auch für die angenehme Arbeitsatmosphäre in unserem Büro, zu der auch Marian Ivanov beigetragen hat.
Diese Arbeit wurde mit einem Stipendium von der Studienstiftung des deutschen Volkes sowie der Bonn-Cologne Graduate School of Physics and Astronomy gefördert. Für diese Förderung bedanke ich mich.
Diese Arbeit wäre nicht möglich gewesen ohne die Unterstützung und den Zuspruch meiner Eltern Barbara und Wolfgang, meiner Schwester Nora und insbesondere meiner Verlobten Mareike Steen. Meiner Verlobten verdanke ich darüberhinaus wichtige Anmerkungen zu dieser Doktorarbeit sowie zu den biochemischen Grundlagen des QuickChange$^{\mathrm{TM}}$Verfahrens. Deshalb ist es mir ein besonderes Vergnügen, mich an dieser Stelle dafür zu bedanken.

Erklärung

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 unten 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 Herrn Prof. Dr. Joachim Krug betreut worden.



Köln, d. 17.3.2012 (Jasper Franke)

Publikationsliste

- J. Franke, G. Wergen und J. Krug, *Records and Sequences of Records from Random Variables with a Linear Trend*, J. Stat. Mech. **P10013** (2010)
- J. Franke, A. Klözer, J. A. G. M. de Visser und J. Krug, *Evolutionary Accessibility of Mutational Pathways*, PLoS Comp. Biol. **7**, 8 (2011)
- G. Wergen, J. Franke und J. Krug, *Correlations Between Record Events in Sequences of Random Variables with a Linear Trend*, J. Stat. Phys. **144**, 6 (2011), pp. 1206-1222; *Erratum to: Correlations Between Record Events in Sequences of Random Variables with a Linear Trend*, J. Stat. Phys. **145**, 5 (2011), pp. 1405-1406
- J. Franke, G. Wergen und J. Krug, *Correlations of record events as a test for heavy-tailed distributions*, Phys. Rev. Lett. **108**, 064101 (2012)