Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature: _____ Date _____

Jakub Otwinowski

Speed of evolution with spatial structure and interacting mutations

By

Jakub Otwinowski
Doctor of Philosophy

Physics

Stefan Boettcher
Advisor

Ilya Nemanman
Advisor

David J. Cutler
Committee Member

H. G. E. Hentschel
Committee Member

Fereydoon Family
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D. Dean of the James T. Laney School of Graduate Studies

---

Date

Speed of evolution with spatial structure and interacting mutations

By

Jakub Otwinowski
MSc., Universiteit van Amsterdam, 2007
B.S. University of Texas at Austin, 2005

Advisor: Stefan Boettcher, PhD
Advisor: Ilya Nemenman, PhD

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Physics
2012

Abstract

Speed of evolution with spatial structure and interacting mutations

By Jakub Otwinowski

Perhaps the simplest question about long term evolutionary adaptation is how quickly do populations adapt to a new environment by incorporating mutations? This question is approached from several different angles. Chapter 1 investigates the speed of evolution when there is a large supply of beneficial mutations and the population has spatial structure. For large system sizes, a speed limit is found on the rate adaptation. The model is analyzed as a surface growth model in physics, which reveals universal properties of the model, such as the distribution of fitnesses. However, neglecting spatial structure, the speed of evolution also depends on how mutations interact with each other. This may be quantified by a fitness landscape, or a genotype-phenotype-fitness map. In chapter 2, the fitness landscape and genotype-phenotype map of an E. coli lac promoter is inferred from a large dataset with 100,000 sequences and fluorescence measurements. The interactions between mutations are quantified using a simple quadratic model, similar to a spin glass Hamiltonian. Chapter 3 describes a toy model based on an overdamped particle in a potential, which demonstrates how a fitness landscape with time dependent interactions between mutations determines the speed of evolution.

Speed of evolution with spatial structure and interacting mutations

By

Jakub Otwinowski
MSc., Universiteit van Amsterdam, 2007
B.S. University of Texas at Austin, 2005

Advisor: Stefan Boettcher, PhD
Advisor: Ilya Nemenman, PhD

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Physics
2012

Acknowledgements

The work about spatially structured populations was done under the supervision of Stefan Boettcher and Joachim Krug. The inference of the genotype-phenotype map was done under the supervision of Ilya Nemenman. The fluctuating epistatic model was done with Sorin Tanase-Nicola, and Ilya Nemenman. I also owe many thanks to Dave Cutler, Martin Tchernookov, Ivan Szendro, and Armita Nourmohammad for helping me with my work, Justin Kinney, Phillip Johnson, and Thierry Mora for comments on the manuscript, and Bruce Levin for getting me interested in microbial evolution.

Contents

List of Figures

Chapter 1

Introduction

1.1 Background

The mathematical theory of evolution, or population genetics, was founded early in the 20th century by geneticists Sewall Wright, J. B. S. Haldane and R. A. Fisher. Population genetics began with understanding the fundamental processes of evolution: natural selection, mutation, genetic drift (or stochasticity), migration, and recombination[1]. Central to natural selection is the concept of *fitness,* that is the ability of an organism to propagate its genes to the next generation. Often it suffices to define fitness as the relative growth rate. Mutation today is known to be a change in genetic code, however a simplified definition of mutation is a change in fitness of an individual. Genetic drift is a name that encompasses the stochastic nature of reproduction and survival. For example an individual could have a random number of children with some probability distribution with a mean defined as the fitness. Migration is the movement of individuals from one population to another, transferring genetic information. Recombination is the shuffling of genes on a chromosomes of offspring in sexual populations.

These processes provided population geneticists decades of research, most of which were devoted to explaining existing variation seen in populations. Existing variation in physical traits was, for a long time, the only type of observable. Eventually genetic data started to appear, but the models were still simple, in that they tracked only one or two locations in the genome.

However, in the last few decades, advances have increased our ability to sequence genetic code dramatically. In just the last few years the costs of sequencing have fallen faster than Moore's law [2]. The ability to know the genetic code, along with automated measurements of the physical traits and fitness, has produced a wave of novel evolutionary experiments and large datasets from existing populations, often including whole genomes instead of a few locations. This new data raised questions that were not amenable with classical population genetics. Variation existed in many locations in a genome, not just one. Beneficial mutations were common in large populations, not rare [3]. Selection was dynamic over many different time scales, not static[4]. These realizations provided a new starting point for theorists.

Here we review some recent trends in mathematical evolution, and present new research in this context.

1.2 Evolution of large populations

The rate of adaptation was traditionally thought to be mutation limited. That is, beneficial mutations would sweep a population one at a time, and the rate of adaptation was simply the rate that mutations appeared and survived. However, when a population is adapting to a new environment, beneficial mutations may be much more common, and this was noticed in microbial evolution experiments [5]. The best known example is Lenski's long-term experiment, which has been running since 1988, and has gone through 50,000 generations [6]. In this experiment, 12 parallel bacterial cultures were grown in liquid medium, and every day 1% were transferred to a new flask with fresh medium. Since they froze samples at many generations in the past, they could go back and sequence them once the technology became more affordable. Instead of a population with one or two variants, they found many beneficial mutations existed at the same time and competed with each other [3].

It turns out it is very difficult to calculate the rate of adaptation of large populations of asexuals. The problem was originally posed by Fisher [7] and Muller [8], but only recently there has been an intense theoretical effort to solve this problem [9, 10, 11, 12, 13, 14]. In large populations there is much genetic variation and therefore a distribution of fitnesses, typically Gaussian. This distribution advances in time as new mutations are discovered,

and can be visualized as a wave traveling towards ever higher fitness. A key insight was that it is important to consider the fluctuations in the leading edge of this traveling wave [15]. The result is that intense competition between beneficial mutations reduces the rate of adaptation, such that it does not depend on the total supply of new mutations, but depends logarithmically on population size and mutation rate.

1.3 Evolution with spatial structure

The previously mentioned studies assumed that populations were *well-mixed*, in reference to microbes living in well-mixed liquid medium. Each individual has the chance to compete with the population as a whole, or, in the language of statistical physics, they are mean-field models that do not take into account any spatial structure.

Spatial structure was considered for simpler situations (without many competing mutations) in classical population genetics. Models had populations on different discrete units called islands, demes, or stepping stones, and every generation a fraction of the population was allowed to migrate to connected units [16, 17]. Also early on, Fisher, Kolmogorov and others formulated a partial differential equation with traveling wave solutions to describe the propagation of a gene in continuous space [18, 19]. Now known as the FKPP equation after the authors, it is the simplest type of reaction-diffusion equation, widely studied in physics and chemistry.

More recently, large geographic surveys of human genomic diversity [20] inspired theorists to consider the evolutionary dynamics of genes on the edge of range expansions of populations [21, 22, 23]. They also revisited stepping stone models to calculate the dynamics of desegregation of multiple variants in one dimension [24], and made inferences of selective sweeps in expanding microbial colonies [25].

Speed of evolution in spatially structured populations

When a spatially structured populations adapts to a new environment, there are many beneficial mutations, just as in well-mixed populations. However, the time for mutations to spread to the whole population is much slower than in well-mixed populations, and it

is more likely that there is competition between beneficial mutations [26, 27]. In chapter 2, we propose a model of competition in spatially structured populations and find that the speed of evolution is qualitatively different when there is spatial structure. Above a critical system size there is a speed limit on adaptation, and the speed does not depend on the total population size, or the total supply of mutations. Furthermore, the evolutionary model is nearly equivalent to a model in surface growth physics that is in a widely studied universality class. We show that the evolutionary model is also in this universality class because it shares the same power law growth of the width of the fitness distribution, and the distribution itself is of the known universal form.

1.4 Epistasis and fitness landscapes

In classical population genetics, mutations were simplified as small changes in fitness whose effects were independent of each other. In reality it is much more complex. Errors in the duplication of genetic code lead to different physical traits, which in turn influence the reproductive success of an organism. The genetic code is very long, and mutations may occur in many locations in the genome, termed *loci*. Mutations at different loci may not have independent effects and may interact with each other, such that the combined effect is different from the sum of the individual effects. These interactions, termed *epistasis,* are interesting for systems and developmental biologists, who would like to determine how genetic elements interact to cause the physical traits and functions of an organism [28, 29, 30, 31, 32, 33]. Large automated studies have generated huge new datasets for this purpose. For example, recently 5.4 million pairs of genes were analyzed in yeast to create a genome scale interaction network [29].

Epistasis also has important evolutionary consequences, but to explain this requires a more general description. Besides interactions between organisms, fitness is determined by a set of physical traits (phenotype), which is in turn determined by the genetic sequence (genotype). Therefore there is a map from the high dimensional discrete genotype space, to the lower dimensional (discrete or continuous) phenotype space, to the one dimensional fitness. This map may also be dependent on the environment. An individual's genome

determines a point in genotype space, and a population is therefore a cloud of points. This cloud moves generally toward higher fitness, and in analogy to energy landscapes, the genotype-fitness map is often called a fitness landscape. A landscape with no epistasis is formulated as a single peak with linear slopes.

From the physics of spin glasses, it is known that with just pair interactions it is possible to have an energy landscape with many local minima, where the system may be trapped for long periods of time. It is possible that fitness landscapes are also rough, and evolution proceeds slowly by jumping from peak to peak.

Kauffman and Levin first studied rough fitness landscapes with their tunably rugged NK landscapes [34], similar to spin glasses. NK landscapes developed into a large body of work that studied the dynamics of adaptation [35, 36, 37], mostly within the physics and computer science communities. Another body of work developed around making more complicated landscapes with more realistic assumptions, such as modeling the folding of RNA molecules [38, 39, 40].

An orthogonal approach is to ignore the high dimensional genotype space, and model mutations as random numbers drawn from some distribution. This was done by Kingman, with uncorrelated adaptive steps [41], and by Gillespie who utilized extreme value theory to identify features that are independent of the underlying fitness landscape [42].

Recent experiments on the adaptation of microbes [43, 44] have renewed interest in the theory of adaptation in (random) fitness landscapes. For example Orr further developed the use of extreme value theory to argue that the distribution of beneficial mutations should be exponential [45, 46, 47]. Plotkin et. al analyzed fitness trajectories in order to infer properties of the underlying fitness landscape [48].

Experimentalists have also begun mapping high-dimensional fitness landscapes, not just epistatic pairs, but whole portions of genotype space designated by all possible combinations of a limited number of loci (for a list of such landscapes see [49]). The goal is to learn not only the genotype-fitness map, but also the evolutionary consequences, such as the reproducibility of trajectories [50, 51, 52] or the accessibility of genotype space [53]. For example Weinreich et. al constructed all possible combinations of five mutations that result in antibiotic resistance of a bacterium, and they found that epistasis between mutations

severely limited the number of evolutionarily accessible paths. The datasets have become larger and larger. Pitt et. al's found the fitness landscape for 10^7 unique sequences of RNA with fitnesses measured by an in vitro selection protocol [54], and Hinkey et. al collected sequences and fitnesses from thousands of HIV samples from which they inferred a fitness landscape [55].

Genotype-phenotype map and fitness landscape of an E. coli *lac* promoter

It is very difficult to measure a complete fitness landscape, because a complete map would require a fitness measurement for every possible sequence, and the number of possible sequences grows exponentially with sequence length. In chapter 3 we reduce the number of required measurements by fitting models of reduced complexity to data. We infer the best linear and quadratic approximations of a landscape from the recently collected 200,000 mutated sequences (of 75 base-pairs) from a bacterial gene regulatory region. We find that the linear or additive mutations account for about 2/3 of the variance in the data, while the pairwise interactions of mutations account for about 10% of the variance. From simulations we find that most of the genotypes are evolutionary accessible, with very few valleys in the fitness landscape. We also discuss the presence of selection on the wild-type sequence, and its optimality under two different environmental conditions. We show that it is feasible to approximate these high dimensional landscapes in a simple way, and we describe some important pitfalls of this type of inference.

1.5 Fluctuating selection

So far the influence of the environment has been neglected. Environmental influence may be on any level: genotype-phenotype, phenotype-fitness, or interactions between individuals, and crucially, it may change over time. A convenient way to describe this effect is to make the fitness landscape time dependent, or in other words selection fluctuates in time.

Classical population genetics considered the rate of evolution and standing variation for fluctuating selection on short timescales, shorter than the timescale of genetic drift, such as fast changes in ecology due to seasons [56, 42, 57, 58, 59]. The related field of

population ecology considered fluctuating selection in density regulated populations [60, 61, 62]. Population dynamics, which modeled explicit interactions between individuals was also analyzed with fluctuating selection [63, 64]. Mustonen and Lässig used theorems from non-equilibrium statistical physics to characterize evolutionary trajectories with fluctuating selection on longer timescales, on the order of timescales of the occurrence of mutations [65, 66, 67].

The effects of fluctuating selection are not easy to generalize, and seem to be heavily context dependent. Experimental efforts have looked for genetic divergence [4] and tradeoffs and optimality [68] under variable environments.

Speed of evolution with fluctuating epistasis

A rough, fluctuating fitness landscape may have large effects on the rate of adaptation. A position in the genotype space may be at a local optimum at one point in time, and on a slope at a different point in time. Kashtan et. al showed that it is possible to have such a speedup in modular, time varying landscapes [69, 70]. Chapter 4 introduces evolution on a dynamic, rough fitness landscape, where it is possible to quantify the dependence of the speedup on the time scales and length scales of the landscape.

To represent an epistatic fitness landscape, we consider a one-dimensional inclined sawtooth function. Environmental change in epistasis is model as a periodic flip in the teeth of the saw. The evolutionary dynamics is modeled as an overdamped Brownian particle in a potential. The particle represents a homogenous population which moves in genotype space towards higher fitness, while the potential is the fitness landscape. We find that the fluctuating epistatic landscape can greatly enhance the motion of the population in genotype space, similar to some models of stochastic ratchets, and we quantify the dependence on the parameters of the landscape. Similarly, the probability to escape local optima is enhanced.

1.6 Final thoughts

The quantitative study of evolution has been accelerating in the last decade, thanks to advancements in microbial experiments, sequencing, and a large interest in closing the

theoretical gaps. Many important topics, aside from the previously mentioned ones, were simmering for decades, and have recently come to a boil. The evolutionary advantages and disadvantages of recombination have been explored in very many contexts [71, 72, 73, 74, 75, 76, 77, 78]. Others have picked apart the apparently contradictory qualities of evolutionary robustness and evolvability [79, 80, 81]. Quantitative genetics, the study of large numbers of loci and the resulting continuous traits, is benefiting from sophisticated analogies to statistical mechanics [82, 83].

Sometimes subjects which were once thought to be well understood, turned out to be quite complicated. For example, the dynamics of a population crossing a fitness valley, has a surprisingly complex phase diagram with a half dozen regimes [84, 78]. Such surprises give the feeling that this is all just scratching the surface, and evolution will continue to provide us with rich phenomena to study for many more decades.

Chapter 2

Speed of evolution in spatially structured populations

2.1 Introduction

The appearance of a beneficial mutation in a population and its fixation is the most basic process of adaptation. This process determines the rate of evolution, or how quickly populations adapt to new environments, and influences the genetic diversity of a population. Traditionally, it was argued that mutations were rare enough, that the adaptation rate was mutation limited. That is, once a new beneficial mutation appeared, it would sweep the whole population quickly, and the next mutation would be sufficiently separated in time as to not interfere.

However, recent microbial experiments suggest beneficial mutations are more common than previously thought [13, 85, 5]. When multiple beneficial mutations coexist in a population in different lineages, and if there is little or no recombination, they must compete with each other. In this regime of mutation competition, few beneficial mutations survive, reducing the rate of evolution, as seen in microbial evolution experiments [86, 87].

The question of the rate of evolution in asexuals originated from the Fisher-Muller hypothesis for the evolutionary advantage of recombination. Fisher [7] and Muller [8] argued that sexual organisms had an advantage in a large population with multiple beneficial

mutations originating in separate lineages. Sexual populations can recombine two beneficial mutations into one organism in one generation, while asexual populations would have to wait until one of the organisms would independently discover the second mutation in the background of the first.

The still open question of the evolutionary advantage of recombination, and pioneering microbial experiments have renewed interest in the theoretical understanding of the rate of evolution in asexuals. Recent theoretical analyses have found the rate of evolution in large populations of asexuals does not depend on the total rate of mutations, but depends weakly on population size and mutation rates [11, 12, 88, 89, 14], consistent with experiments. However, these analyses were limited to well-mixed populations, where each individual competes with the whole population, such as microbes in a test tube.

Many populations are not well-mixed but confined in space such that they only compete with a limited neighborhood population on timescales of a generation. Examples of populations with spatial structure range from plants and animals over large areas of land, to microbes in biofilms [90] to cancer [91, 92]. Spatial structure is often neglected as an inconvenient detail. In periodic selection a single mutant effectively competes with the whole population, and the fixation probability is the same in well-mixed and spatially structured populations [93, 94]. However, as we will see, in conditions of mutation competition, spatial and well-mixed populations diverge in many aspects.

Fisher and Kolmogorov [18, 19] first described the spread of a single beneficial mutation in a spatially continuous population as a traveling wave solution of a partial differential equation of the form

$$\frac{\partial u}{\partial t} = u(1-u) + \frac{\partial^2 u}{\partial x^2},$$

where u is the fraction of the population with the beneficial mutation. The second derivative represents diffusion in space, and the non-linear term is natural selection. As first noted by Fisher, the spread of this wave is much slower than the exponential growth of a beneficial mutation in a well-mixed population, and there must be a large number of simultaneous waves in a large system.

Here we study a simple model of an asexual population with spatial structure and a

steady rate of beneficial mutations. Instead of a continuous space, our model is discrete, similar to classical models of discrete islands or stepping stones with migration between them [16, 17]. First we must introduce some basic concepts and models from classical well-mixed population genetics.

2.2 Wright-Fisher model

The evolutionary model described here consists of a few main features: reproduction with inheritance, selection, and mutation. Time may be considered discrete, as it is a natural analogy for generations of reproduction. A population consists of many individuals, and each individual, i, has a Poisson distributed number of children per generation with mean f_i, which is called the fitness. If f_i is greater than one, the population has a finite probability of increasing forever. To keep the population finite, one may define the joint probability of each individual to have children conditional on the total population size being constant. This conditional distribution is a multinomial, and defines the Wright-Fisher model. It is equivalent to the following process: to create a new generation, each of the N individuals is assigned a parent from the previous generation randomly, but weighted according to their f_i. Individuals then inherit the fitness from the parents.

After a long time, every individual will have the same fitness, so it is necessary to introduce some mutations. In real populations, genetic mutations can have many effects on the physical traits of an organism. However in this context, the only important effect of a mutation is on the fitness. A simple way to model mutations is to have them increase the fitness multiplicatively as

$$f' = f(1+s) \tag{2.1}$$

where s is called the selection coefficient. The chance for a single mutation to sweep through an otherwise homogenous population is called the fixation probability π. A widely used approximation is given by Kimura [95]:

$$\pi = \frac{1-e^{-2s}}{1-e^{-2Ns}}. \tag{2.2}$$

For beneficial mutations ($1/N \ll s \ll 1$)

$$\pi \approx 2s. \tag{2.3}$$

Even beneficial mutations are unlikely to survive due to genetic drift. For example a beneficial mutation with 1% advantage still has 98% chance of going extinct.

When $s \sim 1/N$ or smaller, the fixation probability becomes $\pi = 1/N$, which means the mutation has the same chance of fixation as any other individual. Harmful mutations have a negligible chance of fixating.

2.3 Periodic selection

Periodic selection is when beneficial mutations appears very rarely, and have time to sweep through the whole population before the next mutation appears. In other words, the time for a mutant to spread is much shorter than the mutation time: $t_{\text{fix}} \ll t_{\text{mut}}$. For a population of size N, and a per individual mutation rate U, there will be UN new mutations per generation in the population. The time for a mutation to appear and survive drift is therefore

$$t_{\text{mut}} = (2sUN)^{-1}, \tag{2.4}$$

while the fixation time, is approximately [96]

$$t_{\text{fix}} = \frac{2}{s} \log N, \tag{2.5}$$

for small s. From the separation of timescales, the periodic selection regime occurs when $UN \ll 1$, that is less than 1 mutation appears in the population per generation. The *speed of evolution* is defined as the average rate of fitness increase. Under periodic selection the speed is mutation limited, therefore it is the inverse of the mutation time, multiplied by the magnitude of each mutation,

$$v = 2s^2UN.$$

Notably the speed depends on the total supply of beneficial mutations, UN.

When $UN \gg 1$, mutations compete with each other. This regime may be called *clonal interference*, or *multiple mutations*, depending on the assumptions made. Interestingly, in this regime the speed does not depend on the total supply of mutations. Instead, $v \sim \log N$, for reasonably large (not infinite) N. The details are outside of the scope of this text (for a review see [96]). Our model concerns the competition regime for spatially structured populations.

2.4 Model with spatial structure

Consider a one dimensional lattice of size L and periodic boundary conditions, where each point represents a single organism that occupies a space. The evolution follows standard Wright-Fisher dynamics, that is generations are discrete and the next fitness of each site is chosen randomly from one of the parents in the neighborhood, weighted according to their fitness. The smallest possible neighborhood is in one dimension such that the child in the next generation inherits the fitness from only two possible parents, thats is $f_i(t+1)$ is chosen from either $f_i(t)$ or $f_{i+1}(t)$.

In the case of a homogenous system of fitness 1, where a single mutant appears with fitness $1+s$, the fixation probability for a beneficial mutation is the same as in the well-mixed case, $\pi = 2s$ [93, 94]. Intuitively, the fixation probability is unaffected because a single mutation has ample time to compete with the entire system, regardless of spatial structure. Since the fixation probability is the same, the speed of evolution in periodic selection is the same as in the well-mixed case. What is different is the timescale of fixation.

The boundary between two domains with different fitnesses may be regarded as a particle performing a biased random walk. The speed of this biased walker is the expected value of its position after one time step, which is $c = s/2$ for small s. In the continuum limit, this model corresponds to a special case of the more general stochastic Fisher equation (or SFKPP equation) [97, 98], where it is possible to have traveling waves with speed $c \sim s$ in the strong noise regime, or $c \sim \sqrt{s}$ in the weak noise regime. However, in the end the dependence of the wave speed on s does not change the essential features.

Importantly, the time for fixation may be much longer in the presence of a spatial

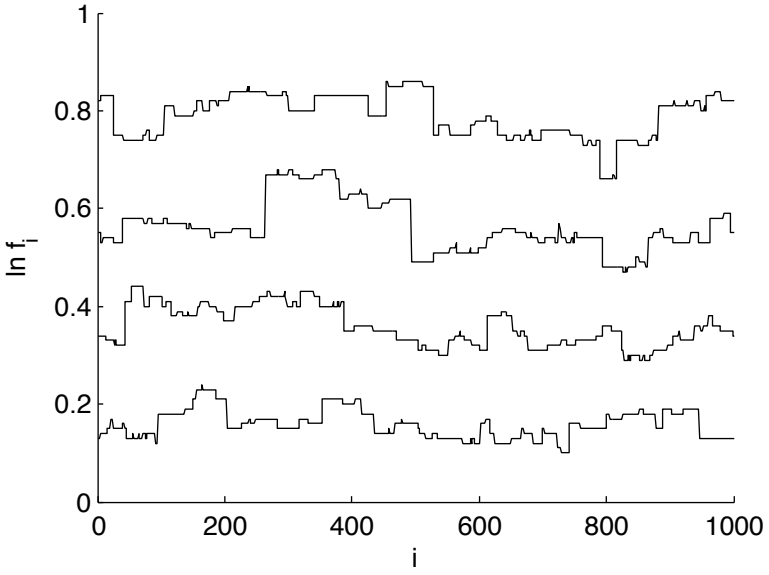

Figure 2.1: The fitness profile at different times during evolution resembles a surface growth process. $N = 1000$, $U = 10^{-3}$, $s = 0.01$. Lines are separated by 10000 generations

structure compared to well-mixed populations. A wave spreading with finite speed c will take time $t_{\text{fix}} \sim L/c$ to cover the whole system (and total population size $N \sim L$), as opposed to a well-mixed population where $t_{\text{fix}} \sim \log(N)$. The slow spread of mutations make it more likely that many exist simultaneously in large systems.

Since we are interested in the rate of evolution during competition, a steady rate of beneficial mutations is supplied, akin to a population adapting to a new environment. Beneficial mutations appear randomly at a steady rate U per site per generation (harmful mutations are unlikely to survive and are neglected). We assume that mutations have independent effects, with no epistasis, and therefore increase the fitness by $\log f' = \log f + s$, where s is a constant on the order of 1%.

2.5 Speed of evolution with spatial competition

Starting from uniform initial conditions, mutations appear in the population, spread, and interfere with each other. Figure 2.1 shows the fitness profile at different time points in the evolution. The average fitness $\bar{f}(t)$ increases exponentially, and a general definition of the

speed of evolution is

$$v = \lim_{t\to\infty} \frac{\langle \ln \bar{f}(t) \rangle}{t}, \tag{2.6}$$

where brackets indicate an ensemble average. While the mean is always moving, the distribution of fitness around the mean reaches a steady state after some time. The distribution will be characterized in a following section.

For very low mutation rates or small system sizes, there should be no competition, and the fitness profile should be constant most of the time. Competition starts when the two timescales are about equal, $t_{\text{mut}} \sim t_{\text{fix}}$, which yields a critical system size for our model,

$$L_c = \frac{1}{2\sqrt{U}}. \tag{2.7}$$

Above L_c, waves of mutations crash into each other, slowing down the rate of adaptation. In simulations is it apparent that above L_c, v saturates and becomes independent of L (figure 2.2). In other words, above the critical system size, the speed of evolution hits a speed limit. Intuitively, when there is competition, the mutations compete with each other on a finite length scale, and not the whole system. This is in contrast to well-mixed populations, where $v \sim \log N$, and never saturates.

In figure 2.3 L is held constant and U is varied, and there is a reduction in v as the transition to competition occurs according to equation 2.7. The standard well-mixed Wright-Fisher model was also simulated. Since the fixation time is $t_{\text{fix}} = 2\ln(N)/s$, the transition happens at a higher rate, $U \sim 1/(4N \ln N)$.

From the observation that v becomes independent of L and dimensional analysis, we can deduce the dependence of v on U. Since $[v] = \frac{1}{[t]}$, $[U] = \frac{1}{[x][t]}$, and the wave speed $[c] = \frac{[x]}{[t]}$, then the only combination of variables without using L is $v \sim (cU)^{1/2}$. From simulations it is difficult to see this exponent because of the slow convergence to asymptotic behavior.

These results describe the mean fitness of the population. In the next section we characterize the variance of the distribution, by an analogy to surface growth models in physics.

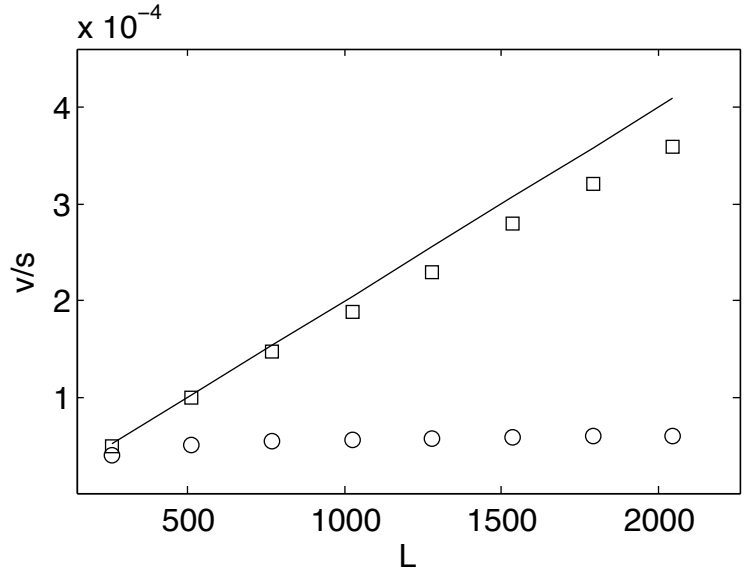

Figure 2.2: The speed of evolution versus system size for the 1D model (circles) and the well-mixed model (squares) with $U = 10^{-5}$. v quickly saturates in 1D but diverges for the well-mixed model. Solid line is $v = 2s^2UL$.

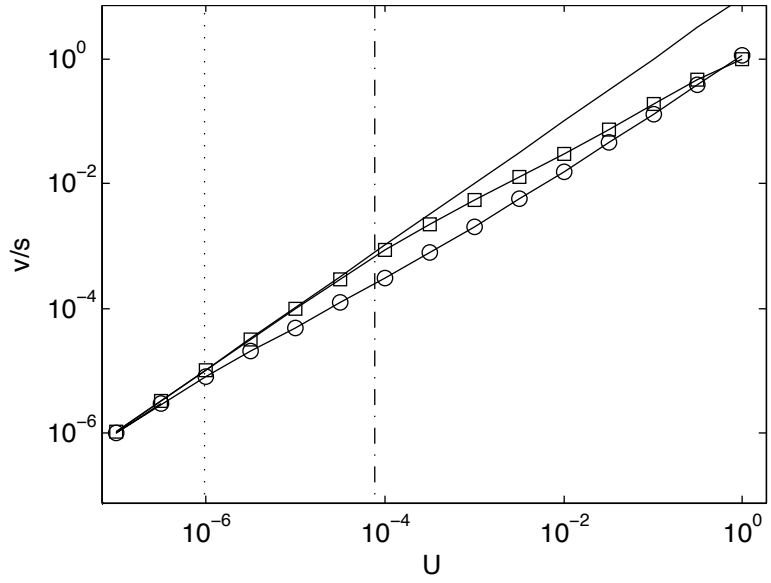

Figure 2.3: The speed of evolution versus mutation rate U with $L = 512$ and $s = 0.01$ averaged over 10^3 fixations for the 1D model (circles) and well-mixed model (squares). The dotted line indicates the transition to competition for the spatial model, and the dot-dashed line indicates the transition for the well-mixed wright-fisher model ($N = 512$). The solid line is the fixation rate for periodic selection ($v = 2s^2UL$).

2.6 Surface growth and universal fitness distributions

If one thinks of the fitness profile as a a physical surface, it becomes apparent that the model described here is similar to surface growth physics [99]. In surface growth physics, particles are deposited on an initially smooth surface randomly, and they may diffuse or stick to each other, gradually forming a rough surface. Many simple models of surface growth were studied by statistical physicists interested in non-equilibrium systems. They discovered that a large number of models share the same properties in the continuum, long-time limit, where many of the microscopic details of the model do not matter, and these classes of models, or universality classes, share the same symmetries.

The evolutionary model defined here is equivalent to a surface growth model called polynuclear growth [100, 101] (PNG), in the limit $s \to \infty$. In PNG, the process of surface growth may be divided into two parts, nucleation (mutation), and spreading (selection). Sites are nucleated with low probability at any point, at a certain rate, U, which corresponds to adding a small block of height to the surface (log fitness). The nucleated site then grows laterally forming a new layer.

While in PNG the spreading is fast and deterministic, in our evolutionary model it is stochastic, and the new layer may even disappear. The boundaries move and collide with each other, and they either annihilate or stack up creating differences in log fitness greater than s. The rate of evolution is equal to the expected increase in log fitness per unit time. Every time step, each fitness difference at position i, contributes to the growth rate on average a rectangle of height Δ_i (from the magnitude of the fitness difference) and width $\Delta_i/2$ (from the expected change in position of the boundary), normalized by L. The total growth rate is the sum of the individual contributions:

$$v = \frac{1}{2L} \sum_i \Delta_i^2.$$

While, the final distribution of height differences is difficult to calculate, it is more instructive to look at the distribution of fitnesses.

In surface growth phenomena, starting from flat initial conditions, the standard devia-

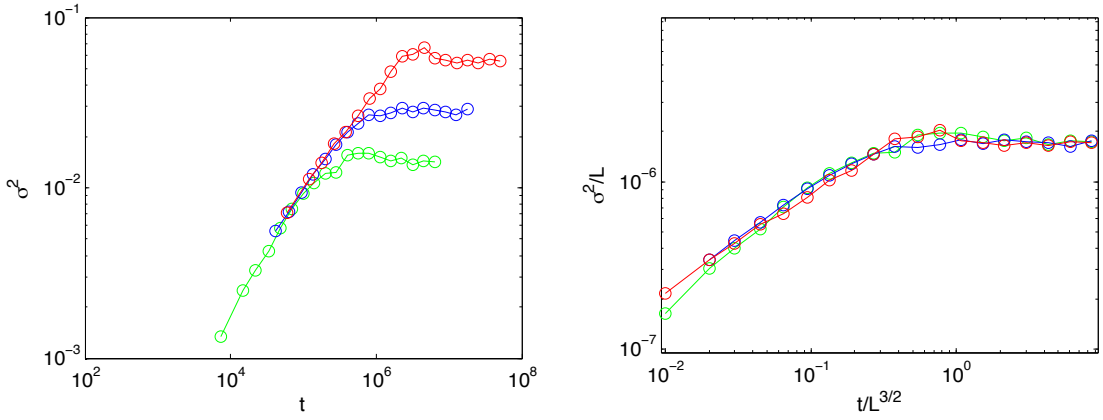

Figure 2.4: (a) The standard deviation of the log fitness distribution as a function of time for different system sizes, $L = 2^{13}$ (green), $L = 2^{14}$ (blue), and $L = 2^{15}$ (red). $s = 0.05$, $U = 10^{-5}$. After a transient regime, σ saturates at a value that depends on L. (b) When the data is rescaled as σ^2/L and $t/L^{3/2}$ it collapses onto the same curve, indicating that in fact $\sigma(t) \sim t^{1/3}$ and $\sigma(t \to \infty) \sim L^{1/2}$, which are the PNG (and KPZ) exponents.

tion of the interface height distribution grows in time as $\sigma(t) \sim t^{\beta}$, where β is the growth exponent, then reaches a steady state when the correlation length reaches the size of the system [99, 102]. In the steady state, $\sigma(t \to \infty) \sim L^{\alpha}$ where α is the saturation exponent. Similarly, in our model the standard deviation of the log fitness distribution also follows power-laws. Figure 2.4a shows the growth of σ for different system sizes. The crossover time is where saturation sets in (the elbow), and it scales as $L^{\alpha/\beta}$. One may try to measure the exponents from the simulations, but since our model is essentially PNG, one can guess that it has the same exponents, $\alpha = 1/2$, and $\beta = 1/3$. Figure 2.4b shows that the data collapses when plotted as σ^2/L and $t/L^{3/2}$, in agreement with the PNG exponents. From surface growth physics, it is known that the distribution is Gaussian in the stationary regime, but in the transient regime it is not, and its form was recently discovered.

The PNG model belongs to the broader Kardar-Parisi-Zhang (KPZ) class of models [103], which share universal scaling exponents and universal distributions. It is known that in the KPZ class in the non-stationary regime, the fitnesses grow as [104]

$$\log f_i(t) = v_\infty t + (\Gamma t)^{1/3} \chi, \tag{2.8}$$

where χ is a random variable from one of the Tracy-Widom distributions, v_∞ is the long-

time growth rate, and Γ is related to the parameters of the KPZ equation. From eq. 2.8 we fine the width of the distribution:

$$\sigma^2 = \mathrm{var}(\log f_i) = (\Gamma t)^{2/3}\mathrm{var}(\chi). \tag{2.9}$$

The Tracy-Widom distributions were first discovered in random matrix theory [105], and their connection to KPZ is itself an interesting story (see [104]). The TW distributions come from the distributions of the largest eigenvalues from different ensembles of random matrices. Different random matrix ensembles (i.e. Gaussian unitary, Gaussian orthogonal) correspond to different geometries in surface growth (i.e. droplet, flat).

Here we show numerically that the distribution of fitnesses in the non-stationary regime is one of the non-Gaussian Tracy-Widom distributions characteristic of PNG and the KPZ class (in this case, the flat initial condition means that the distribution is from the Gaussian orthogonal ensemble). One signature of the TW distributions can be seen by measuring higher moments, that is its skewness, $\langle\left(\frac{\log f - \langle\log f\rangle}{\sigma}\right)^3\rangle$, and (excess) kurtosis, $\langle\left(\frac{\log f - \langle\log f\rangle}{\sigma}\right)^4\rangle - 3$, which do not depend on the parameters v_∞ and Γ. Figure 2.5 shows that the skewness and kurtosis of the fitness distributions are non-zero, indicating non-Gaussianality, and they approach the known values of the GOE Tracy-Widom distribution.

It is possible to compare the fitness distribution directly to the Tracy-Widom distribution. The parameters, v_∞, and Γ can be found by fitting the mean of eq. 2.8 to the mean log fitness from simulations (adding a constant accounts for the very short time behavior and makes the fit better). The fitnesses from the simulation are then rescaled (Fig. 2.6):

$$\chi_{sim} = \frac{\log f_i - v_\infty t}{(\Gamma t)^{1/3}}. \tag{2.10}$$

Figure 2.6 shows that in the non-stationary regime, the fitnesses fall onto the universal distribution, which is skewed towards higher fitnesses, with the right tail $P(\chi)_{\chi\to\infty} \sim \chi^{-3/2}$. To demonstrate the robustness of this result, simulations were also made with fitness difference s generated from an exponential distribution, and the resulting distribution is identical.

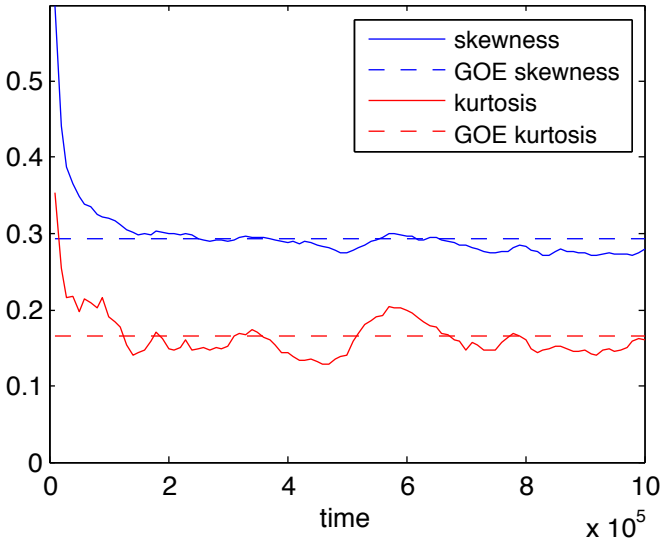

Figure 2.5: Skewness and kurtosis of the fitness distributions (solid lines) and known values for the GOE Tracy-Widom distribution (dashed lines). $L = 2^{14}$, $s = 0.05$, $U = 10^{-5}$.

In PNG with different initial conditions, it is possible to get one of the other universal distributions. In the droplet geometry, the initial condition is a single nucleation site, outside of which no other nucleations are allowed to occur. The initial site spreads out, and on top of the first layer more nucleations are allowed. This results in a curved profile, and the height distribution around this curve is a Tracy-Widom distribution from the Gaussian unitary ensemble.

In the evolutionary model, it is also possible to create the droplet formation. This could correspond to a mutation that raises the mutation rate significantly (a mutator strain), and competes with a population that has essentially no mutations. The shape of the resulting profile is not spherical like in PNG. This presents a difficulty because to find the distribution the unknown shape of the curve must be subtracted to find the distribution. The alternative is to look at only at the origin, which has no curvature, but then the statistics are too few.

2.7 Discussion

The concept of effective population size has long been useful in population genetics in many contexts, as a quantity that may be inferred from an idealized model. When considering

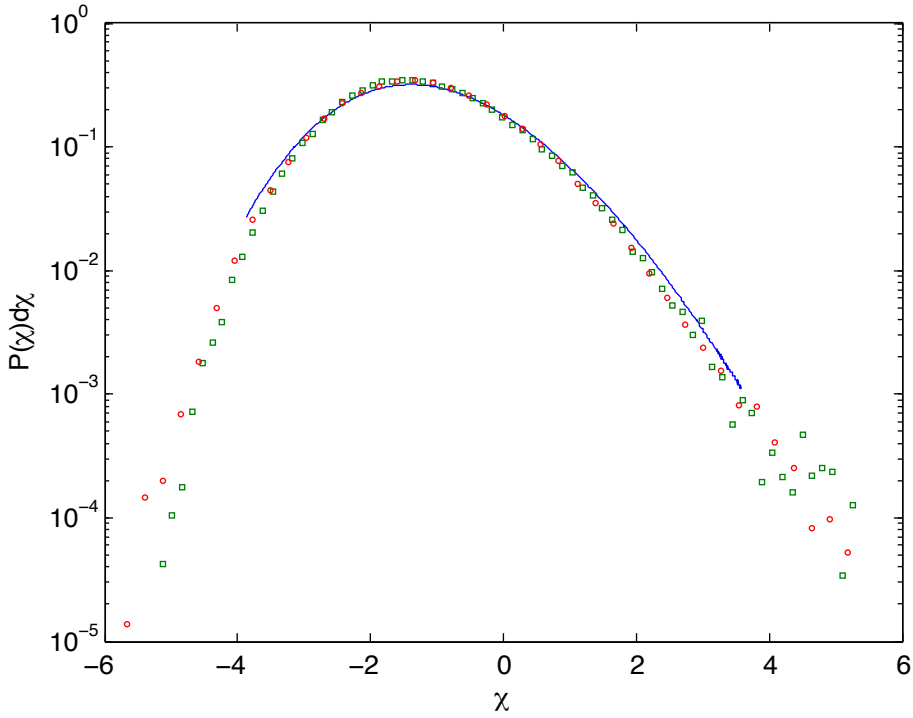

Figure 2.6: (blue line) Tracy-Widom GOE distribution. (red circles) scaled fitness distribution χ_{sim} for constant s (red circles) and exponentially distributed s (green squares) after 10^6 time steps averaged 100 times, $L = 2^{18}$. $U = 10^{-5}$, $\langle s \rangle = 0.05$.

the effective population size with spatial structure one is faced with two natural choices: the total population size, and a population size per length (or area). Our results indicate that the right answer depends on the situation. If the system is small enough, individuals have time to compete with everyone, and the system is effectively well-mixed. Above L_c, the rate of evolution does not depend on the total population, and it is appropriate to consider only a neighborhood population: L_c times the density. It is true that $v \sim UN_{\text{eff}}$, but it is misleading, because the effective population size depends on U itself, and in the end $v \sim U^{1/2}$, and does not depend on the total supply of mutations.

Fisher's fundamental theorem states that the speed of evolution is equal to the variance of the fitness distribution (appendix A). In the spatial model, there is a speed limit for large system sizes, while the variance grows linearly with L, and $v \neq \sigma^2$. It may seem as though Fisher's fundamental theorem is violated. However, in this case it makes more sense to consider the local population rather than the total population. The variance of the fitness distribution for the local neighborhood of size L_c, should not change with L, and Fisher's theorem still holds.

Martens and Hallatschek [27] generalized the critical system size to arbitrary wave speed c and dimension d

$$L_c = \left(\frac{c}{2sU}\right)^{\frac{1}{d+1}} . \tag{2.11}$$

Using estimates of typical mutation rates, bacterial densities and motility, they estimated that this length scale for a two dimensional habitat is on the order of 4 to 800 cell lengths. Thus competition may be ubiquitous in microbial colonies, and may be tested experimentally by varying the system sizes. The wave speed is related to the migration rate as $c \sim \sqrt{m}$ or $c \sim m$. Therefore in the long term populations with higher migration rates have less mutation competition and should be able to adapt more quickly to new environments.

The model presented here has the scaling exponents and universal distribution that belong to the KPZ universality class. This is not so surprising given the similarity to the PNG model, which is known to be in KPZ. Universality implies that the model is robust, because many of the details, such as the wave speed, and the distribution of s, do not change the scaling behavior and the fitness distribution. Unfortunately, these quantities appear only

asymptotically for long times, and since it is difficult to observe them in simulation, it is much more difficult to observe them in nature, where other forces may interfere.

In higher dimensions it is known that in the KPZ class the scaling exponents are different, and have been estimated only numerically. The height distributions are also different. Their exact form is unknown, but they are also skewed.

Chapter 3

Inference of a genotype-phenotype map

3.1 Introduction

The relationships between genotype, phenotype, fitness, and environment, to a large extent determine the evolutionary fate of a population. These relationships often involve complex and collective effects [106], which are difficult to untangle. One approach is to measure the fitness of many different genotypes, and build a *fitness landscape*, a high dimensional map from genotype or phenotype to reproductive fitness. This concept was first introduced by Sewell Wright in 1932 [107]. Evolutionary dynamics and adaptation depend crucially on features of the fitness landscape, and many studies have quantified large scale features of landscapes, including genetic interactions [49, 43, 53, 44, 51, 108, 109, 50], the presence of stabilizing selection [110, 111], or the reproducibility of evolutionary paths [52, 51]. Unfortunately, the large dimensionality of genotype space has made fitness landscapes difficult to measure, and only a few large landscapes have been characterized [54, 55].

A major difficulty that has precluded mapping of large fitness landscapes, either directly or through phenotypes, is *epistasis*, which is the dependence of fitness effects of a mutation on the presence of other mutations. Epistasis makes the inference of landscapes combinatorially complex. This problem has attracted substantial attention. For example, millions

of interactions between gene pairs have been measured from genetic knockout experiments [28, 29, 30, 31, 32, 33]. Higher order epistatic interactions, that is those involving more than two loci at a time, have also been investigated for small fitness landscapes [49]. Unfortunately, these pioneering studies have not yet provided the full picture of genotype to phenotype or to fitness mapping for genetic sequences of larger lengths, and most such large maps are modeled without epistasis (see, e. g., [112]). Indeed, a complete landscape would be defined not by genes or specific loci, but by all possible nucleotide sequences. However with $\sim 4^L$ different sequences of length L, it had been impractical to measure the landscapes for sequences of relatively large length until next generation sequencing technologies dramatically lowered the cost [113]. Nonetheless, measuring phenotypes of a large number of sequences is still tricky, and only a few large fitness landscapes have been quantified. For example, Pitt et al. measured the fitness landscape of $\sim 10^7$ RNA sequences with an in vitro selection protocol [54]. Similarly, Mora et al. studied frequencies of genetic sequences of IgM molecules in zebrafish B cells (which are related to fitnesses), but they imposed a translational symmetry of the sequence [114]. Finally, Hinkley et al. analyzed 70,000 HIV sequences and their *in vitro* fitnesses, built a fitness landscape defined on different amino acids of certain HIV genes, and then investigated large scale properties of the ensuing fitness landscape [55, 115]. However, even in these high throughput studies, the data did not contain all possible pairs of mutations, potentially biasing the results, especially far from the wild type sequences (see Discussion).

In this article, we start the process of reconstructing a large, yet detailed bacterial genotype to phenotype map, including quantifying the epistatic interactions in the ensuing fitness landscape. We are seeking a complete landscape based on long contiguous nucleotide sequences, which additionally allows quantifying phenotypes of transcriptional regulation in addition to those of enzymatic activity. This permits fitnesses to be defined over both coding and non-coding DNA. To map the landscape far from the wild type genotype, we would like sampling of the sequence data that is unbiased by selection.

Recent experiments by Kinney et al. [116] have collected a dataset that comes close to satisfying these criteria. The data consists of mutagenized transcriptional regulatory sequences from the *E. coli* (MG1655 and TK310 strains) *lac* promoter. In total $\sim 200,000$

lac promoter sequences were mutagenized in a 75 nucleotide region containing the cAMP receptor protein (CRP) and RNA polymerase (RNAP) binding sites (-75:-1). The transcriptional activity induced by the mutagenized promoters was measured through fluorescence of the transcribed gene products and FACS sorted according to the transcriptional activity into up to nine logarithmically spaced categories. All categories were then independently sequenced, so that the quantitative (on the scale of 1 to 9) phenotypic effect of each sequence is known to within a certain accuracy. Thus the data can be used to reconstruct the genotype-to-phenotype map. However, the promotor activity is directly related to lactose metabolism and thus is correlated with growth rate or fitness under conditions where lactose is the preferred energy source. Therefore, the fluorescence can also be viewed as a proxy for fitness of this sequence. Having noted this, we will not be making this distinction again in much of what follows, and will be using *fitness landscape* and *genotype to phenotype map* interchangibly when this causes no confusion.

In summary, the Kinney et al. [116] dataset provides simultaneous measurements of sequences and their phenotype (approximate fitnesses). Crucially, the data set is dense, so that every pair of mutations has occurred at least 20 times, each time in a different genetic backgrounds of about 5 other random mutations. We use these sequence and transcriptional activity data to infer the detailed genetic landscape for the 75 nucleotide DNA sequence, quantifying pairwise epistatic interactions among all of the nucleotides to the accuracy afforded by the data. This is done by constructing a linear-nonlinear regression model that connects sequences to their phenotypes. Since the number of possible epistatic interactions is comparable with the number of sampled sequences, we control the complexity of the models by L_1 regularization, to avoid overfitting. L_1 regularizaton also imposes sparsity on the inferred epistatic interactions, which is reasonable considering the limited number of known binding sites. We then analyze the statistics of epistatic effects in the inferred landscape. Finally, analysis of the landscapes obtained under different environmental conditions provides evidence that the wild-type sequence of the *E. coli lac* promoter is close to optimal in the ecological niche that the bacterium occupies.

3.2 Results

3.2.1 Inferring the non-epistatic genotype to phenotype map

The simplest model of a genotype to phenotype map is one where each locus contributes a fixed amount to the phenotype, regardless of the state of other loci. We fit this additive map using linear regression of the fluorescence values y (integers 1 to 9) on the predictors x_i, which encode the presence of mutations ($x_i = 1$ when a mutation is present, and $x_i = 0$ otherwise). Since there are four nucleic acids, each locus has three binary numbers for each of the possible mutations from the wild-type, and the sequence length is effectively tripled. In other words, for each locus, 000 represents the wild-type, and 001, 010, 100 represent the three mutations. The simplest statistical model is then

$$y = \beta_0 + \sum_{j=1}^{3L} \beta_j x_j + \varepsilon, \tag{3.1}$$

where ε is the statistical noise. To make inferences on the largest dataset possible, we combined the data from three experiments done by Kinney et al. [116] (fullwt, crpwt, rnapwt, 129,000 sequences total), which differ only by the regions in which mutations were allowed to take place. Fullwt was mutagenized over the whole sequence (-75:-1), while crpwt and rnapwt were mutagenized only over the CRP binding area and RNAP binding area. We then find β's by ordinary least squares regression, e.g. coefficients that minimize $\langle \varepsilon^2 \rangle$ in Eq. (3.1).

Since the wild-type is a sequence of all zeros, β_0 is the predicted phenotype of the wild type. The coefficient $r^2 = 1 - \sigma_\varepsilon^2/\sigma_y^2$ measures the goodness of fit, or how much of the variance in the data, σ_y^2, is explained by the model. Some variation in the data is experimental noise, such as background fluorescence and cell-to-cell variability, which sets an upper bound on r^2. We estimate this intrinsic noise by averaging the variance of y for identical sequences with different recorded fluorescence values. The ratio of this intrinsic variance to the total variance of y is $2.3/6.5 = 0.35$ (recall that y ranges between 1 and 9), so only about 65% of the total variability of the data can be explained by any statistical model, even an arbitrarily complex model. The linear model yields $r^2 = 0.464 \pm 0.002$,

which is about 70% of the explainable variance.

Part of the genotype-phenotype map may be non-linear (on average) due to the first five bins being dominated by the background fluorescence. While epistasis is sometimes defined as a non-linearity of the *mean* fitness as a function of the number of mutations, here were are interested in interactions between loci. To identify these interactions in the background on an arbitrary mean nonlinear fitness, we introduce a generalized model:

$$f(y) = \beta_0 + \sum_j \beta_j x_j + \varepsilon, \tag{3.2}$$

where $f(y)$ is a monotonically increasing, nonlinear function of y. The function is found by maximizing the fit (r^2), which corresponds to minimizing[1]

$$f(y) = \arg\min_{g(y)} \frac{\mathrm{var}\left(g(y) - \beta_0 - \sum_j \beta_j x_j\right)}{\mathrm{var}\left(g(y)\right)}. \tag{3.3}$$

We add the constraints that $f(9) = 9$, and $f(1) = 1$ to keep $\mathrm{var}\left(g(y)\right)$ finite. The function $g(y)$ is defined over only 9 values of y, and a constrained non-linear optimization procedure (`fmincon` from MATLAB) finds an optimal $f(y)$ quickly (Fig. 3.1). While this improves the linear fit, $r^2 = 0.501 \pm 0.002$, statistical tests suggest that the linear-nonlinear model is imperfect, indicating the need for inclusion of epistatic interactions. Since the nonlinear mapping de-emphasizes noisy bins with small y, the intrinsic noise estimated with $f(y)$ is lower, $1.8/7.6 = 0.24$. Thus the linear model accounts for about 66% of the explainable variance.

Examination of the coefficients β_j with the largest magnitude reveals the consensus locations of the CRP and RNAP binding sites (Fig. 3.2), which validates the modeling approach. Interestingly, the wild type does not contain the "consensus" binding sequences: $\mathrm{TGTGA(N)_6TCACA}$ for CRP [119] and $\mathrm{TTGACA(N)_{18}TATAAT}$ for RNAP [120], but the wild type is only four mutations away. Four of the large positive coefficients in Fig. 3.2 (positions -54, -34, -9, -8, marked red) correspond to the mutations needed to get the

[1]This method resembles a type of generalized linear model called ordinal probit regression[117], and is also similar to the inference of non-linear filters in computational neuroscience using information-theoretic tools [118].

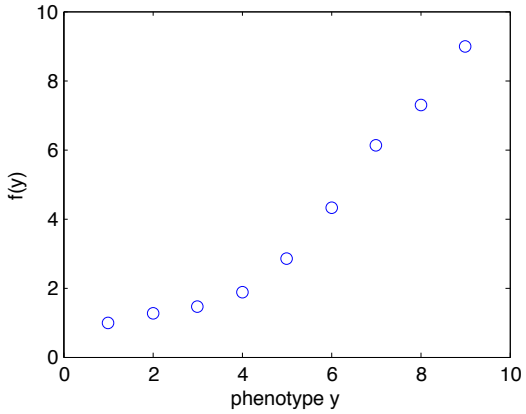

Figure 3.1: Generalizing the fitted function by replacing the output values y with a non-linear function $f(y)$ improves the least squares fit. Constrained non-linear optimization found the optimal $f(y)$ for the linear model with $r^2_{opt} = 0.501 \pm 0.002$. The non-linearity is due to the first few bins being dominated by background fluorescence and not gene expression.

consensus sequences.

3.2.2 Inferring epistatic contributions to fitness

The simplest model with epistatic interactions between all pairs of nucleotides can be written as:

$$f(y) = \beta_0 + \sum_j \beta_j x_j + \sum_{i<j} \beta_{ij} x_i x_j + \varepsilon. \tag{3.4}$$

The last sum is over all nucleotide pairs. Note that we keep $f(y)$ the same as in the previous section, which maximizes the maximum explanatory power of the nonepistatic terms and minimizes that for the epistatic terms. The number of epistatic terms in this statistical model ($\sim L^2$) should be contrasted with typical biophysical models of protein-DNA interactions, which include only a single free energy term describing interactions between the CRP and RNAP proteins [121, 116].

The total number of coefficients β_0, β_i, and β_{ij} in the quadratic epistasis model, Eq. (3.4), is 25,201, thus the linear system is too large to be solved exactly. A standard iterative algorithm [122] converges to a solution with $r^2 = 0.79$. However, overfitting is a concern since the number of observations, 129,000, is not much larger than the number of coefficients. To prevent overfitting, we minimize the residuals variance in Eq. (3.4) subject to a regularizing

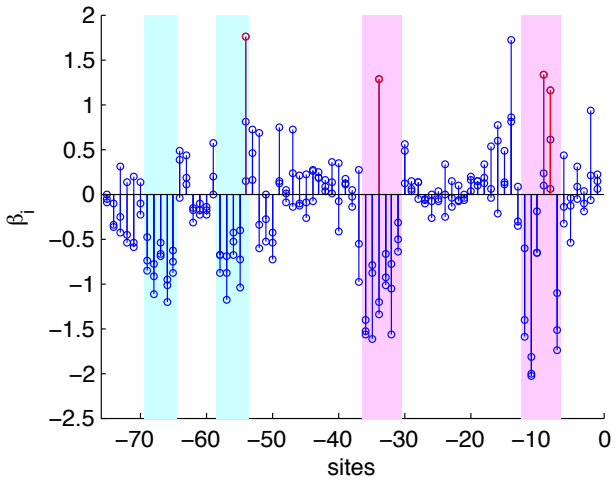

Figure 3.2: Stem plot of the linear coefficients. Three circles on each stem represent the changes in fitness for each of the three possible mutations per site. The cyan and the magenta areas correspond to the consensus bindings site locations of CRP and RNAP.

constraint

$$\beta^* = \arg\min_{\beta} \left(\langle \varepsilon^2 \rangle + \lambda \|\beta\| \right), \tag{3.5}$$

where β is the concatenated vector of all the regression coefficients, $||\beta||$ is its norm, and λ is a free parameter (Lagrange multiplier). Regularization constrains the statistical complexity of the model by minimizing the norm of the coefficients [123]. When the L_1 norm is used, $\|\beta\| = \sum |\beta_i|$, this regression is called the Least Absolute Shrinkage and Selection Operator (LASSO) [124]. LASSO favors sparse solutions, which here is a reasonable assumption since most of the β's are interaction terms, and interactions are presumed to be mainly between the relatively small CRP and RNAP binding sequences. Thanks to an efficient implementation of the algorithm [125], we can compute the LASSO solution for 100 values of λ, from the maximum value (where the solution is all β's equal to zero), to four orders of magnitude smaller.

However, choosing the *best* solution is ambiguous. A common method of model selection is cross-validation. Figure 3.3 shows that solutions with large λ are a poor fit, while small λ values have less predictive power, as seen through cross-validation. Typically one chooses the best model as the one with the maximum r^2 (r^2_{CV}) [124]. However, both the training and the cross-validation data are sequences with an average of only 6.8 mutations from the wild-

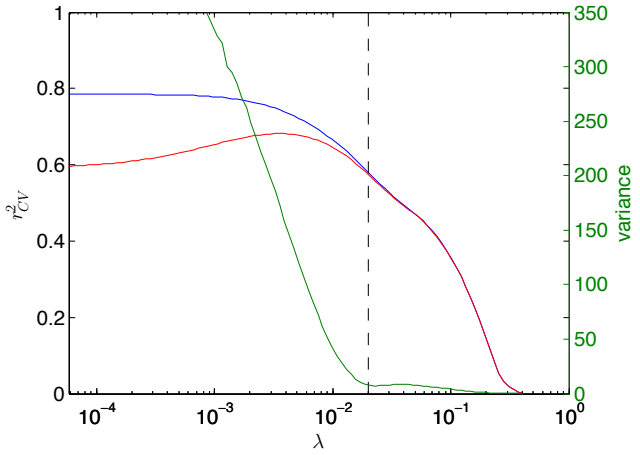

Figure 3.3: The LASSO solution of the quadratic model was computed for 100 values of λ. Blue is the r^2 value, and red is the 10-fold cross-validated r^2_{CV}. The green curve is the variance of $f(y)$ for randomly generated sequences. The variance is too large even for values of λ that are larger than the optimal value predicted by the maximum of the r^2_{CV} curve. We choose the model with $\lambda = 0.016$ (dashed line) for further analysis. This model has $\sim 10^3$ non-zero coefficients, most of which are epistatic.

type (9% mutated sites). Thus cross-validation may not ensure predictability for sequences farther away in the genotype space. Indeed, the variance of the fitted values of $f(y)$ for the experimental data is not sensitive to changes in λ (not shown). Nonetheless, Fig. 3.3 shows that the variance of $f(y)$ for random sequences blows up for less constrained models (low λ), where unrealistically high fitted values of y or $f \sim 50 \ldots 100$ emerge. This indicates overfitting due to uneven sampling of the genotype space and the resulting correlations in the training and the test data. We thus limit λ to the range where the variance of the fitted values for random sequences is comparable to that for the experimental data and is insensitive to λ. Incidentally, this is also the place where r^2 and r^2_{CV} curves split in Figure 3.3 (dashed line, $\lambda = 0.016$, $\sim 10^3$ non-zero coefficients). The results in the following Section are from this set of coefficients.

3.2.3 Properties of the inferred genomic landscape

The distribution of f values for randomly generated sequences (Fig. 3.4) shows that the random sequences are typically not very functional (presumably because the binding sites

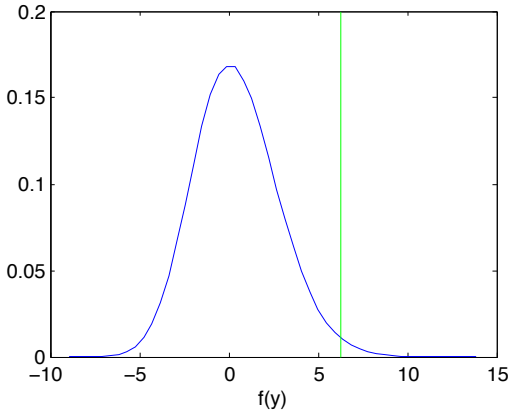

Figure 3.4: Histogram of phenotype $f(y)$ values of 10^5 uniformly random sequences for the $\lambda = 0.016$ model. Random sequences have very low phenotype values because of the specificity of binding sites. The peak of the distribution indicates what phenotype values evolve under neutral conditions. The the wild-type value, β_0 (green line), is much higher than the neutral value indicating selective pressure.

loose specificity). The peak near $f = 0$ represents the most common sequence that would be observed under neutral evolution, and the relatively high value for the wild-type ($f_{\mathrm{wt}} = 6.2$) compared to the random sequences indicates that it is under strong selection. Note that we can assert this without any comparative genomics or population genetics data, which would typically be required.

The fraction of variance explained by the pairwise epistatic model is $r^2_{\mathrm{CV}} = 0.569 \pm 0.007$ (although it is sensitive to λ, cf. Fig. 3.3). Comparing to the non-epistatic model with $r^2 = 0.501$, and taking into account the intrinsic experimental noise of $\sim 25\%$, we see that about 10% of the explainable variance is due to the pairwise epistasis. However, we have not yet reached the point where adding new data does not improve the model anymore. Thus the magnitude of the epistatic effects may become larger when more data allows us to explore smaller values of λ without overfitting.

For the chosen λ, the coefficients corresponding to the linear terms are about 70% non-zero, but the interaction terms are very sparse, with only about 3% non-zero. By simultaneously looking at magnitudes of all coefficients belonging to one nucleotide, we can find the nucleotides that affect the fitness the most. These turn out to be the nucleotides within the CRP and RNAP binding sites (see Fig. 3.5). Thus this kind of data allows for identification of binding sites without a biophysical model of protein-DNA interactions,

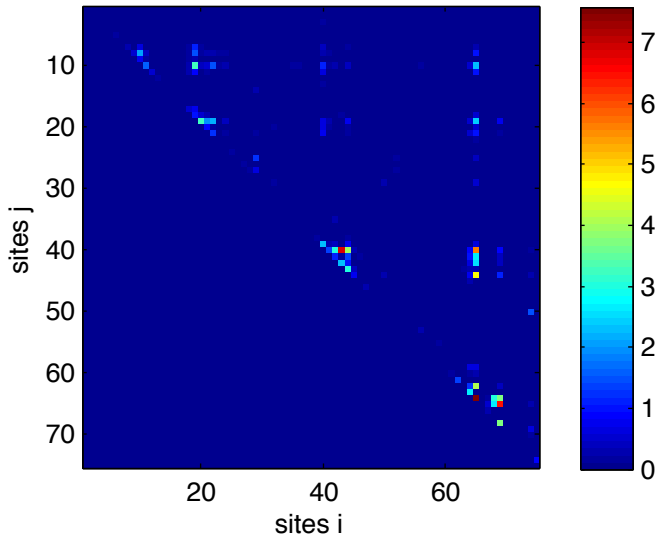

Figure 3.5: Matrix of the sum of the absolute values of the pair interaction coefficients for each pair of sites i, j (3 mutations per site equals 9 interactions). The clusters near the diagonal are interactions within the RNAP and CRP binding sites, and the off-diagonal clusters are interactions between the binding sites.

as is done traditionally [126, 127]. More importantly, as Fig. 3.5 shows, the model can infer functional interactions between amino acid / nucleic acid binding over a much longer range than can be computed from biophysical and structural biology approaches [128]. This validates our approach to regularize the model by LASSO regression. Alternative methods that instead limit the number of β's by constraining the range of interactions in Eq. (3.5) or by allowing interactions only between consensus sites would either miss the long-range effects, or the small (but statistically significant) interactions away from the binding sites seen in Fig. 3.5.

Epistatic interactions may be classified into several categories (see Table 3.1): synergistic epistasis (the effect of two same-sign mutations is larger than the sum of effect of each one separately), antagonistic epistasis (the effect of two same-sign mutations is smaller than the sum of their individual effects), and other epistatic effects (the individual effects of two mutations have opposite signs, while epistasis is present). We find that most of the interactions in the *E. coli lac* promoter are antagonistic (406/675=60%). We found only one case of a severe type of antagonistic epistasis (reciprocal sign epistasis), where the individual effects are both harmful, but the total effect is beneficial. It is known that reciprocal sign

type	β_i	β_j	β_{ij}	% of non-zero β_{ij}
synergistic	$\pm$	$\pm$	$\pm$	7%
antagonistic	$\pm$	$\pm$	$\mp$	60%
other	$\pm$	$\mp$	$\pm$ or $\mp$	33%

Table 3.1: Epistasis may be categorized as: (synergistic) the effect of two mutations is greater than the sum of the two mutations individually, (antagonistic) the total effect is less than the sum, and (other). In our data, the epistasis is mostly antagonistic.

epistasis is a necessary (but insufficient) condition for a multi-peaked landscape [129], and hence we expect this landscape to be fairly smooth (at most two maxima).

While there are many ways to measure the roughness of the landscape [49], we verified this by directly exploring the accessibility of the local fitness peaks of the inferred landscape. We used a greedy random walk similar to the evolution of a large monomorphic population, which can move towards higher fitness and cannot escape local maxima. Starting from the wild-type sequence, the algorithm only chooses mutations which are higher in fitness, with probability proportional to the fitness difference. Out of 1000 random walks, the population ends up in only two very similar sequences which differ by 2 mutations, and they are 40 and 39 mutations away from the wild type (compare to the average of ~ 6.7 mutations per sequence). Since the sequences are so far away from the training data, their predicted fitness value are likely not reliable.

3.2.4 Landscape in two environments

In addition to the data from the three experiments analyzed above, Kinney et al. [116] performed experiments with a different strain of bacteria (TK310) that is unable to control its intracellular cAMP levels. Because CRP is activated by cAMP, varying extracellular cAMP levels controls the active intracellular concentration of CRP. *E. coli* prefers to metabolize glucose over lactose, so cAMP is inhibited by the presence of glucose, and *lac* expression is suppressed when glucose is present. We inferred fitness landscapes using the non-epistatic model as in the Section 2.1 for two conditions, no cAMP and $500\mu M$ cAMP, representing an environment with glucose and no glucose. The datasets are smaller ($\sim 25,000$ sequences), and distinguish only 5 levels of fluorescence, but they are otherwise very similar, so the same linear-nonlinear r^2 optimization was used. The following results were found with the

non-epistatic model, but because the pair interactions account for a smaller fraction of the variation, the epistatic model produces very similar fitted values.

As expected, when CRP is not active there is little binding at the CRP sites, and the associated coefficients are almost all small (Fig. 3.6). Because of the lack of CRP binding, expression for the wild type sequence, and sequences close to the wild-type, is lower when there is glucose (Fig. 3.7). However, there are some changes to the RNAP binding site coefficients. Random sequences are not functional in the no-glucose environment, but they have some low functionality, comparable to the wild-type, in the glucose environment (Fig. 3.7), suggesting that there is less specificity in the RNAP binding.

In the no cAMP (glucose) environment, *lac* expression should decrease the growth rate because the cell is metabolizing glucose instead of lactose, and *lac* expression costs resources [130, 131]. Therefore we expect sequences under selection, such as the wild type, to have relatively high expression with cAMP, and low expression without cAMP, compared to sequences not under selection (random sequences). Figure 3.7 shows that there exist very few sequences which are better than the wild type in both environments, i.e. simultaneously higher expression with cAMP, and lower expression without cAMP. The non-elliptical shape of the fitted values for the experimental sequences suggests again that the wild type is under a strong selection towards the top left corner of the plot. Finally, we point out that, even when lactose is being metabolized, too high expression of *lac* genes is costly, possibly because cellular resources are pulled to *lac* transcription and translation and away from production of essential proteins [130]. This may make sequences in the top right corner of Fig. 3.7 less fit than our $f(y)$ model assumes, making the wild type even closer to optimality.

3.3 Discussion

We constructed a genotype-to-phenotype mapping, including effects of all pairwise epistatic interactions, by analyzing functional properties of $\sim 200,000$ randomly mutated sequences in the vicinity of the wild type *E. coli lac* operon. Our approach is generally similar to those in Refs. [114, 55]. However, there are substantial differences. Our alleles are nucleotides in a regulatory region of a bacteria, instead of amino acid variants. Our landscape is more

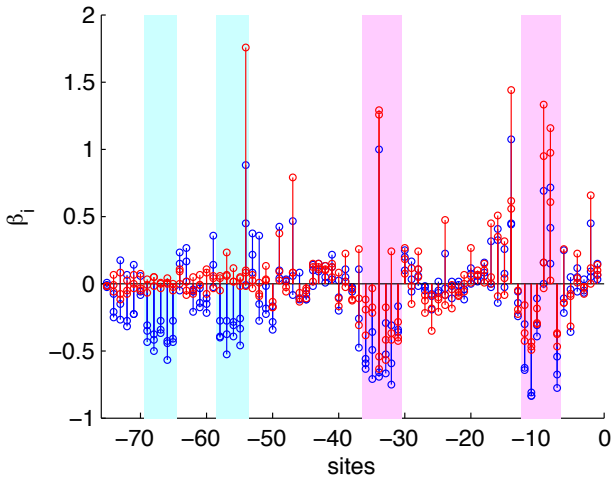

Figure 3.6: (blue) coefficients β_i for the non-epistatic model with no-glucose (normal levels of cAMP) (red) with glucose (no cAMP). CRP is activated by cAMP and does not bind without it.

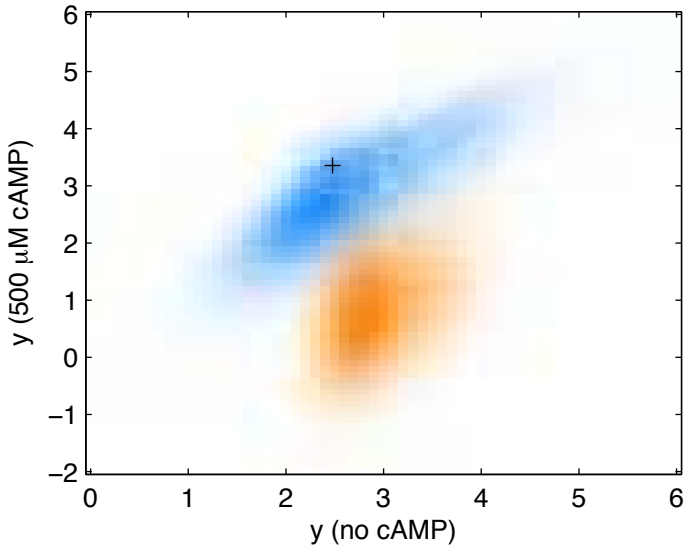

Figure 3.7: 2D histogram of expression for the two environments, no cAMP (glucose), and cAMP (no glucose) for 10^5 random sequences (orange), and sequences from the experiment (blue), which are closer to the wild type (plus sign). The wild-type is nearly Pareto optimal in that very few sequences have both higher expression with cAMP and lower expression without cAMP (above and to the left of the plus sign). The phenotype values range from 1 to 5 in these experiments.

complete, in that interaction among all pairs of nucleotides in the sequence are estimated from the data that includes each such pair at least 20 times in different genetic backgrounds. In particular, we have relaxed the condition [114] that the interaction terms β_{ij} can depend only on the distance between the loci, rather than on the specific positions of the loci. Mora et al. [114] have used maximum entropy approaches to infer the fitness landscape, while, along with Hinkley at al. [55], we have focused on linear regression inference (though with different regularization constraints and different nonlinear mapping between the fitness and the observed phenotype). The quadratic model, Eq. (3.4), is the same in the regression and the maximum entropy approach. However, the philosophical basis behind the approaches is different, and so are the criteria used to specify the coefficients β. Maximum entropy methods choose them to constrain observable correlation functions, while regression attempts to approximate the entire fitness function. It remains to be seen which of the two frameworks provides a better model for genomic data.

Possibly the largest difference from the previous approaches that considered epistatic interactions for many mutations is that we found a genotype-phenotype map, rather than the true fitness landscape. While we expect the phenotype and the fitness to be strongly correlated when lactose is being metabolized (and anti-correlated otherwise), the relation between the fitness and either the observed fluorescence or its nonlinearly reparameterized form, $f(y)$, is likely nontrivial. Ideally, a second experiment would measure the phenotype-to-fitness map to complete the reconstruction of the fitness landscape. In fact, Dekel and Alon[130] have completed this second step for the *lac* regulatory sequence. However, we cannot use their findings to complete the reconstruction since their *E. coli* strains and growth environments were slightly different from those of Kinney et al. [116], which provided us with the fitness data.

Our observations have revealed a few cautionary notes regarding using genome frequency in a population as a measure of fitness [114, 55]. In such experiments, all sequence data (including whatever part of it that is left for cross-validation) are localized near the wild type, near-optimal sequences due to selection. Carefully inferred models (whether regression or maximum entropy based) perform well for the observed data, but will generalize badly for sequences far away from the wild type. Our approach samples the genotype space

more evenly without selection, and therefore is better suited for making inferences about the global landscape properties, such as its ruggedness. Nonetheless, even in our data, with each sequence ~ 7 mutations away from the wild type, extrapolation to much larger genotypic differences produces absurd results, even if cross-validation fails to notice problems, cf. Fig. 3.3.

In our inferred landscape, epistasis accounted for about 10% of the explainable variance. Most of the epistasis was antagonistic, but the landscape was essentially single peaked. This is in contrast to the work on HIV fitness landscape [115], which has observed more substantial epistasis and many more local maxima. It is possible that more epistatic components would be observed for *E. coli* as well if more data were available, and it was possible to use a smaller λ without overfitting. However, more intriguing is the observation that, at $\lambda \approx 4 \cdot 10^{-3}$, which maximizes r_{CV} in Fig. 3.3, our inferred landscape is also extremely epistatic. Yet it is clearly wrong and overfits, even if cross-validation does not notice this due to the sampling bias towards the wild type. We thus conclude that inferences about the global roughness properties of fitness landscapes from sequence data may be misleading. The severity of the problem correlates with the nonuniformity of the genotype sampling, making the data from populations under strong selection especially suspect. To allow studying global properties of landscapes, an ideal experiment would sample the sequence space much more uniformly.

In addition to the weak epistasis, we also found that the wild-type *E. coli lac* regulatory region is optimal for the two environments measured. That is, it is on the front of possible sequences which maximize expression when it is beneficial, and minimize expression when it is harmful. If under the growth conditions the fitness is a nonmonotonic function of the transcriptional activity and decreases at high expression [130], the wild type operon may be closer to the optimal front. To investigate this, experiments are needed that would study fitnesses of many sequences under selection in fluctuating environments.

For over ten years, genome wide studies of transcription factor binding sites have been using genomic statistics and population genetics models to infer phenotype-fitness landscapes with binding energy as the phenotype [132, 133, 134, 135]. Our method allows similar findings, but is in some sense simpler because we use no population genetics. It

is also complementary to traditional methods because we infer a genotype-phenotype map and found the non-averaged type of epistasis. The difference is that previous studies assumed that binding energies were additive, and found fitness as a non-linear function of binding energy, while our method tries to infer a non-linear map between sequence and transcription, without making any strong assertions about the phenotype-fitness map. The connections between the two approaches need to be developed further. Of special interest is understanding the effective dimensionality of the epistatic interactions: can the epistasis in the statistical model be explained by a single parameter describing binding between RNAP and cAMP, or are additional biophysical mechanisms important as well?

The ability of our method to identify protein binding sites and epistatic interactions among them raises an important point. These epistatic interactions, inferred by either of the methods we have mentioned in this work, especially interactions over long ranges, may not correspond to true biophysical interactions between amino acids and nucleotides, but are likely *effective* interactions resulting from collective effects of many other epistatic terms, including higher order terms. Our approach may not be as useful where there is enough information to build a detailed biophysical model, but there are few places in the genome where this is the case. Our approach can detect long distance epistasis where a priori it is unclear that interactions exist. When working on the genome scale, effective models that can make accurate *predictions* of phenotype or fitness for previously unobserved sequences may be useful regardless of their lack of microscopic accuracy. They may be closer to the right level of description of the problem [136], by striking a balance between microscopic biophysically relevant detail, and power to describe the richness of phenomena emerging on the genomic scale.

Chapter 4

Speeding up evolutionary search by small fitness fluctuations

4.1 Introduction[1]

Organisms adapt to their environment by sequential fixation of beneficial mutations. This process is often visualized as motion of a population, specified by multi-dimensional genomic variables corresponding to the dominant genotype in the population, in the fitness landscape, where the height of the landscape corresponds to the reproductive fitness of an average individual in the population [107]. Fitness landscapes are believed to be rough with many local maxima, and a population may get stuck in one, so that every plausible mutation is deleterious [137, 138, 52, 73]. In such cases adaptation to a global fitness maximum requires the rare fixation of deleterious mutations. Even when there is a path of only neutral or weakly selective mutations to the global optimum, it may be difficult to find it, and navigating such paths can be slow due to the low fixation probability ($\approx 1/N$ for a neutral mutation in a population of N individuals [95]). In view of this, one of the central questions of evolutionary biology is how the current diversity of the living world has emerged in a short few billion years since the life on Earth has started, especially since thousands of generations of laboratory evolutionary experiments lead to only a few dozen

[1]This work was published as: Otwinowski, J., Tanase-Nicola, S., & Nemenman, I. (2011). Speeding up Evolutionary Search by Small Fitness Fluctuations. Journal of Statistical Physics

fixated mutations [3].

It has been recognized that temporal fluctuations in the fitness landscape can drive the population out of a local fitness maximum, thereby accelerating evolutionary processes. For example, the maximum of fitness at one time may be on a fitness slope at another time, allowing the population to leave the area. The effects of fluctuating selective pressure on mutation accumulation dynamics have been studied extensively [42, 66], starting with the introduction of the concept of adaptive topography by Wright [107]. More recently the evolutionary dynamics of density regulated populations in fluctuating environments have been elucidated in more ecologically realistic models [60, 61, 62], bridging the gap between the classical population dynamics [63, 64] and population genetics models.

In a recent pioneering numerical evolution experiment [69], these ideas were further developed to show that certain types of fluctuating environmental pressures may speed up evolution many times. However, it remains unknown to what extent these results generalize. Is the speedup a general property? How does it depend on the spatiotemporal structure of the fluctuating environment? Can a population escape any local maximum? How does the motion in the genotype space depend on time? Do fitness fluctuations have to be dramatic, as in Ref. [69], or can small fluctuations still speed the evolution up?

In this article, we answer some of these questions in the context of a model of evolutionary dynamics that is simple enough to allow a thorough analytical and numerical treatment, but is at the same time general enough so that at least some of our predictions hold for a wide class of evolutionary models. We consider the limit of a weak mutation rate, when the time scales are well separated. The time between successive mutations is longer that the typical fixation time, and the characteristic time scale of the fitness landscape changes is the longest. Such a situation is relevant for microbial populations under seasonal environmental changes, or for host-pathogen interactions, where environmental changes may correspond to phases of transmission, unhindered growth in a new host, and activation of the host immune response. Further, we assume that the total population size is independent of the genotype (though not necessarily fixed), so that the evolutionary dynamics depends only on the relative fitness differences between the genotypes. We consider adaptation in a highly epistatic genotypic space, such that the evolution takes place on a one dimensional

pathway with large, local fitness differences. Under these assumptions, we show that the evolutionary search can be sped up substantially when only a small component of the fitness landscape undergoes temporal variations.

4.2 The model

Our model of a fluctuating environment is based on an overdamped particle in a potential. The position x is some generalized coordinate that describes the dominant genotype in the population, and hence a change in x is a fixation event. This genotype changes according to an equivalent Langevin dynamics given by

$$\frac{dx}{dt} = -\frac{1}{\gamma}\frac{\partial U(x,t)}{\partial x} + \eta, \tag{4.1}$$

where U is the potential, γ is the "friction"-like scale factor, η is a white Gaussian noise with variance $2D$, where D is the intrinsic diffusivity, presumably related in the population biology context to the population size [42]. Notice that in the usual physics language, the process will minimize the potential so that U is the negative fitness. Motion to minimize U represents fixation of beneficial mutations, while Langevin noise allows low probability fixation of neutral and deleterious mutations. The first phenomenon is called natural selection/drift in evolutionary/physics languages, and the second is unfortunately referred to as drift/diffusion, respectively. To avoid confusion, in the remainder of the article we use the physics terminology.

We write the potential as

$$U(x,t) = U_0(x) + \Phi(x)S(t), \tag{4.2}$$

and we focus on the following range of parameters:

$$\operatorname{var} S(t) \sim 1, \tag{4.3}$$

$$\max[\Phi(x)] - \min[\Phi(x)] \ll \max[U_0(x)] - \min[U_0(x)] \tag{4.4}$$

This models the emergence of novel functions in a population. Namely, the fitness is largely independent of time, as described by U_0. However, small temporal changes in fitness are allowed. For example, acquiring a new enzyme is generically advantageous if its substrate is present, but deleterious if it is absent due to generic costs associated with protein overproduction [139, 140]. We model this by adding a small fitness component $\Phi(x)$ that fluctuates as $S(t)$, representing, for example, changes in the availability of the metabolite due to seasonal or geological variations. Finally, we choose to separate the global, almost non-epistatic, fitness from the local, possibly highly-epistatic (but small) effects by making the gradient of U_0 smaller than that of Φ, even though the scale of Φ itself is smaller than that of U_0. For example, this agrees with the observation that the ability of proteins to bind to DNA or to metabolic substrates is highly sensitive to the details of the protein sequence [137, 138, 52, 73].

With the conditions above, we can redefine U_0, Φ, and S without much loss of generality, so that $\langle S \rangle_t = 0$. We then consider the simplest form of $\Phi(x)$ and $S(t)$ that satisfies these conditions, and we will discuss how our results generalize to some other forms of the functions in Discussion. Namely, we choose Φ to be a zero-mean periodic saw-tooth potential, and S to be a zero-mean periodic telegraph signal. These considerations allow us to write near a particular point x in the genotype space

$$\frac{1}{\gamma}\frac{\partial U}{\partial x} = -v + \phi(x)s(t), \tag{4.5}$$

$$\phi(x) \equiv \frac{h}{L}\ \mathrm{sign}\left[\sin\frac{\pi x}{L}\right], \tag{4.6}$$

$$s(t) \equiv \mathrm{sign}\left[\sin\frac{\pi t}{T}\right], \tag{4.7}$$

where v is the intrinsic drift (in physics terms) or bias, defined as positive for the drift to the right, see Fig. 4.1. We always assume that $|h|/L > |v|$, so that the fluctuating component of the potential can actually create local maxima and minima on top of the global landscape $U_0(x)$. In what follows, we denote by T the time between subsequent potential flips (the half-period of the fluctuations), and L is half of the spatial period.

This model is similar to various stochastic ratchets considered in the literature [141,

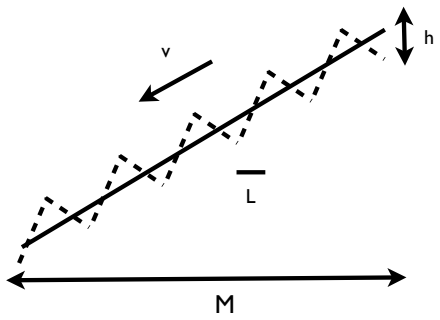

Figure 4.1: The potential $U(x,t)$ at a fixed time. An oscillatory, symmetric, sawtooth perturbation is added on top of the average linear potential that creates a drift velocity of v.

142, 143, 144]. Thus, the question of whether the fitness fluctuations can speed up the evolutionary search is a question similar to whether a rectified or a high-variance motion can appear due to ratcheting. We know from prior analysis [145] that any unbiased spatially variable but temporally constant potential cannot give rise to rectified motion, and it will always slow down diffusion. Hence temporal fluctuations are an essential component of the model.

4.2.1 Rescaling of the equation of motion

Using the choices above, we can rewrite the equation of motion, Eq. (4.1) as

$$\frac{dx}{dt} = [-\phi(x)s(t) + v] + \sqrt{2D}\eta. \tag{4.8}$$

where η is a Gaussian white noise of unit variance. In Eq. (4.8), the dynamics explicitly depends on five different parameters L, T, v, h, and D. Nevertheless, by rescaling the time, the space, and the potential as $x/L \to x$, $t/T \to t$, $\frac{L}{h}\phi \to \phi$, $\frac{L}{h}v \to v$, we can reduce the number of parameters to only three: the ratio of the typical diffusion time over half the spatial period to half of the temporal period, $\omega = \frac{L^2}{2DT}$, the height of the fluctuating barriers in diffusivity (temperature) units, $\beta = \frac{h}{D}$, and the ratio between the slope of the average, large scale potential to the slope of the fluctuating perturbation, v. In physical terms, if

ω is large, the particle has time to explore the entire valley of ϕ before the potential flips. Further, β measures the difficulty of crossing the peaks by diffusion. Finally, the condition that the perturbation induces local optima is $|v| < 1$.

Using the rescaled variables, the dynamics becomes

$$\frac{dx}{dt} = \frac{\beta}{2\omega}\left[-\phi(x)s(t) + v\right] + \sqrt{\frac{1}{\omega}}\eta. \tag{4.9}$$

We will use these rescaled variables in the rest of the article, unless noted otherwise. From this equation, it is easy to recover the dynamics in the original, non-scaled units by simple multiplications. In what follows we present simulation results obtained using first order Euler integration scheme of the dynamics defined in rescaled variables, Eq. (4.9).

4.3 Fluctuating potentials enhances diffusion and drift

Numerical simulations suggest that the behavior of $x(t)$ at large times is diffusive, and anomalous scaling is not seen [146, 147, 148, 149]. This is consistent with general analytical results obtained by multi-scale techniques [145]. We can characterize the genotype coordinate motion by effective drift and diffusion constants, which depend on the spatial and the temporal periods of the fluctuations and the barrier height. To quantify the enhancement or the suppression of the motion at long spatial and temporal scales compared to the intrinsic diffusivity and drift, we define

$$r_D \equiv \frac{D_{\text{eff}}}{D} = \frac{\text{var}_t(x)}{2t}\frac{L^2}{DT} = \frac{\text{var}_t(x)}{t}\omega, \tag{4.10}$$

$$r_v \equiv \frac{v_{\text{eff}}}{v} = \frac{\langle x(t)\rangle}{t}\frac{L}{vT} = \frac{\langle x(t)\rangle}{t}\frac{2\omega}{v\beta}, \tag{4.11}$$

where the time-dependent mean and variance of the trajectories $x(t)$ are obtained numerically. As seen in the Fig. 4.2, both the effective drift and the effective diffusion can be enhanced with respect to the intrinsic values, this enhancement having a maximum for fluctuation periods comparable to the average time to travel between two extrema.

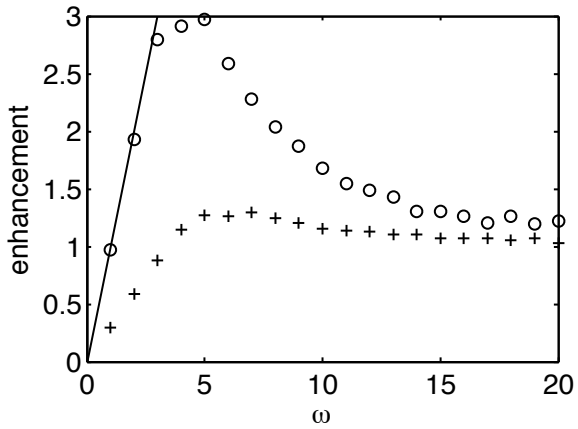

Figure 4.2: Enhancement of diffusion, $r_D = \frac{D_{\text{eff}}}{D}$ (circles) and drift, $r_v = \frac{v_{\text{eff}}}{v}$ (pluses) as a function of the relative flipping frequency $\omega = \frac{L^2}{2DT}$, with $\beta = 10$ and $v = 1/10$. At low ω, the particle has time to reach the minima, and $D_{\text{eff}} \approx \frac{L^2}{2T}$, or $r_D \approx \omega$. Simulations are averaged over 1000 trajectories and 1000 time steps. Solid line indicates $r_D = \omega$. Error bars are smaller than the symbols.

4.3.1 Building intuition

When $\beta \ll 1$, the sawtooth peaks are very small, and the diffusion has no trouble crossing them. When β is larger, the behavior is more interesting. For $\omega \to 0$, the particle has ample time to fall into a minimum of ϕ before $s(t)$ flips. When the potential flips, the particle can now go either left or right, with unequal but comparable probabilities, which creates a biased random walk behavior with the effective diffusion coefficient $\approx L^2/(2T)$, or, equivalently, $r_D \approx \omega$. The fluctuating potential allows the particle to diffuse against the drift, so that the speed of the evolutionary search is strongly enhanced when the environment oscillates.

For low ω, $r_D \approx \omega$ is also small, so the diffusion is suppressed compared to the internal value. However, for $h \gg 1$ and without the changing sign of $s(t)$, the particle would get stuck at a minimum of ϕ almost immediately, and the overall diffusion would be essentially zero. It is hard to exit a deep potential well only with the help of diffusion. Whether r_D is greater or less than 1, the fact that oscillations make it non-zero in the long term is our most important finding, suggesting that temporal fluctuations can make fixation of rare-to-fixate mutations a much more common process. Figure 4.2 demonstrates these findings for different values of ω and for $\beta = 10$ and $v = 1/10$.

When the flipping is fast compared to the diffusion time (large ω in the Figure), the

fluctuating potential averages out to zero. The only motion is due to the internal diffusion, and D and r_D go to one.

4.3.2 Analytical treatment at $\beta \to \infty$

In the limit $\beta \gg 1$, the fluctuating peaks are very high, the particle almost never crosses them due to noise, and analytical progress can be made. First consider a particle that starts close to a local minimum x_0 of the sawtooth. After some time $\sim D/v^2$, which goes to zero as $\beta \to \infty$, the particle equilibrates near x_0 with a probability density

$$p(x|x \gtrless x_0) \propto e^{-(1\mp v)|x|}. \tag{4.12}$$

Then the ratio between the probability, $k_>$, that a particle is located to the right of x_0 and will move to the right after the potential flips to the probability, $k_<$, that it is to the left of x_0 and will move to the left is

$$\frac{k_>}{k_<} = \frac{1-v}{1+v}. \tag{4.13}$$

When $s(t)$ changes sign, for a small D the particle then glides down with a constant velocity in its chosen direction, reaching the next minimum to the right ($>$) or to the left ($<$) in time

$$\tau_{\gtrless} = \frac{2\omega}{\beta}\frac{1}{(1\mp v)}. \tag{4.14}$$

If $\tau_{\gtrless} < 1$, then the particle has the time to reach the minimum on either the left or the right hand side (assuming, as always, that the sawtooth actually forms the local minima, i.e., $0 < v < 1$). Then when the potential flips the next time, the process repeats. This results in a discrete random walk between the extrema of the sawtooth, and (in unscaled variables)

$$D_{\text{eff}} = \left[(k_> + k_<) - (k_> - k_<)^2\right]\frac{L^2}{2T}, \tag{4.15}$$

$$v_{\text{eff}} = (k_> - k_<)\frac{L}{T}. \tag{4.16}$$

In dimensionless units,

$$r_D = \left[(k_> + k_<) - (k_> - k_<)^2\right] \omega, \tag{4.17}$$

$$r_v = (k_> - k_<) \frac{2\omega}{v\beta}. \tag{4.18}$$

Using Eq. (4.13) results in:

$$r_D = \left(1 - v^2\right) \omega, \tag{4.19}$$

$$r_v = \frac{2\omega}{\beta}. \tag{4.20}$$

Notably, $r_D \propto \frac{1}{D} \to \infty$ when $D \to 0$, but $r_D/\beta = D_{\text{eff}}/h$ is finite.

In Fig. 4.3, we compare the analytical results to the numerically estimated r_D and r_v (here r_D is normalized by $\beta/2$). As $\beta \to \infty$, the agreement is clearly seen for small ω, that is, for an infrequently flipping potential.

When the potential changes faster, and $\tau_> > 1 > \tau_<$, the particle fails to make it to the minimum to the right and, after a subsequent flip, always comes back to where it started from. However, it always reaches the left minimum before the flip. When $\tau_> > \tau_< > 1$, it does not reach the left minimum either, but goes the distance $\frac{\beta}{2\omega}(1 + v)$ to the left, then reverses and travels $\frac{\beta}{2\omega}(1 - v)$ to the right, reverses again and repeats until it reaches the left minimum. After one flip, it moves $\frac{\beta}{2\omega}(1 + v)$, after three flips it moves $\frac{\beta}{2\omega}(1 + 3v)$, and so on. Eventually, when $\beta/(2\omega)[(2n+1)v)+1] > 1$, the particle reaches the next minimum. Thus every time $2\omega/(\beta v)$ crosses an odd integer, more periods are needed to travel between the nearby extrema, and the diffusive behavior changes, resulting in the discontinuities in Fig. 4.3. These dynamics can be described by a master equation

$$P_i(t + 1) = (1 - k_>)P_{i+1}(t - 2n) + k_> P_i(t - 1), \tag{4.21}$$

where $P_i(t)$ stands for being at an extremum i at the end of the flip t, and exactly $2n + 1$ flips are needed to travel between the $i + 1$'th and the i'th extrema. Solving this for the

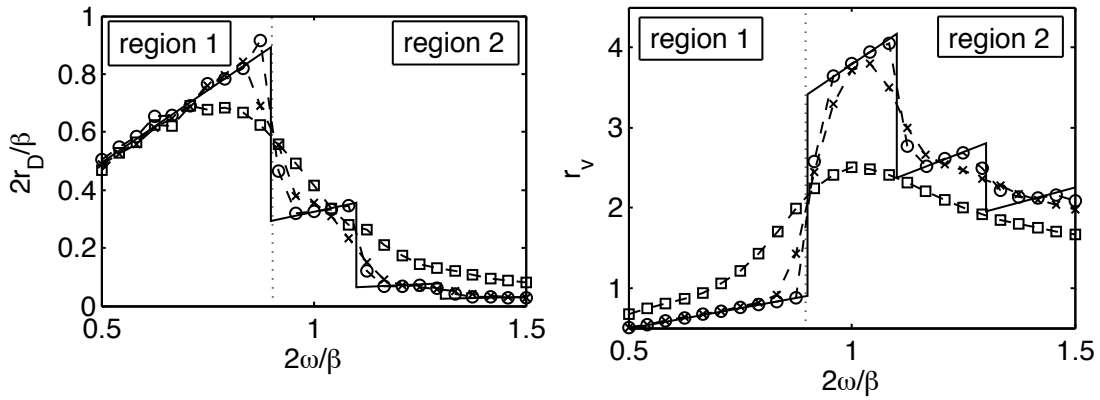

Figure 4.3: Effective diffusion (left) and effective drift (right) versus the period of the fluctuations. Notice that r_D is normalized by $2/\beta$, and it remains finite even if $D \to 0$ and $\beta \to \infty$. The data are obtained for $v = 1/10$ and (in decreasing order of noise strength) $\beta = 100$ (squares), $\beta = 1000$ (crosses), $\beta = 10,000$ (circles), $\beta \to \infty$ (solid line, analytical result). In the small noise case, the behavior of the diffusing particle is markedly different between Region 1 and Region 2. Region 1 corresponds to small ω when a particle can always travel between two extrema of the potential, performing an effective biased random walk. Region 2 corresponds to large ω, when the particle spends most of the time traveling between minima but rarely reaching them. Simulations are averaged over 1000 trajectories and 1000 time steps.

drift and the diffusion (see Appendix) gives

$$r_v = \frac{1-v}{3+v+2n(1-v)} \frac{2\omega}{v\beta}, \tag{4.22}$$

$$r_D = \frac{8(1-v^2)}{(3+v+2n(1-v))^3}\,\omega. \tag{4.23}$$

The analytical results and the numerical simulations verifying them are shown in Fig. 4.3 with ω normalized by $\beta/2$.

4.4 Fluctuating potential shortens the fitness barrier crossing time

We have shown that the typical diffusive/drift behavior of the system is enhanced by the fluctuating component of the potential. However, what kind of an effect does this enhancement have on the probability of rare, atypical events, such as escape from a suboptimal fitness maximum? To model a barrier in fitness, we place the overdamped particle in a

flipping sawtooth potential and a constant force v, and we observe the mean first passage time for the particle to reach an unscaled distance $M = L\mathcal{M}$, against the drift, with a reflecting boundary at $x = 0$, see Fig. 4.1.

4.4.1 Fluctuation-activated escape from the minimum is possible even at zero internal diffusion

If during one half-period the particle is able to travel between the two extrema (independent of the direction and without the help of the noise) then $\tau_{\gtrless} < 1$, and the probability that the particle travels over multiple periods against the (effective) drift is finite even at a very low noise. As seen in Fig. 4.4, the escape time over the average barrier in U_0 depends on the length of the barrier $\mathcal{M}$. In order to understand this dependence, we consider our model in the limit of zero fluctuating potential, which allows us to use an analytical expression for the escape time [150]

$$\langle t \rangle_D = \frac{\mathcal{M}^2}{D} \left[\frac{1}{2Pe} - \frac{1}{4Pe^2} \left(1 - e^{-2Pe} \right) \right], \tag{4.24}$$

where Pe is the Péclet number

$$Pe = \left| \frac{vL\mathcal{M}}{2D} \right|. \tag{4.25}$$

In Fig. 4.4, we show the simulation results of the exit time for slow fluctuations. We compare the results to what we would have expected under a continuous diffusion with the drift and the diffusion constant given by the values of the effective parameters as obtained in the previous section. We observe that the escape times are well approximated with the "effective" continuous diffusion model. The conclusion is valid even for a very small intrinsic noise $D \to 0$, which implies that the average time to escape over the global barrier is fluctuation activated (as opposed to noise activated), and is much faster.

We emphasize again that the fact that even rare escape events in the fluctuating potential are modeled well with the drift/diffusion approximation is different from the earlier observation that the typical behavior in this system is diffusive.

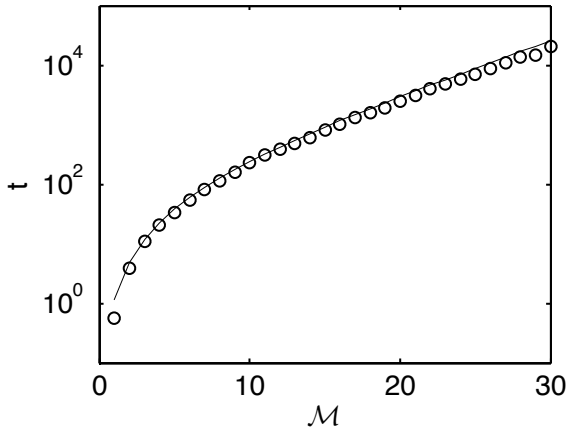

Figure 4.4: Mean first passage times for crossing a barrier of width $\mathcal{M}$ in our model (simulation, circles) and in the corresponding continuous diffusion coarse-grained model (solid line, Eq. (4.24)) with effective parameters obtained from simulation. The parameters are $\beta = 10^4$, $\omega = 2.5$, $v = -0.1$, and they correspond to the leftmost point in Fig. 4.3. The small discrepancy between simulation and analytical approximation is due the continuous nature of the coarse-grained model. A better approximation has been obtained with a model of discrete jumps between minima, which we don't show here.

4.4.2 Fluctuations enhance escape even for steep barriers

The qualitative behavior of particle trajectories changes if the fluctuating potential is fast enough such that the particle can only move less than one period in one direction in the absence of the noise. Even though, on average, the variance of the particle position grows linearly, and one can define a proper effective diffusion coefficient, the particle never crosses a barrier against drift in the absence of the intrinsic noise, $\beta \to \infty$. Hence the probability of rare excursions against the effective drift cannot be described using the same effective drift/diffusion model.

Fig. 4.5 reports results of numerical simulations at different values of the additive noise. The escape time as a function of $\mathcal{M}$ still can be fitted well with a drift/diffusion model, Eq. (4.24), with a noise dependent effective Péclet number $Pe(\beta)$. We conclude that, for small noises, the effective Péclet number is approximatively inversely proportional to the noise strength (the plot seems to reach a constant for $\beta \to \infty$). This implies that the escape times will become infinite at zero noise. The dependence is consistent with a model without any fluctuating potential, but with some effective parameters. The parameters are such that, with the fluctuations, the escapes are significantly faster. Indeed as shown in Fig. 4.5,

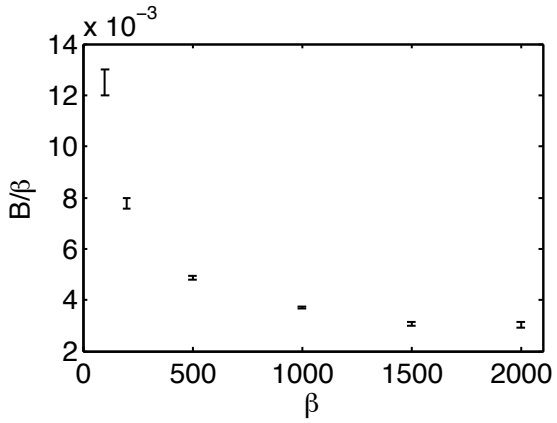

Figure 4.5: Dependence of the effective Peclet number, $Pe_{\text{eff}} = B\mathcal{M}$ on the intrinsic noise. We simulate exit times for systems of lengths $\mathcal{M} = 1, 2, 3, 4, 5$ for different β. For each dataset with the same β, we performed a weighted least squares fit for the average escape time of the form $t = A[\exp(B\mathcal{M}) - 1 - B\mathcal{M}]$, fitting for A and B. The error bars indicate the confidence bounds of the fits. We show B/β versus β decreases sublinearly and reaches a constant at $\beta \to \infty$. Thus the mean exit time diverges, but much slower than for diffusion without the fluctuating potential. For example, for the parameter values used here ($2\omega/\beta = 1$, $v = 0.1$), we have $B/\beta = 0.1$ with only the intrinsic diffusion, and it is $\sim 10^{-3}$ in the fluctuating potential model as seen in the Figure.

the effective Péclet number, defined as $Pe_{\text{eff}} = B\mathcal{M}$, is always smaller than the equivalent quantity in Eq. (4.24), such that

$$B \ll \left| \frac{vL}{2D} \right| . \tag{4.26}$$

4.5 Discussion

Using a one dimensional model of diffusion in the presence of a constant force perturbed by small periodic fluctuations, we have shown that time dependent potentials can significantly speed up the large scale drift, diffusion, and barrier escape times. This conclusion is valid even for a very small intrinsic diffusivity, the long time statistics of the particle trajectories being mainly determined by the properties of the potential fluctuations, but not of the Langevin noise.

Our model is a caricature of evolutionary dynamics in the limit of low mutation rates and constant population sizes. Even in this limit, there are several simplifying assumptions in our model that can be relaxed.

First, the periodicity of the potential time dependence is not crucial. Based on the similarity with Brownian motor models [141, 142], we expect that our conclusions will still be valid for a nonperiodic $s(t)$: the nonperiodic flipping will mix together and average behaviors from the different regions in Fig. 4.3. Further, Dubkov et al. [143] studied a randomly flipping sawtooth with no drift. Their potential flipped with dichotomous Markovian noise with rate v. They found an analytic expression for the diffusion enhancement r_D which grows slower than in our model for small ω, and the peak of r_D is lower and at larger ω. This is consistent with the averaging over different regions in Fig. 4.3. Similarly, if the potential is not spatially periodic, this will introduce a quenched noise and is likely to result in emergence of regions in x that are very hard to cross, similar to [146].

Our model is based on a piecewise linear periodic potential with discontinuous first derivatives (sharp minima and maxima). We expect that our conclusions stay valid for smoother perturbations as long as, upon a flip, a particle can leave the vicinity of a maximum in a time much faster than a typical travel time between the extrema.

The piecewise-constant perturbation in our model is symmetric with respect to reflection, thus no mean rectified motion can be created [141, 145, 142]. Simply put, the system is not a ratchet, and crossing of large scale barriers instead is achieved by creating a large effective diffusion. Asymmetry of the perturbation may create additional rectification, but the results in Fig. 4.3 suggest that the diffusive effects will not change qualitatively. This is crucial for evolutionary modeling since crossing barriers by ratcheting requires tuning the ratchet in a specific direction in the genotype space, while diffusion will explore the space isotropically.

The genomic space is expected to be high dimensional. For simplicity we have considered a one-dimensional model of mutation accumulation, which could still be relevant for highly epistatic genomic landscapes with a large proportion of highly deleterious or non-viable genotypes. Moreover, we expect that systematic application of multiscale homogenization methods [145], would allow one to extend the results to these more general formulations.

If the assumption of well separated time scales and constant population size are not satisfied, the Langevin microscopic dynamics used here is not valid any more. Then different fitnesses can give rise to different population sizes, and hence to varying fixation rates and

to a space dependent D in our language, which would require more complex models with position and time dependent noise strengths. Such models would modify the probability of individual trajectories [151, 149], and more work is needed in order to identify the regimes in which such diffusion models and their predictions apply to population genetics dynamics.

Appendix A

Fisher's Fundamental Theorem

The starting point for learning about adaptation historically has been Fisher's fundamental theorem of natural selection. Here we present a simplified derivation (for asexuals) which shows that in essence it is a generic statement about exponential growth. Consider a subpopulation, n_i, that grows exponentially in time with a constant rate f_i:

$$\dot{n_i} = f_i n_i,$$

with a total population size that also grows in time:

$$N = \sum_i n_i.$$

We can define the mean fitness

$$F = \frac{1}{N} \sum f_i n_i,$$

and the rate of change of average fitness

$$\dot{F} = \frac{1}{N} \sum_i f_i \dot{n_i} - \frac{\dot{N}}{N^2} \sum_i f_i n_i$$

$$\dot{F} = \frac{1}{N} \sum_i f_i^2 n_i - \frac{1}{N^2} \left(\sum_i f_i n_i \right)^2$$

equals the variance of f_i. A similar result also holds for discrete times, and diploid populations.

Appendix B

Diffusion and drift in the no-noise limit

If it takes $2n+1$ flips to go from site $i+1$ to site i, and, going rightwards, one returns back in two flips, then

$$P_i(t+1) = (1-b)P_{i+1}(t-2n) + bP_i(t-1). \tag{B.1}$$

Multiplying by i and summing over it, we get

$$\langle i(t+1)\rangle = (1-b)[\langle i(t-2n)\rangle + 1] + b\langle i(t-1)\rangle. \tag{B.2}$$

Assuming that the average can be written as

$$\langle i(t)\rangle = c + v_{\text{eff}}t, \tag{B.3}$$

we obtain

$$v_{\text{eff}} = \frac{1-b}{1+2n(1-b)+b}. \tag{B.4}$$

Now multiplying Eq. (B.1) by i^2 and summing over i, we get

$$\langle i^2(t+1)\rangle = (1-b)[\langle i^2(t-2n)\rangle + 2\langle i(t-2n)\rangle + 1] + b\langle i^2(t-1)\rangle. \tag{B.5}$$

This allows to write for the variance of i at moment $t+1$

$$\sigma^2(t+1) = (1-b) + (1-b)\sigma^2(t-2n) + b\sigma^2(t-1) + 2(1-b)\langle i(t-2n)\rangle \\ - (1-b)[\langle i(t-2n)\rangle + \langle i(t+1)\rangle]v(2n+1) \\ - 2bv[\langle i(t-2n)\rangle + \langle i(t+1)\rangle]. \quad \text{(B.6)}$$

Now assuming

$$\sigma^2(t) = C + D_{\text{eff}}t, \quad \text{(B.7)}$$

we get

$$D_{\text{eff}} = \frac{4(1-b)b}{(1+b+(1-b)2n)^3}. \quad \text{(B.8)}$$

Bibliography

[1] John H. Gillespie. *Population Genetics: A Concise Guide*. JHU Press, 2004.

[2] KA Wetterstrand. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing Program, 2012.

[3] Jeffrey E Barrick, Dong Su Yu, Sung Ho Yoon, Haeyoung Jeong, Tae Kwang Oh, Dominique Schneider, Richard E Lenski, and Jihyun F Kim. Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature*, 461(7268):1243–7, October 2009.

[4] Tim F Cooper and Richard E Lenski. Experimental evolution with E. coli in diverse resource environments. I. Fluctuating environments promote divergence of replicate populations. *BMC evolutionary biology*, 10:11, January 2010.

[5] Paul D Sniegowski and Philip J Gerrish. Beneficial mutations and the dynamics of adaptation in asexual populations. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 365(1544):1255–63, April 2010.

[6] Richard E. Lenski. E. coli Long-term Experimental Evolution Project, 2012.

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

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

[9] Philip Gerrish and Richard Lenski. The fate of competing beneficial mutations in an asexual population. *Genetica*, 102-103(0):127–144, 1998.

[10] Eric Brunet, Igor M Rouzine, and Claus O Wilke. The stochastic edge in adaptive evolution. *Genetics*, 179(1):603–20, May 2008.

[11] Igor M Rouzine, Eric Brunet, and Claus O Wilke. The traveling-wave approach to asexual evolution: Muller's ratchet and speed of adaptation. *Theoretical population biology*, 73(1):24–46, February 2008.

[12] Michael M Desai and Daniel S Fisher. Beneficial mutation selection balance and the effect of linkage on positive selection. *Genetics*, 176(3):1759–98, July 2007.

[13] Michael M Desai, Daniel S Fisher, and Andrew W Murray. The speed of evolution and maintenance of variation in asexual populations. *Current Biology*, 17(5):385–94, March 2007.

[14] Oskar Hallatschek. The noisy edge of traveling waves. *Proceedings of the National Academy of Sciences of the United States of America*, 108(5):1783–7, December 2010.

[15] Daniel S Fisher. Leading the dog of selection by its mutational nose. *Proceedings of the National Academy of Sciences of the United States of America*, 108(7):2633–2634, February 2011.

[16] S Wright. Isolation by distance. *Genetics*, (March), 1943.

[17] Motoo Kimura and GH Weiss. The stepping stone model of population structure and the decrease of genetic correlation with distance. *Genetics*, (480):561–576, 1964.

[18] R A Fisher. The wave of advance of advantageous genes. *Annals of Eugenics*, 7(Part 4):355–369, June 1937.

[19] A N Kolmogorov, I G Petrovskii, and N S Piskunov. Study of the diffusion equation with growth of the quantity of matter and its application to a biological problem. *Moscow Univ Bull Math*, 1(1):1–25, 1937.

[20] Lori J Lawson Handley, Andrea Manica, Jérôme Goudet, and François Balloux. Going the distance: human population genetics in a clinal world. *Trends in genetics : TIG*, 23(9):432–9, September 2007.

[21] Oskar Hallatschek, Pascal Hersen, Sharad Ramanathan, and David R Nelson. Genetic drift at expanding frontiers promotes gene segregation. *Proceedings of the National Academy of Sciences of the United States of America*, 104(50):19926–30, December 2007.

[22] Oskar Hallatschek and David R Nelson. Gene surfing in expanding populations. *Theoretical population biology*, 73(1):158–70, March 2008.

[23] Oskar Hallatschek and David R Nelson. Life at the front of an expanding population. *Evolution; international journal of organic evolution*, 64(1):193–206, January 2010.

[24] K. S. Korolev, Oskar Hallatschek, and David R. Nelson. Genetic demixing and evolution in linear stepping stone models. *Reviews of Modern Physics*, 82(2):1691–1718, May 2010.

[25] Kirill S Korolev, Melanie J I Müller, Nilay Karahan, Andrew W Murray, Oskar Hallatschek, and David R Nelson. Selective sweeps in growing microbial colonies. *Physical biology*, 9(2):026008, April 2012.

[26] Jakub Otwinowski and Stefan Boettcher. Accumulation of beneficial mutations in one dimension. *Physical Review E*, 84(1):1–6, July 2011.

[27] Erik a Martens and Oskar Hallatschek. Interfering waves of adaptation promote spatial mixing. *Genetics*, 189(3):1045–60, November 2011.

[28] Daniel Segrè, Alexander Deluna, George M Church, and Roy Kishony. Modular epistasis in yeast metabolism. *Nature genetics*, 37(1):77–83, January 2005.

[29] Michael Costanzo, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D Spear, Carolyn S Sevier, Huiming Ding, Judice L Y Koh, Kiana Toufighi, Sara Mostafavi, Jeany Prinz, Robert P St Onge, Benjamin VanderSluis, Taras Makhnevych, Franco J Vizeacoumar, Solmaz Alizadeh, Sondra Bahr, Renee L Brost, Yiqun Chen, Murat Cokol, Raamesh Deshpande, Zhijian Li, Zhen-Yuan Lin, Wendy Liang, Michaela Marback, Jadine Paw, Bryan-Joseph San Luis, Ermira Shuteriqi, Amy Hin Yan Tong, Nydia van Dyk, Iain M Wallace, Joseph a Whitney, Matthew T Weirauch, Guoqing

Zhong, Hongwei Zhu, Walid a Houry, Michael Brudno, Sasan Ragibizadeh, Balázs Papp, Csaba Pál, Frederick P Roth, Guri Giaever, Corey Nislow, Olga G Troyanskaya, Howard Bussey, Gary D Bader, Anne-Claude Gingras, Quaid D Morris, Philip M Kim, Chris a Kaiser, Chad L Myers, Brenda J Andrews, and Charles Boone. The genetic landscape of a cell. *Science (New York, N.Y.)*, 327(5964):425–31, January 2010.

[30] Jason H Moore. A global view of epistasis. *Nature genetics*, 37(1):13–4, January 2005.

[31] Patrick C Phillips. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. *Nature reviews. Genetics*, 9(11):855–67, November 2008.

[32] Chris Stark, Bobby-Joe Breitkreutz, Andrew Chatr-Aryamontri, Lorrie Boucher, Rose Oughtred, Michael S Livstone, Julie Nixon, Kimberly Van Auken, Xiaodong Wang, Xiaoqi Shi, Teresa Reguly, Jennifer M Rust, Andrew Winter, Kara Dolinski, and Mike Tyers. The BioGRID Interaction Database: 2011 update. *Nucleic acids research*, 39(Database issue):D698–704, January 2011.

[33] Anastasia Baryshnikova, Michael Costanzo, Yungil Kim, Huiming Ding, Judice Koh, Kiana Toufighi, Ji-Young Youn, Jiongwen Ou, Bryan-Joseph San Luis, Sunayan Bandyopadhyay, Matthew Hibbs, David Hess, Anne-Claude Gingras, Gary D Bader, Olga G Troyanskaya, Grant W Brown, Brenda Andrews, Charles Boone, and Chad L Myers. Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nature methods*, 7(12):1017–24, December 2010.

[34] Stuart Kauffman and Simon Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, 128(1):11–45, September 1987.

[35] CA Macken and AS Perelson. Protein evolution on rugged landscapes. *Proceedings of the National . . .*, 86(August):6191–6195, 1989.

[36] Henrik Flyvbjer and Benny Lautrup. Evolution in a rugged fitness landscape. *Physical Review A*, 46(10):6714–6723, 1992.

[37] Stuart A. Kauffman and Edward D. Weinberger. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, 141(2):211–245, November 1989.

[38] A S Perelson and C A Macken. Protein evolution on partially correlated landscapes. *Proceedings of the National Academy of Sciences of the United States of America*, 92(21):9657–9661, 1995.

[39] M E J Newman and R Engelhardt. Effects of selective neutrality on the evolution of molecular species. *Proceedings of the Royal Society B Biological Sciences*, 265(1403):1333–1338, 1998.

[40] Matthew C Cowperthwaite and Lauren Ancel Meyers. How Mutational Networks Shape Evolution: Lessons from RNA Models. *Annual Review of Ecology Evolution and Systematics*, 38(1):203–230, 2007.

[41] J. F. C. Kingman. A Simple Model for the Balance between Selection and Mutation. *Journal of Applied Probability*, 15(1):1–12, 1978.

[42] John H. Gillespie. *The Causes of Molecular Evolution*. Oxford University Press, 1994.

[43] Hsin-Hung Chou, Hsuan-Chao Chiu, Nigel F Delaney, Daniel Segrè, and Christopher J Marx. Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science (New York, N.Y.)*, 332(6034):1190–2, June 2011.

[44] Aisha I Khan, Duy M Dinh, Dominique Schneider, Richard E Lenski, and Tim F Cooper. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science (New York, N.Y.)*, 332(6034):1193–6, June 2011.

[45] HA Orr. The population genetics of adaptation: the adaptation of DNA sequences. *Evolution*, 56(7):1317–1330, 2002.

[46] H Allen Orr. The distribution of fitness effects among beneficial mutations. *Genetics*, 163(4):1519–1526, 2003.

[47] Paul Joyce, Darin R Rokyta, Craig J Beisel, and H Allen Orr. A general extreme value theory model for the adaptation of DNA sequences under strong selection and weak mutation. *Genetics*, 180(3):1627–1643, 2008.

[48] Sergey Kryazhimskiy, Gasper Tkacik, and Joshua B Plotkin. The dynamics of adaptation on correlated fitness landscapes. *Proceedings of the National Academy of Sciences of the United States of America*, 106(44):18638–43, November 2009.

[49] I.G. Szendro, M.F. Schenk, J. Franke, J. Krug, and J. de Visser. Quantitative analyses of empirical fitness landscapes. *Arxiv preprint arXiv:1202.4378*, 2012.

[50] Mark Lunzer, Stephen P Miller, Roderick Felsheim, and Antony M Dean. The biochemical architecture of an ancient adaptive landscape. *Science (New York, N.Y.)*, 310(5747):499–501, October 2005.

[51] Daniel M Weinreich, Nigel F Delaney, Mark A Depristo, and Daniel L Hartl. Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science (New York, N.Y.)*, 312(5770):111–4, April 2006.

[52] Frank J Poelwijk, Daniel J Kiviet, Daniel M Weinreich, and Sander J Tans. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445(7126):383–6, January 2007.

[53] Jasper Franke, A. Klözer, J.A.G.M. de Visser, and Joachim Krug. Evolutionary accessibility of mutational pathways. *PLoS computational biology*, 7(8):e1002134, 2011.

[54] Jason N Pitt and Adrian R Ferré-D'Amaré. Rapid construction of empirical RNA fitness landscapes. *Science (New York, N.Y.)*, 330(6002):376–9, October 2010.

[55] Trevor Hinkley, João Martins, Colombe Chappey, Mojgan Haddad, Eric Stawiski, Jeannette M Whitcomb, Christos J Petropoulos, and Sebastian Bonhoeffer. A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nature genetics*, 43(5):487–9, May 2011.

[56] M Kimura. Process Leading to Quasi-Fixation of Genes in Natural Populations Due to Random Fluctuation of Selection Intensities. *Genetics*, 39(3):280–95, May 1954.

[57] Naoyuki Takahata and Motoo Kimura. Maintenance of genetic variability under mutation and selection pressures in a finite population. *Proceedings of the National Academy of Sciences of the United States of America*, 74(6):5813–5817, 1977.

[58] N. Takahata. Genetic Variability Maintained in a Finite Population under Mutation and Autocorrelated Random Fluctuation of Selection Intensity. *Proceedings of the National Academy of Sciences*, 76(11):5813–5817, November 1979.

[59] Naoyuki Takahata. Genetic Variability and Rate of Gene Substitution in a Finite Population under Mutation and Fluctuating Selection. *Genetics*, 98(2):427–440, 1981.

[60] D G Heckel and J Roughgarden. A species near its equilibrium size in a fluctuating environment can evolve a lower intrinsic rate of increase. *Proceedings of the National Academy of Sciences of the United States of America*, 77(12):7497–500, December 1980.

[61] Robert H. MacArthur and Edward O. Wilson. *The theory of island biogeography*. Princeton University Press, 2001.

[62] Russell Lande, Steinar Engen, and Bernt-Erik Saether. An evolutionary maximum principle for density-dependent population dynamics in a fluctuating environment. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 364(1523):1511–8, June 2009.

[63] Toshiyuki Namba. Competitive Co-existence in a seasonally fluctuating environment. *Journal of Theoretical Biology*, pages 369–386, 1984.

[64] J M Cushing. Periodic Lotka-Volterra competition equations. *Journal of mathematical biology*, 24(4):381–403, January 1986.

[65] Ville Mustonen and Michael Lässig. Adaptations to fluctuating selection in Drosophila. *Proceedings of the National Academy of Sciences of the United States of America*, 104(7):2277–82, February 2007.

[66] Ville Mustonen and Michael Lässig. Molecular Evolution under Fitness Fluctuations. *Physical Review Letters*, 100(10):4–7, March 2008.

[67] Ville Mustonen and Michael Lässig. From fitness landscapes to seascapes: non-equilibrium dynamics of selection and adaptation. *Trends in genetics : TIG*, 25(3):111–9, March 2009.

[68] Frank J Poelwijk, Marjon G J de Vos, and Sander J Tans. Tradeoffs and Optimality in the Evolution of Gene Regulation. *Cell*, 146(3):462–470, July 2011.

[69] Nadav Kashtan, Elad Noor, and Uri Alon. Varying environments can speed up evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 104(34):13711–6, August 2007.

[70] Nadav Kashtan, Avi E Mayo, Tomer Kalisky, and Uri Alon. An analytically solvable model for rapid evolution of modular structure. *PLoS computational biology*, 5(4):e1000355, April 2009.

[71] Lutz Becks and Aneil F Agrawal. Higher rates of sex evolve in spatially heterogeneous environments. *Nature*, 468(7320):89–92, November 2010.

[72] Nick Colegrave. Sex releases the speed limit on evolution. *Nature*, 420(6916):664–6, December 2002.

[73] J Arjan G M de Visser, Su-Chan Park, and Joachim Krug. Exploring the effect of sex on empirical fitness landscapes. *The American naturalist*, 174 Suppl(July 2009):S15–30, July 2009.

[74] Joseilme Gouveia, Viviane de Oliveira, Caio Sátiro, and Paulo Campos. Rate of fixation of beneficial mutations in sexual populations. *Physical Review E*, 79(6):1–5, June 2009.

[75] Peter D Keightley and Sarah P Otto. Interference among deleterious mutations favours sex and recombination in finite populations. *Nature*, 443(7107):89–92, September 2006.

[76] Yuseob Kim and H Allen Orr. Adaptation in sexuals vs. asexuals: clonal interference and the Fisher-Muller model. *Genetics*, 171(3):1377–86, November 2005.

[77] R a Neher, B I Shraiman, and D S Fisher. Rate of adaptation in large sexual populations. *Genetics*, 184(2):467–81, February 2010.

[78] Daniel B Weissman, Marcus W Feldman, and Daniel S Fisher. The rate of fitness-valley crossing in sexual populations. *Genetics*, 186(4):1389–410, December 2010.

[79] N Colegrave and S Collins. Experimental evolution: experimental evolution and evolvability. *Heredity*, 100(5):464–70, May 2008.

[80] Jeremy a Draghi, Todd L Parsons, Günter P Wagner, and Joshua B Plotkin. Mutational robustness can facilitate adaptation. *Nature*, 463(7279):353–5, January 2010.

[81] Joanna Masel and Meredith V Trotter. Robustness and evolvability. *Trends in genetics : TIG*, 26(9):406–14, September 2010.

[82] Michael Lässig. From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC bioinformatics*, 8 Suppl 6:S7, January 2007.

[83] Richard Neher and Boris Shraiman. Statistical genetics and evolution of quantitative traits. *Reviews of Modern Physics*, 83(4):1283–1300, November 2011.

[84] Daniel B Weissman, Michael M Desai, Daniel S Fisher, and Marcus W Feldman. The rate at which asexual populations cross fitness valleys. *Theoretical population biology*, 75(4):286–300, June 2009.

[85] Lília Perfeito, Lisete Fernandes, Catarina Mota, and Isabel Gordo. Adaptive mutations in bacteria: high rate and small effects. *Science (New York, N.Y.)*, 317(5839):813–5, August 2007.

[86] J.A.G.M. de Visser, Clifford W. Zeyl, Philip J. Gerrish, Jeffrey L. Blanchard, and Richard E. Lenski. Diminishing Returns from Mutation Supply Rate in Asexual Populations. *Science*, 283(January):404–406, 1999.

[87] Aaron C Shaver, Peter G Dombrowski, Joseph Y Sweeney, Tania Treis, Renata M Zappala, and Paul D Sniegowski. Fitness Evolution and the Rise of Mutator Alleles in Experimental Escherichia coli Populations. *Genetics*, 566(October):557–566, 2002.

[88] Su-Chan Park and Joachim Krug. Clonal interference in large populations. *Proceedings of the National Academy of Sciences of the United States of America*, 104(46):18135–40, November 2007.

[89] Su-Chan Park, Damien Simon, and Joachim Krug. The Speed of Evolution in Large Asexual Populations. *Journal of Statistical Physics*, 138(1-3):381–410, January 2010.

[90] Hilary M. Lappin-Scott and J. William Costerton. *Microbial Biofilms*. Cambridge University Press, 2003.

[91] Lauren M F Merlo, John W Pepper, Brian J Reid, and Carlo C Maley. Cancer as an evolutionary and ecological process. *Nature reviews. Cancer*, 6(12):924–35, December 2006.

[92] Erik a Martens, Rumen Kostadinov, Carlo C Maley, and Oskar Hallatschek. Spatial structure increases the waiting time for cancer. *New Journal of Physics*, 13(11):115014, November 2011.

[93] Takeo Maruyama. On the fixation probability of mutant genes in a subdivided population. *Genetical Research*, 15(02):221–225, April 1970.

[94] T Maruyama. A simple proof that certain quantities are independent of the geographical structure of population. *Theoretical population biology*, 154:148–154, 1974.

[95] Motoo Kimura. On the probability of fixation of mutant genes in a population. *Genetics*, 47(6):713–9, June 1962.

[96] Su-Chan Park, Damien Simon, and Joachim Krug. The Speed of Evolution in Large Asexual Populations. *Journal of Statistical Physics*, 138(1-3):381–410, January 2010.

[97] C Doering, C Mueller, and P Smereka. Interacting particles, the stochastic Fisher-Kolmogorov-Petrovsky-Piscounov equation, and duality. *Physica A: Statistical Mechanics and its Applications*, 325(1-2):243–259, July 2003.

[98] Oskar Hallatschek and K. S. Korolev. Fisher Waves in the Strong Noise Limit. *Physical Review Letters*, 103(10):108103, September 2009.

[99] Albert-László Barabási and Harry Eugene Stanley. *Fractal concepts in surface growth*. Cambridge University Press, 1995.

[100] Nigel Goldenfeld. Kinetics of a model for nucleation-controlled polymer crystal growth. *Journal of Physics A: Mathematical and General*, 2807, 1984.

[101] Michael Praehofer and Herbert Spohn. Universal Distributions for Growth Processes in 1+1 Dimensions and Random Matrices. *Physical Review Letters*, 84(21):4, 1999.

[102] F Family and T Vicsek. Scaling of the active zone in the Eden process on percolation networks and the ballistic deposition model. *Journal of Physics A: Mathematical and General*, 18(2):L75–L81, February 1985.

[103] Mehran Kardar, Giorgio Parisi, and Yi-Cheng Zhang. Dynamic Scaling of Growing Interfaces. *Physical Review Letters*, 56(9):889–892, 1986.

[104] Thomas Kriecherbauer and Joachim Krug. A pedestrian's view on interacting particle systems, KPZ universality and random matrices. *Journal of Physics A: Mathematical and Theoretical*, 43(40):403001, October 2010.

[105] Craig A. Tracy and Harold Widom. Fredholm determinants, differential equations and matrix models. *Communications in Mathematical Physics*, 163(1):33–72, June 1994.

[106] Nigel Goldenfeld and Carl Woese. Life is Physics: Evolution as a Collective Phenomenon Far From Equilibrium. *Annual Review of Condensed Matter Physics*, 2(1):375–399, March 2011.

[107] S Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc. 6th Int. Congress Genetics*, 1:356–365, 1932.

[108] David W Hall, Matthew Agan, and Sara C Pope. Fitness epistasis among 6 biosynthetic loci in the budding yeast Saccharomyces cerevisiae. *The Journal of heredity*, 101 Suppl(Supplement 1):S75–84, 2010.

[109] Jack da Silva, Mia Coetzer, Rebecca Nedellec, Cristina Pastore, and Donald E Mosier. Fitness epistasis and constraints on adaptation in a human immunodeficiency virus type 1 protein region. *Genetics*, 185(1):293–303, May 2010.

[110] J G Kingsolver, H E Hoekstra, J M Hoekstra, D Berrigan, S N Vignieri, C E Hill, a Hoang, P Gibert, and P Beerli. The strength of phenotypic selection in natural populations. *The American naturalist*, 157(3):245–61, March 2001.

[111] Ruth G Shaw and Charles J Geyer. Inferring fitness landscapes. *Evolution; international journal of organic evolution*, 64(9):2510–20, September 2010.

[112] Rachel B Brem and Leonid Kruglyak. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proceedings of the National Academy of Sciences of the United States of America*, 102(5):1572–7, February 2005.

[113] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145, 2008.

[114] Thierry Mora, Aleksandra M Walczak, William Bialek, and Curtis G Callan. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 107(12):5405–10, March 2010.

[115] Roger D. Kouyos, Gabriel E. Leventhal, Trevor Hinkley, Mojgan Haddad, Jeannette M. Whitcomb, Christos J. Petropoulos, and Sebastian Bonhoeffer. Exploring the Complexity of the HIV-1 Fitness Landscape. *PLoS Genetics*, 8(3):e1002551, March 2012.

[116] Justin B Kinney, Anand Murugan, Curtis G Callan, and Edward C Cox. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *PNAS*, 2010.

[117] William H Green. *Econometric Analysis*. Prentice Hall, 2003.

[118] Tatyana O Sharpee, Hiroki Sugihara, Andrei V Kurgansky, Sergei P Rebrik, Michael P Stryker, and Kenneth D Miller. Adaptive filtering enhances information transmission in visual cortex. *Nature*, 439(7079):936–42, February 2006.

[119] O G Berg and P H von Hippel. Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. *Journal of molecular biology*, 200(4):709–23, April 1988.

[120] C B Harley and R P Reynolds. Analysis of E. coli promoter sequences. *Nucleic Acids Research*, 15(5):2343–2361, 1987.

[121] Thomas Kuhlman, Zhongge Zhang, Milton H Saier, and Terence Hwa. Combinatorial transcriptional control of the lactose operon of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 104(14):6043–8, April 2007.

[122] Christopher C. Paige and Michael A. Saunders. LSQR: An Algorithm for Sparse Linear Equations and Sparse Least Squares. *ACM Transactions on Mathematical Software*, 8(1):43–71, March 1982.

[123] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

[124] R Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (*, 1996.

[125] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal Of Statistical Software*, 33(1), 2010.

[126] O G Berg and P H von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *Journal of molecular biology*, 193(4):723–50, February 1987.

[127] Marko Djordjevic, Anirvan M Sengupta, and Boris I Shraiman. A biophysical approach to transcription factor binding site discovery. *Genome research*, 13(11):2381–90, November 2003.

[128] Amy L Bauer, William S Hlavacek, Pat J Unkefer, and Fangping Mu. Using Sequence-Specific Chemical and Structural Properties of DNA to Predict Transcription Factor Binding Sites. *PLoS Computational Biology*, 6(11):13, 2010.

[129] Frank J Poelwijk, Sorin Tanase-Nicola, Daniel J Kiviet, and Sander J Tans. Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *Journal of theoretical biology*, 272(1):141–4, March 2011.

[130] Erez Dekel and Uri Alon. Optimality and evolutionary tuning of the expression level of a protein. *Nature*, 436(July):588–592, 2005.

[131] L Perfeito, S Ghozzi, Johannes Berg, Karin Schnetz, and Michael Lässig. Nonlinear fitness landscape of a molecular pathway. *PLoS Genetics*, 7(7):1–10, 2011.

[132] Ulrich Gerland and Terence Hwa. On the selection and evolution of regulatory DNA motifs. *Journal of molecular evolution*, 55(4):386–400, October 2002.

[133] Johannes Berg, Stana Willmann, and Michael Lässig. Adaptive evolution of transcription factor binding sites. *BMC evolutionary biology*, 4:42, October 2004.

[134] Ville Mustonen and Michael Lässig. Evolutionary population genetics of promoters: predicting binding sites and functional phylogenies. *Proceedings of the National Academy of Sciences of the United States of America*, 102(44):15936–41, November 2005.

[135] Ville Mustonen, Justin Kinney, Curtis G Callan, and Michael Lässig. Energy-dependent fitness: a quantitative model for the evolution of yeast transcription factor binding sites. *Proceedings of the National Academy of Sciences of the United States of America*, 105(34):12376–81, August 2008.

[136] Nigel Goldenfeld and Leo P Kadanoff. Simple Lessons from Complexity. *Science*, 284(5411):87–89, 1999.

[137] S J Schrag, V Perrot, and B R Levin. Adaptation to the fitness costs of antibiotic resistance in Escherichia coli. *Proceedings of the Royal Society B Biological Sciences*, 264(1386):1287–91, September 1997.

[138] DM Weinreich and RA Watson. Perspective: sign epistasis and genetic constraint on evolutionary trajectories. *Evolution*, 59:1165–1174, 2005.

[139] D Allan Drummond and Claus O Wilke. The evolutionary consequences of erroneous protein synthesis. *Nature reviews. Genetics*, 10(10):715–24, October 2009.

[140] M. Scott, C. W. Gunderson, E. M. Mateescu, Z. Zhang, and T. Hwa. Interdependence of Cell Growth and Gene Expression: Origins and Consequences. *Science*, 330(6007):1099–1102, November 2010.

[141] Marcelo O. Magnasco. Forced thermal ratchets. *Physical Review Letters*, 71(10):1477–1481, September 1993.

[142] M B Tarlie and R D Astumian. Optimal modulation of a Brownian ratchet and enhanced sensitivity to a weak external force. *Proceedings of the National Academy of Sciences of the United States of America*, 95(5):2039–43, March 1998.

[143] AA Dubkov and Bernardo Spagnolo. Acceleration of diffusion in randomly switching potential with supersymmetry. *Physical Review E*, 72(4):1–8, October 2005.

[144] N. Sinitsyn and Ilya Nemenman. Universal Geometric Theory of Mesoscopic Stochastic Pumps and Reversible Ratchets. *Physical Review Letters*, 99(22):1–4, November 2007.

[145] M Vergassola and M Avellaneda. Scalar transport in compressible flow. *Physica D: Nonlinear Phenomena*, 106(1-2):148–166, July 1997.

[146] G Weiss and S Havlin. Some properties of a random walk on a comb structure. *Physica A: Statistical and Theoretical Physics*, 134(2):474–482, January 1986.

[147] R Metzler. The random walk's guide to anomalous diffusion: a fractional dynamics approach. *Physics Reports*, 339(1):1–77, December 2000.

[148] Ariel Lubelski, Igor Sokolov, and Joseph Klafter. Nonergodicity Mimics Inhomogeneity in Single Particle Tracking. *Physical Review Letters*, 100(25):250602, June 2008.

[149] Golan Bel and Ilya Nemenman. Ergodic and non-ergodic anomalous diffusion in coupled stochastic processes. *New Journal of Physics*, 11(8):083009, August 2009.

[150] Sidney Redner. *A guide to first-passage processes*. Cambridge University Press, 2001.

[151] Irina V Gopich and Attila Szabo. Theory of the statistics of kinetic transitions with application to single-molecule enzyme catalysis. *The Journal of chemical physics*, 124(15):154712, April 2006.