© 2014

Michael Manhart

ALL RIGHTS RESERVED

BIOPHYSICS AND STOCHASTIC PROCESSES IN MOLECULAR EVOLUTION

By

MICHAEL MANHART

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Physics and Astronomy

written under the direction of

Professor Alexandre V. Morozov

and approved by

New Brunswick, New Jersey

October, 2014

ABSTRACT OF THE DISSERTATION

BIOPHYSICS AND STOCHASTIC PROCESSES IN MOLECULAR EVOLUTION

By MICHAEL MANHART

Dissertation Director:

Professor Alexandre V. Morozov

Evolution is the defining feature of living matter. It occurs most fundamentally on the scale of biomolecules such as DNA and proteins, which carry out all the processes of cells. How do the physical properties of these molecules shape the course of evolution? We address this question using a synthesis of biophysical models, theoretical tools from stochastic processes, and high-throughput data. We first review some basic features of population and evolutionary dynamics, focusing especially on fitness landscapes and how they determine accessible pathways of evolution. We then derive a universal scaling law describing time reversibility and steady state of monomorphic populations on arbitrary fitness landscapes. We use this result to study the evolution of transcription factor (TF) binding sites using high-throughput data on TF-DNA interactions and genome-wide site locations. We find that binding sites for a given TF appear to be subjected to universal selection pressures, independent of the properties of their corresponding genes, and their binding energy-dependent fitness is consistent with a simple functional form inspired by a thermodynamic model. We next consider the properties of evolutionary pathways. We develop a general approach for calculating statistical properties of the path ensemble in a stochastic process. We first demonstrate this approach on a series of simple examples, including evolution on a neutral network and two reaction rate problems. We then apply these techniques to a model of how proteins evolve new binding interactions while maintaining folding stability. In particular we show how the structural coupling of protein folding and binding results in protein traits emerging as evolutionary “spandrels”: proteins can evolve strong binding interactions that confer no intrinsic fitness advantage but merely serve to stabilize the protein if misfolding is deleterious. These observations may explain the abundance of apparently nonfunctional interactions among proteins observed in high-throughput assays. When there are distinct selection pressures on both folding and binding, evolutionary paths of proteins can be tightly constrained so that folding stability is first gained and then partially lost as the new binding function is developed. This suggests the evolution of many natural proteins is highly predictable at the level of biophysical traits.

Preface

The matter making up living things consists of the same basic stuff — ordinary atoms of protons, neutrons, and electrons — as all the non-living baryonic matter in the universe. And yet, living matter produces behaviors far more complex than anything exhibited by non-living matter: living matter can make copies of itself, it can respond to its environment in complicated ways, it can remember and learn, it can even think about itself and write these very sentences. So what makes living matter so fundamentally different?

Erwin Schrödinger was one of the first to seriously contemplate biology from the first principles of physics. In his famous book *What is Life?* [1], Schrödinger speculated on the role of entropy in living matter as well as the physical basis of genetics. In the decades since that prescient work, the interface between physics and biology has become one of the most fertile areas of scientific research. We now view the astonishing uniqueness of living matter through the lens of emergence [2–4]: life emerges from the collective behavior of the diverse molecules making up a cell. But emergence provides only a conceptual framework, so characterizing the vast range of phenomena in biological systems requires a program of targeted experiments and modeling. This is the essence of "physical biology," which aims to characterize biological phenomena in terms of the underlying physico-chemical processes [5, 6], and is the philosophy underlying this dissertation.

Here we focus our attention on evolution. Indeed, evolution is the defining feature of living matter; Theodosius Dobzhansky famously wrote that "Nothing in biology makes sense except in the light of evolution" [7]. However, a logical addendum to this statement might be that nothing in evolution makes sense except in the light of physics: evolution occurs most fundamentally on the scale of biomolecules such as DNA and proteins, which carry out all the processes of cells. How do the physical properties of these molecules shape the course of evolution?

This is the major question driving my Ph.D. research, a synthesis of models from molecular biophysics, mathematical and computational techniques from stochastic processes and statistical physics, and data from high-throughput experiments. This dissertation summarizes the majority of my work on these topics during graduate school, drawing from the following papers:

- [8] Manhart M, Haldane A, Morozov AV (2012) A universal scaling law determines time reversibility and steady state of substitutions under selection. *Theor Popul Biol* 82:66–76.
- [9] Haldane A, Manhart M, Morozov AV (2014) Biophysical fitness landscapes for transcription factor binding sites. *PLoS Comput Biol* 10:e1003683.
- [10] Manhart M, Morozov AV (2013) Path-based approach to random walks on networks characterizes how proteins evolve new functions. *Phys Rev Lett* 111:088102.
- [11] Manhart M, Morozov AV (2014) in *First-Passage Phenomena and Their Applications*, eds. Metzler R, Oshanin G, Redner S. (World Scientific, Singapore).
- [12] Manhart M, Morozov AV (2014) Protein folding and binding can emerge as evolutionary spandrels through structural coupling. arXiv:1408.3786.

Material from these papers is reproduced in full or in part, with some small changes and rearrangements made to produce a consistent, coherent whole. I have excluded some additional work that is not yet published or does not fit into the main narrative, especially my collaboration with the group of Gábor Balázsi on the evolution of a synthetic gene network.

We begin Chapter 1 with an introduction to some basic elements of population and evolutionary dynamics, drawn in part from Ref. 11. We focus especially on the concept of fitness landscapes and how they determine accessible pathways of evolution. Chapter 2 reproduces Ref. 8 and discusses time reversibility and properties of evolutionary steady state under natural selection. We derive a universal scaling law that implies a general form for the steady state distribution of monomorphic populations on arbitrary fitness landscapes. In Chapter 3 we use these results to study the evolution of transcription factor

(TF) binding sites, short segments of DNA that regulate nearby genes by binding to TF proteins. Using high-throughput *in vitro* measurements of TF-DNA binding interactions and a large collection of genomic binding site locations, we find that binding sites for a given TF appear to be subjected to universal selection pressures, independent of the properties of their corresponding genes. We are thus able to infer fitness landscapes for these sites as functions of their binding energy alone. Their energy-dependent fitness is consistent with a simple functional form inspired by a thermodynamic model, although the inferred parameters indicate selection pressures beyond the biophysical constraints imposed by TF-DNA interactions. This chapter is reproduced from Ref. 9.

We next take a detour through stochastic processes and statistical physics to understand the properties of evolutionary pathways. In Chapter 4 we develop a general path-based approach to stochastic processes on discrete state spaces, along with an efficient numerical algorithm for calculating statistical properties of the path ensemble. We demonstrate this approach on a series of simple examples, including evolution on a neutral network and two reaction rate problems. This chapter reproduces material from Refs. 10, 11. In Chapter 5 we apply these techniques to a model of how proteins evolve new binding interactions while maintaining folding stability. We characterize the fitness landscape for these proteins and the resulting evolutionary dynamics. In particular we show how the structural coupling of protein folding and binding results in protein traits emerging as evolutionary "spandrels": proteins can evolve strong binding interactions that confer no intrinsic fitness advantage but merely serve to stabilize the protein if misfolding is deleterious. This chapter reproduces material from Ref. 12. Finally, in Chapter 6 we conclude and provide an outlook for future work in this area.

Acknowledgments

My Ph.D. experience would not have been possible without a large supporting cast; I fear this one small section in my dissertation hardly does justice to the impact they've had on my life. But I will do my best to earnestly express my gratitude to all of them.

First of all, having Alex Morozov as my advisor and collaborator has been a joy. I knew almost nothing about biophysics and biology at the beginning of graduate school, but Alex nevertheless took a chance on me and opened my eyes to this incredible area of science. He has always supplied wisdom and guidance when needed, but he also gave me the freedom to develop my own interests and style, even from the very beginning. His kindness, dedication, and tenacity are an inspiration to me and model the kind of scientist I wish to become.

I am also grateful for the support of my other committee members: Anirvan Sengupta, Natan Andrei, Gyan Bhanot, and Kevin Chen. Anirvan has been a constant source of wisdom and knowledge throughout my graduate studies; he has always patiently listened to my questions and wild ideas. I am grateful to Natan who guided me through a year-long reading course in statistical physics, which was not only highly productive but a true pleasure. I thank Gyan and Kevin for their conversations and insightful feedback on my work.

The many students in the biophysics group during my time here — Manjul Apratim, Aatish Bhatia, Răzvan Chereji, Adel Dayarian, Dai Wei, Bill Flynn, Allan Haldane, Dave Hassan, Pasha Khromov, Willow Kion-Crosby, Junyi Li, George Locke, Ted Malliaris, Mohammad Ramezanali, Julie Tsitron, Kshitij Wagh, Tahir Yusufaly — have been an essential source of moral, scientific, and technical support. I have learned far more from our conversations than I could in any class!

I am especially grateful to Allan for his research collaboration, which is well-represented in this dissertation. I am also grateful for my collaborators at the University of Texas MD Anderson Cancer Center — Gábor Balázsi, Caleb González, Christian Ray, Dmitry Nevozhay, and Rhys Adams. As a theorist, working closely with this talented group of experimentalists has been invaluable to my scientific growth.

Although it is not explicitly represented in this document, teaching has been a substantial component of my graduate education. It has ignited my passion for science and matured my knowledge. I especially thank Aatish Bhatia, Deepak Iyer, Simon Knapen, and Saurabh Jha for a rewarding experience co-teaching Physics 106. I thank Eric Gawiser for his continued stewardship of DELTA P, the TA training program originally organized by Simon, Heather Briggs, and myself. I had the pleasure of serving as a TA for Suzanne Brahmia; she and Eugenia Etkina have been enthusiastic supporters of all my teaching efforts, and I am grateful to both for their encouragement. I also thank Jill Knapp at Princeton for her indefatigable leadership of the Prison Teaching Initiative, in which I am humbled to have played a small role.

I would be remiss if I did not specifically thank Ron Ransome and Ted Williams, the two graduate program directors during my first few years in the department. I had a difficult first year in many ways; Ron's kindness and palpable commitment to all the graduate students were a major source of strength for me. In particular, he championed DELTA P, SSPAR, and Physics 106, and I cannot thank him enough for giving me those opportunities for personal and professional growth. Ted Williams continued to be a consistent advocate of these efforts after he succeeded Ron as GPD. I am especially grateful to Ted and the other GPDs for their generous financial support of my graduate studies. Shirley Hinds, our tireless graduate administrator, has been a patient and indispensable help for navigating the whole graduate program.

Among my other classmates in the department, there are too many to specifically thank here, but they know who they are! I have learned so much from taking classes together, teaching together, attending seminars together, and socializing together. My friends here at Rutgers and around the world made graduate school not merely manageable but downright enjoyable.

Finally, above all others I am grateful to my family and my dear Maryam. Neither words nor equations can express how much they mean to me.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction to Population and Evolutionary Dynamics

We begin with an introduction to the dynamics of populations and evolution. This will establish concepts and models needed for later chapters as well as provide important context. Our approach here is theoretical rather than empirical, and reproduces some material from our review in Ref. 11. For empirical discussions of evolution we refer the reader to standard expositions both at the organismic scale [13, 14] and at the molecular scale [15, 16]. Additional treatments of the theory can be found in several standard texts [15–20].

1.1 Elementary processes of population dynamics

The fundamental property of living matter is its ability to self-replicate. When an organism replicates, the information that it transmits to its offspring, which identifies the offspring as a copy of the parent, is the **genotype**. We will denote genotype as σ. In natural systems this genotype is physically encoded in DNA, RNA, or a protein, all of which are chains of elementary molecular building blocks. Thus we can represent the genotype σ as a sequence of L characters from an alphabet of size k, where the set of possible molecular building blocks defines the alphabet; in the absence of deletions and insertions, the total number of possible genotypes is k^L. DNA and RNA molecules store the genotype as a sequence of the nucleotide bases adenine (A), cytosine (C), guanine (G), and thymine or uracil (T or U). So a genotype representing a DNA or RNA sequence has $k = 4$. If the relevant genetic information is a gene, i.e., it codes for a protein molecule, then we can (almost) equivalently consider the sequence of amino acids making up the protein. In that case, there are $k = 20$ possible amino acids at each position in the sequence. A coarser description of genotype, applicable when we are considering many different genes across the genome, is to simply describe the binary state of each gene ($k = 2$) as either mutated or not mutated relative to some reference genotype.

A population consists of a large number of organisms, each of which can have a distinct genotype. Figure 1.1 shows a schematic of a population, where color indicates genotype. Since individual organisms only survive for short periods, it is not the state of the individuals but the state of the population that matters: the major goal of population genetics, and more broadly evolutionary theory, is to describe how the genetic state of a population changes over time.

In general there are four fundamental processes through which the population changes, illustrated in Fig. 1.1. The first is asexual birth, in which an organism produces a copy of itself with the same genotype. The second is death, in which an organism is somehow degraded or at least can no longer replicate. The third process is mutation, where the genotype of an organism spontaneously changes; this could happen as an independent event (e.g., if an organism is exposed to radiation), as shown in the figure, or it could occur as part of the birth process (e.g., due to errors in DNA replication). The fourth process is recombination; this can occur in many different ways, but in general it is some process that mixes two existing genotypes in the population. It is important to note that all of these processes are stochastic and therefore subject to random fluctuations.

Natural selection is an emergent phenomenon arising from the birth and death processes. Some genotypes will on average undergo more birth and less death than other genotypes. These genotypes will therefore dominate the population over long time scales, thus appearing to have been "selected" [13]. A genotype that is present in 100% of the population is said to have **fixed**. It is convenient to define a parameter s, called the **selection coefficient**, that characterizes the relative ability of a genotype to dominate a population. We will provide a more precise definition of s in the context of a specific model below, but in general it satisfies the operational definition as the exponential growth rate of one

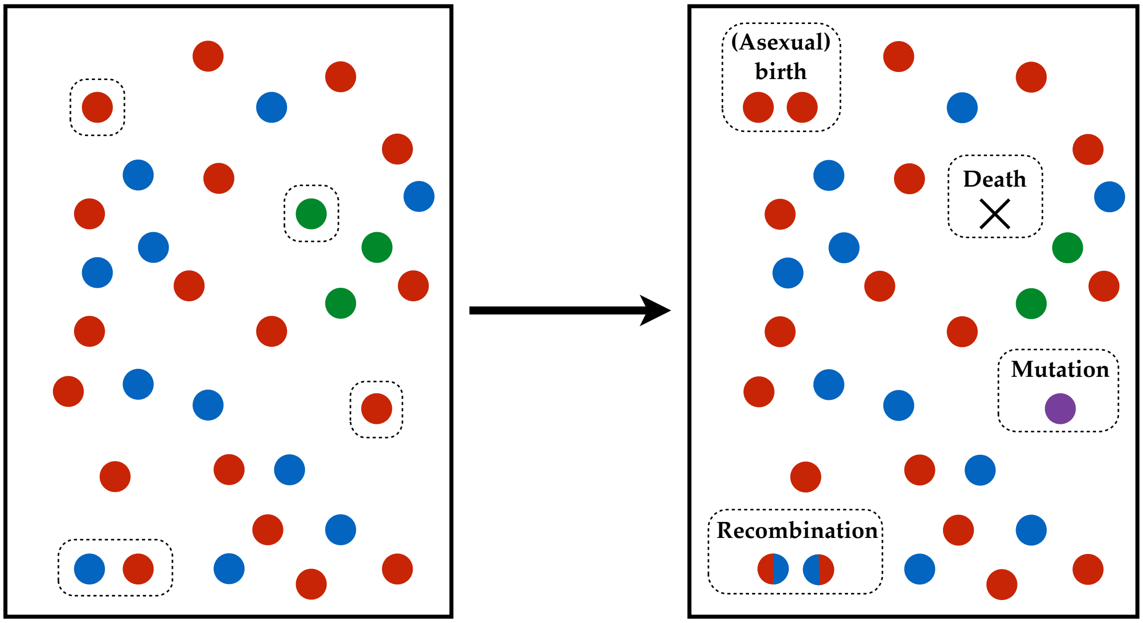


Figure 1.1: **Schematic of population dynamics.** Each organism is a dot, with different colors indicating different genotypes. Organisms in dashed boxes undergo the processes of asexual birth, death, mutation, and recombination.

genotype σ' in the presence of another genotype σ. Positive s means the genotype σ' grows faster and eventually dominates, while negative s means the genotype σ grows faster and will dominate. The case of $s = 0$ means that σ and σ' are selectively **neutral**, or equivalent under natural selection.

Each of the fundamental population processes has a characteristic rate or time scale. The characteristic rate of natural selection is s (consistent with the operational definition of the selection coefficient). Let u denote the characteristic rate of mutation and r the characteristic rate of recombination. These have corresponding time scales s^{-1}, u^{-1}, and r^{-1}. An additional important time scale is given by the population size N. This is the time scale of the stochastic fluctuations intrinsic to the population, known as **genetic drift**.

These basic time scales are important because their relative values determine the qualitative dynamics of a population to a large extent. In particular, processes with substantial separations of time scales decouple, dramatically simplifying the dynamics. For instance, if selection occurs more rapidly than the other processes ($s \gg u, r, N^{-1}$), then new beneficial mutations will rapidly and assuredly fix in the population. Mutation is therefore too slow to produce additional mutations before the first one has fixed, which precludes the existence of many genotypes simultaneously competing in the population (known as "clonal interference"), and genetic drift is too slow for a random fluctuation to cause the beneficial mutation to go extinct. Estimating these time scales and using them to simplify models is a recurring theme in the population genetics and evolutionary literature [21], and this work will be no exception.

1.2 A "theory of everything" for evolution

It is possible to write down a very general model of population dynamics that incorporates all these processes. Let $\mathbf{n} = \{n_\sigma\}$ be a list of "occupation numbers," where n_σ is the number of organisms in the population with genotype σ. The total population size is therefore $N = \sum_\sigma n_\sigma$. It is convenient to define shorthand notation for changes to the population via birth, death, mutation, and recombination. Let $\mathcal{B}_\sigma$ be a "birth operator" such that $\mathcal{B}_\sigma \mathbf{n} = \{n'_{\sigma'}\}$, where $n'_{\sigma'} = n_{\sigma'} + 1$ if $\sigma' = \sigma$ and $n'_{\sigma'} = n_{\sigma'}$ for $\sigma' \neq \sigma$. We can similarly define a death operator $\mathcal{D}_\sigma$ and a mutation operator $\mathcal{M}_{\sigma \to \sigma'}$. Note that $\mathcal{B}_\sigma$ and $\mathcal{D}_\sigma$ are inverses of each other, while $\mathcal{M}^{-1}_{\sigma \to \sigma'} = \mathcal{M}_{\sigma' \to \sigma}$. We will denote by $\mathcal{R}_{\sigma + \sigma' \to \sigma''}$ a recombination operator that creates a recombined genotype σ'' from σ and σ', with inverse operator $\mathcal{R}^{-1}_{\sigma + \sigma' \to \sigma''}$. Let $P(\mathbf{n}, t)$ be the probability that the population is in state $\mathbf{n}$ at time t. This obeys the following master equation:

$$
\begin{aligned}
\frac{\partial}{\partial t} P(\mathbf{n}, t) = & \underbrace{\sum_{\sigma} b_\sigma(\mathcal{D}_\sigma \mathbf{n}, t)(n_\sigma - 1) P(\mathcal{D}_\sigma \mathbf{n}, t)}_{\text{birth}} \quad + \quad \underbrace{\sum_{\sigma} d_\sigma(\mathcal{B}_\sigma \mathbf{n}, t)(n_\sigma + 1) P(\mathcal{B}_\sigma \mathbf{n}, t)}_{\text{death}} \\
& + \underbrace{\sum_{\sigma, \sigma'} u_{\sigma \to \sigma'}(n_\sigma + 1) P(\mathcal{M}_{\sigma' \to \sigma} \mathbf{n}, t)}_{\text{mutation}} \quad + \quad \underbrace{\sum_{\sigma, \sigma'} r_{\sigma + \sigma' \to \sigma''} n_\sigma n_{\sigma'} P(\mathcal{R}^{-1}_{\sigma + \sigma' \to \sigma''} \mathbf{n}, t)}_{\text{recombination}} \\
& - \sum_{\sigma} (b_\sigma(\mathbf{n}, t) + d_\sigma(\mathbf{n}, t)) n_\sigma P(\mathbf{n}, t) - \sum_{\sigma, \sigma'} (u_{\sigma', \sigma} n_\sigma + r_{\sigma + \sigma' \to \sigma''} n_\sigma n_{\sigma'}) P(\mathbf{n}, t),
\end{aligned}
\tag{1.1}
$$

where $u_{\sigma \to \sigma'}$ is the rate of an organism with genotype σ mutating into genotype σ' and $r_{\sigma + \sigma' \to \sigma''}$ is the rate at which two organisms with genotypes σ and σ' recombine into σ''. The birth and death rates $b_\sigma(\mathbf{n}, t)$ and $d_\sigma(\mathbf{n}, t)$ for an organism with genotype σ capture the effects of natural selection as previously described. As indicated by the notation, ecological interactions between subpopulations may cause these rates to depend on the state of the population (e.g., leading to frequency-dependent selection), and they may also change with time. Although not included here, it is straightforward to adapt this general model for other phenomena such as spatial and demographic structure.

In some sense, Eq. 1.1 is a “theory of everything” for evolution. In principle it provides a complete description of a population’s dynamics over all time scales. However, its role is similar to that of the many-body Schrödinger equation in physics, which similarly offers an essentially complete description of the quantum mechanical behavior of an arbitrary collection of atoms [3]. However, for any system containing more than just a few particles, it is presently impossible to solve the Schrödinger equation, either analytically or computationally. Thus the equation is relegated to a largely symbolic role, and effective phenomenological models are needed to study specific systems. Similarly, the generalized evolution equation (Eq. 1.1) is too complicated to be tractable in most cases, and we must develop more specialized, simplified models tailored to the specific phenomena under consideration.

But there is also a critical difference between the evolution equation and the Schrödinger equation. Although the Schrödinger equation is too complicated to solve, we know all of its terms: we can write explicit mathematical expressions for the kinetic terms and pairwise Coulomb interactions between the electrons and the nuclei. This is not the case for the evolution equation. While simple models for mutation and recombination may suffice, the birth and death terms (the functions $b_\sigma(\mathbf{n}, t)$ and $d_\sigma(\mathbf{n}, t)$) are complex: the dependence of these terms on an organism's underlying genotype σ must capture how different DNA and protein sequences affect the complex molecular machinery that enables cells to reproduce and interact with their environment. Understanding this relationship between genotype and **phenotype** — an organismic-scale property such as birth or death rate — is a major challenge facing modern molecular biology and a primary focus of this work. This particular relationship between genotype and reproductive success is typically understood through the notion of a fitness landscape, which we explore in the next section.

1.3 Fitness landscapes

The **fitness** of a genotype σ is an abstract quantity [13, 22–24] whose precise definition varies across models and experiments, but in general it is some number that characterizes the reproductive success of an organism with that genotype. Most notions of fitness fall into two classes. The first class can be thought of as "probabilistic" fitness, since it can usually be interpreted as the relative probability of an organism to reproduce [19]. We will mainly use this type of fitness in this work and denote it as $\mathcal{F}$. Another notion of fitness is "Malthusian" fitness, which satisfies an operational definition as the relative exponential growth rate of a population with that genotype. In most models Malthusian fitness is simply the logarithm of the probabilistic fitness ($\log \mathcal{F}$). Moreover, most models depend only on relative fitness values, i.e., ratios of probabilistic fitnesses or differences of Malthusian fitnesses (equivalent under the logarithmic mapping). In particular, the selection coefficient as previously defined corresponds to a difference of Malthusian fitnesses.

The mapping of all possible genotypes σ to their fitness values $\mathcal{F}(\sigma)$ is known as a genotypic **fitness landscape** [22]. We imagine a population as a distribution of genotypes on this fitness landscape. Natural selection tends to drive the population toward higher fitness values, while mutation, recombination, and genetic drift tend to randomize the population. This is analogous to a distribution of particles obeying Brownian dynamics on a potential energy landscape, where particles experience both a gradient force toward lower energies along with stochastic fluctuations. Thus, although other forces shape the evolution of a population as well, the fitness landscape plays a major role. It is therefore critical to understand its structure over the space of all genotypes.

1.3.1 Global landscape properties

Figure 1.2 shows schematic landscapes illustrating some key global properties. Perhaps the most salient feature of a landscape is the number of maxima: this is defined as the number of genotypes where all their mutational neighbors have lower fitness. This property therefore depends on the available mutational moves an organism can make. For example, a genotype may be a local fitness maximum with respect to all mutations of single positions in its sequence, but it may not be a maximum if recombinations or insertions are possible. In general, a greater number of possible moves increases the connectivity of the genotype space and will reduce the number of local fitness maxima. The schematic landscapes in Fig. 1.2 should thus not be taken too seriously, since their two-dimensional structure belies the high-dimensional space of genotypes.

In Fig. 1.2A we show a landscape with a single global maximum; we therefore expect populations to tend toward genotypes near this point. In contrast, Fig. 1.2B shows a landscape with multiple local maxima. In this case populations may experience divergent outcomes if they end up at different maxima. This has important consequences for the

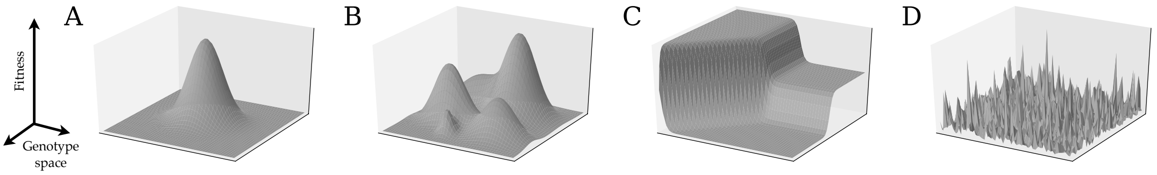


Figure 1.2: **Global properties of fitness landscapes.** Schematic fitness landscapes over genotype space: (A) smooth landscape with a single maximum, (B) landscape with multiple local maxima, (C) landscape with large flat (neutral) regions, (D) highly rugged (epistatic) landscape. Note that real genotype spaces are very high dimensional compared to these two-dimensional schematics. All plots were made with `matplotlib` [25].

repeatability of evolution, as discussed in Sec. 1.4.2.

Besides having a well-defined set of maxima, another possibility for global landscape structure is a series of large flat regions of genotypes with the same or similar fitness values. These regions are known as **neutral networks** [26, 27]; "neutral" indicates that genotypes in these regions are equivalent under selection. Figure 1.2C provides an example of such a landscape. Populations may have very different dynamics on landscapes with neutral regions compared to landscapes with well-defined maxima; they may spend long times wandering flat areas until they reach gradients that drive them up toward a region of higher fitness.

1.3.2 Ruggedness and epistasis

The first three landscape examples in Fig. 1.2 have different global features but locally are all rather smooth. Figure 1.2D in contrast shows a landscape with very rugged local structure, resulting in a large number of local minima and maxima. It is clear that a highly rugged landscape will result in very different evolutionary dynamics. Landscape ruggedness is equivalent to a genetics concept known as **epistasis**. Let the genotype σ be represented as $(\sigma_1, \sigma_2, \dots, \sigma_L)$, where σ_i is the letter at position $i \in \{1, \dots, L\}$. In general the probabilistic fitness function $\mathcal{F}(\sigma)$ cannot be decomposed into a product of independent contributions from each position i or, equivalently, the Malthusian fitness $\log \mathcal{F}(\sigma)$ cannot be decomposed into a sum. This means that the fitness effect of a mutation at a given position may depend on the state of other positions. If this is true, the positions will be correlated, which can

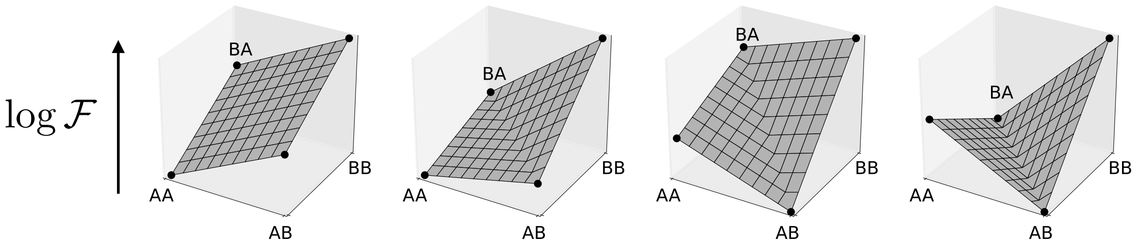


Figure 1.3: **Qualitative types of epistasis.** The four qualitative types of epistasis for a two-letter, two-position model. From left to right: *no epistasis*, where each mutation has the same effect on additive fitness ($\log \mathcal{F}$) regardless of the other position, yielding a linear landscape; *magnitude epistasis*, where the magnitude (but not the sign) of a mutation's additive fitness effect depends on the other position; *sign epistasis*, where the sign of a mutation's fitness effect (beneficial or deleterious) depends on the other position; *reciprocal sign epistasis*, where multiple instances of sign epistasis can lead to multiple local fitness maxima.

be thought of as a coupling between the positions. Mathematically this is reminiscent of a Hamiltonian for a system of interacting particles.

Epistasis is precisely this interactive coupling in the context of genotypic sequences. Following convention, we use Malthusian fitness here ($\log \mathcal{F}$), and categorize types of epistasis according to the qualitative differences in the additive fitness effects of mutations. We summarize the four possible cases using a two-letter, two-position model ($k = 2$, $L = 2$) in Fig. 1.3 in which sequence AA evolves into sequence BB, which has the highest fitness. In the first case on the left of Fig. 1.3, there is no epistasis: the fitness effect of the A → B substitution at position 2 is the same regardless of the state of position 1, and vice versa. Thus the Malthusian fitness can be decomposed into a sum of contributions from each site: $\log \mathcal{F}(\sigma) = \log \mathcal{F}_1(\sigma_1) + \log \mathcal{F}_2(\sigma_2)$, i.e., the Malthusian fitness landscape is linear in genotype space. A population should therefore be able to easily evolve from AA to the global maximum at BB.

In the second case of Fig. 1.3, the fitness effect of A → B at position 2 differs in magnitude but not in sign depending on whether position 1 has A or B. This situation is known as **magnitude epistasis**. Note that there are two kinds of magnitude epistasis: an "amplifying" or "super-additive" case in which the fitness benefit of a mutation is enhanced by the presence of other mutations (as shown in the second panel of Fig. 1.3), and the "diminishing returns" or "sub-additive" case in which the fitness benefit is decreased by other mutations.

The third case of Fig. 1.3 shows how the $\mathsf{A} \to \mathsf{B}$ substitution at position 2 can have opposite effects on fitness depending on the state of position 1: it is deleterious if $\sigma_1 = \mathsf{A}$, but beneficial if $\sigma_1 = \mathsf{B}$. Since the sign of the fitness effect depends on the other position, this situation is known as **sign epistasis**. It is equivalent to frustration in the spin models of statistical physics. Sign epistasis can significantly affect accessibility of genotypes on the landscape, since the pathway $\mathsf{AA} \to \mathsf{AB} \to \mathsf{BB}$ requires a deleterious substitution. When sign epistasis exists at multiple positions, it is known as **reciprocal sign epistasis**, as shown in the fourth case of Fig. 1.3. Reciprocal sign epistasis is a necessary condition for the existence of multiple local maxima [28] (cf. Fig. 1.2B,D). These cases straightforwardly generalize to higher-dimensional genotype spaces with additional positions and letters.

There is no universal measure to quantify the epistatic ruggedness of fitness landscapes [24]. One commonly-used measure is simply the number of local fitness maxima, since multiple local maxima are indicative of sign epistasis [28]. For binary alphabets ($k = 2$), deviations of the Malthusian fitness function from linearity can be quantified by fitting a linear function and calculating the sum of squares of residuals, known as a roughness parameter [24]. A third option is to consider all pairs of positions and all pairs of possible letters at those positions, and then classify the resulting sub-landscape for each combination like those shown in Fig. 1.3. Weinreich and coworkers have more recently proposed a scheme for quantifying higher-order epistasis that goes beyond the pairwise couplings described here [29].

1.3.3 Empirical landscape reconstructions

What is the structure of fitness landscapes for real organisms? Explicit reconstruction of a fitness landscape entails engineering organisms with each genotype and measuring their relative fitnesses. Although great progress has been made in recent years (see Ref. 24 for a review), the enormous number of possible genotypes limits this to a very small number of positions (typically 4–9) and possible mutations (typically binary at each position). Thus fitness measurements are usually only obtained for a few tens or hundreds of genotypes. In addition, because individual genotype survivability is not directly accessible in experiments, proxy measures of fitness are employed, such as growth rates and antibiotic resistance.

Many studies have attempted to characterize these empirical landscapes in terms of their epistatic features, accessibility to adaptation, and correspondence to theoretical models [24, 30–37]. For example, magnitude epistasis appears to be widespread [36–39], while substantial sign epistasis appears in fewer examples [31, 32]. The emerging picture supports a mostly smooth but nonlinear landscape, with limited sign epistasis and therefore a single or small number of maxima [24, 33, 40]. However, the limited nature of these reconstructions requires a cautious interpretation. While these experiments may provide insight into local landscape properties such as epistasis, a clear picture of global properties requires much more exhaustive sampling of the landscape.

1.3.4 Simple model landscapes

The difficulty of experimentally reconstructing landscapes has spurred the development of theoretical models. A few simple models have traditionally dominated the field; they generally consider sequences with binary alphabets ($k = 2$), in which case genotype space is a unit hypercube. In Kauffman's NK (or "LK") model [34, 41, 42], each of the L positions in the gene (or genes in the genome) interacts with K other sites chosen by random sampling. The Malthusian fitness of genotype σ is given by

$$\log \mathcal{F}(\sigma) = \sum_{i=1}^{L} \log \mathcal{F}_i(\sigma_i, \sigma_{n_1(i)}, \ldots, \sigma_{n_K(i)}), \tag{1.2}$$

where $n_1(i), \ldots, n_K(i)$ are the randomly-chosen interaction partners of position i. The single-site fitnesses $\mathcal{F}_i$ are obtained by sampling from a continuous distribution; each combination of 2^{K+1} possible states of the argument corresponds to an independent sampling. When $K = 0$, positions do not interact and the landscape is non-epistatic. Because in this limit the landscape is smooth and has a single maximum, it is sometimes called the "Mount Fuji" model [43]. The amount of epistasis can be tuned by increasing K to the maximum value of $L - 1$. With $K = L - 1$, the fitnesses of different genotypes are uncorrelated; this model is called the "House of Cards" due to the unpredictable fitness effects of mutations [44]. Realistically, closely-related genotypes should have correlated fitnesses, so this limit serves mainly as a null model. Many properties of the NK landscape are known [41, 45–48]; for example, in the $K = L - 1$ limit the average number of local maxima is $k^L/(L(k-1)+1)$ [24].

Another class of models starts from a non-epistatic (Mount Fuji) landscape and adds a tunable amount of noise to it. For example, in the "rough Mount Fuji" model [43], genotype σ_0 is arbitrarily selected as the global maximum and the fitness of genotype σ is given by

$$\log \mathcal{F}(\sigma) = -\theta d(\sigma, \sigma_0) + \eta(\sigma), \tag{1.3}$$

where $d(\sigma, \sigma_0)$ is the **Hamming distance** (number of positions that are different) between genotypes σ and σ_0, θ is the average additive fitness cost of each deleterious mutation, and $\eta(\sigma)$ is a zero-mean random variable sampled independently for each sequence σ. The ruggedness of the landscape is controlled by the ratio of θ to the standard deviation of the distribution from which the random variables $\eta(\sigma)$ are sampled [24].

Another class of models aims to construct landscapes with neutral networks (e.g., Fig. 1.2C). These are inspired by observations that most mutations either have no fitness effect or are strongly deleterious and thus rapidly removed from the population [15]. One way to construct neutral networks is equivalent to percolation models in physics [49]: each genotype is randomly assigned fitness 1 with probability p and fitness zero with probability $1 - p$ [34]. The neutral network is then the connected subset of viable genotypes.

1.3.5 *Ab initio* biophysical fitness landscapes

Although the aforementioned models have produced a wealth of theoretical results and have facilitated some analysis of data, their purely phenomenological nature allows for little interpretation of their parameters and includes no basis in the underlying molecular processes governing cells. From a physical perspective, a major weakness of attempting to directly construct the genotypic fitness landscape $\mathcal{F}(\sigma)$, either experimentally or theoretically, is that it relates properties of organisms at very different scales: the genotype σ is essentially a "microscopic" property of individual DNA or protein molecules, while fitness is a "macroscopic" property of a whole organism. Moreover, macroscopic properties of an organism do not directly depend on microscopic details of genotype anyway; fitness depends on the properties of at least whole molecules like proteins, and more likely on whole collections of molecules (e.g., forming metabolic or signaling networks).

This suggests that instead of directly measuring or postulating a genotypic fitness landscape, a better approach is to bridge the length scales via intermediate molecular phenotypes that are closer in scale to genotypes, and therefore easier to relate, and are also more closely tied to the actual reproductive ability of the whole organism. Besides its practical advantages, this approach has the conceptual advantages of incorporating explicit physico-chemical properties of molecules into evolutionary models, which enables more precise experiments and unambiguous interpretation of model parameters. Building *ab initio* models of evolution based on biophysics has been explored in several contexts over the last decade [50, 51]. It has been extensively studied for individual proteins [10–12, 27, 35, 52–63] and protein interaction networks [64–66], as well as for DNA regulatory sites [9, 67–72] and small molecular networks [73–77]. This approach is a central focus of this work, and thus we will discuss it further in subsequent chapters, especially Chapters 3 and 5.

1.4 Pathways of evolution

We now turn to considering how fitness landscapes affect evolutionary dynamics of populations. Of particular interest are evolutionary paths of populations, which are essential for understanding the history of natural populations as well as predicting future evolutionary outcomes. We first define a simple model of evolutionary dynamics to make precise the notion of evolutionary paths and discuss specific properties of interest.

1.4.1 Simple model of evolutionary dynamics

In general, a population will occupy a distribution of genotypes (see the model in Sec. 1.2). However, under certain conditions a population may be genetically homogeneous, or **monomorphic**, most of the time. This occurs when new mutations are sufficiently rare that when a new mutation arises, it is likely to either fix or go extinct before a second mutation arises. Thus the population is always monomorphic except for these relatively short transition periods.

A simple argument shows under which conditions this occurs. Let u be the mutation rate, defined as the probability of mutation per position per generation. Assuming a genotype of length L and a population of size N, the average time until a mutation arises at any position in any organism is of order $(LNu)^{-1}$. We want to compare this to the average time for a single mutant to fix or go extinct. For a selectively-neutral mutant, fixation and extinction have conditional average times of order N and $\log N$ [78, 79], respectively, with probabilities $1/N$ and $(N-1)/N$ [20, 80]. To leading order in N, the average time until *either* fixation or extinction of a mutant is thus $\log N$. The monomorphic limit therefore holds if this time is much shorter than the mutation waiting time:

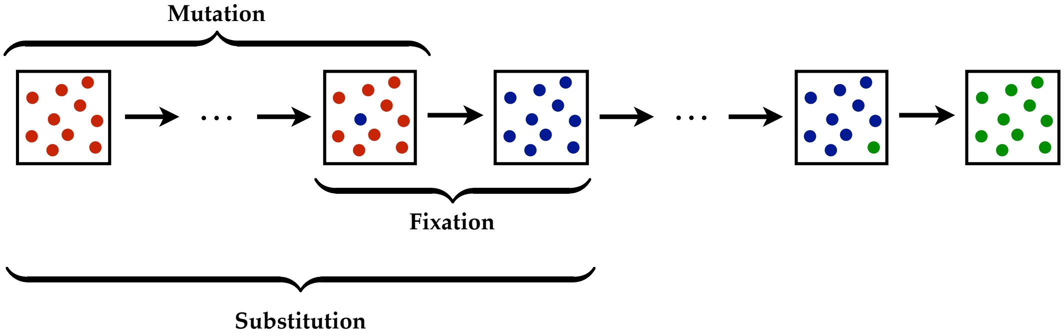


Figure 1.4: **Substitution process of a monomorphic population.** A monomorphic population evolves by sequential substitutions, which consist of a single mutation arising followed by fixation of the mutant. In the figure the population makes one substitution from red to blue and then another from blue to green.

$$\log N \ll \frac{1}{LNu}, \tag{1.4}$$

or, equivalently, $u \ll (LN \log N)^{-1}$ [81, 82]. The monomorphic limit thus requires sufficiently low mutation rates and small population sizes. This is believed to hold for many natural populations, especially among eukaryotes [15, 83].

Monomorphic populations have dramatically simpler evolutionary dynamics. We can describe their state at any given time by a single genotype σ (rather than a whole list of occupation numbers). Consider a vector space with basis elements $\{|\sigma\rangle\}$, one for each possible genotype. The probability distribution of an ensemble of populations can be expressed as a vector $|\pi(t)\rangle$ in this space, where the components $\langle\sigma|\pi(t)\rangle = \pi(\sigma, t)$ give the probability that a population has genotype σ at time t. This is normalized so that $\sum_\sigma \pi(\sigma, t) = 1$. The population changes genotype when a mutation occurs producing a new genotype σ', which then fixes in a population of σ; this combined process of mutation and fixation is known as **substitution** (see Fig. 1.4). We can define a substitution rate matrix $\mathbf{W}$ with elements [15]

$$\langle\sigma'|\mathbf{W}|\sigma\rangle = Nu(\sigma'|\sigma)\ \phi(\sigma'|\sigma), \tag{1.5}$$

where $u(\sigma'|\sigma)$ is the rate of genotype σ mutating into σ' and $\phi(\sigma'|\sigma)$ is the fixation probability of a single σ' mutant in a population of σ. The substitution rate is thus the rate of producing a single mutant times the probability that the mutant fixes. The probability distribution of populations therefore obeys the master equation

$$\frac{d}{dt}\pi(\sigma', t) = \sum_{\sigma \neq \sigma'} [\langle \sigma' | \mathbf{W} | \sigma \rangle \; \pi(\sigma, t) - \langle \sigma | \mathbf{W} | \sigma' \rangle \; \pi(\sigma', t)]. \tag{1.6}$$

If we define the diagonal elements of $\mathbf{W}$ as the negative of the total substitution rate out of that genotype

$$\langle \sigma | \mathbf{W} | \sigma \rangle = - \sum_{\sigma' \neq \sigma} \langle \sigma' | \mathbf{W} | \sigma \rangle, \tag{1.7}$$

then we can also write the master equation in vector notation, mathematically equivalent to the Schrödinger equation:

$$\frac{d}{dt}|\pi(t)\rangle = \mathbf{W}|\pi(t)\rangle. \tag{1.8}$$

The general solution to this equation is therefore expressed as

$$|\pi(t)\rangle = e^{t\mathbf{W}}|\pi(0)\rangle, \tag{1.9}$$

where $|\pi(0)\rangle$ is the initial distribution over genotypes.

The exact form of the fixation probability $\phi(\sigma'|\sigma)$, and hence the substitution rates (through Eq. 1.5), depends on the underlying population dynamics (Eq. 1.1). In most models the fixation probability depends only on the relative fitness between the two genotypes. A convenient parameterization of relative fitness is the selection coefficient. For probabilistic fitness this is defined as

$$s = \frac{\mathcal{F}(\sigma')}{\mathcal{F}(\sigma)} - 1. \tag{1.10}$$

This ranges from $+\infty$ when $\mathcal{F}(\sigma')$ is infinitely larger than $\mathcal{F}(\sigma)$ (infinitely beneficial mutation) to -1 when $\mathcal{F}(\sigma')$ is infinitely smaller (infinitely deleterious). When the fitnesses are exactly equal, $s = 0$. Note that when $\mathcal{F}(\sigma')$ and $\mathcal{F}(\sigma)$ are both close to 1 (the scale-invariance of probabilistic fitness means it is always possible to scale at least one fitness value to 1), $s \approx \mathcal{F}(\sigma') - \mathcal{F}(\sigma)$.

In general we will assume $\phi(\sigma'|\sigma) = \phi(s)$. For example, the approximate fixation probability for the Wright-Fisher model of population dynamics is [17, 20, 80]

$$\phi_{\mathrm{WF}}(s) = \frac{1 - e^{-2s}}{1 - e^{-2Ns}}. \tag{1.11}$$

The Moran population model, on the other hand, yields

$$\phi_{\mathrm{Moran}}(s) = \frac{1 - (1+s)^{-1}}{1 - (1+s)^{-N}}. \tag{1.12}$$

We will examine the differences in these fixation dynamics in Chapter 2, but here we simply describe a few relevant limits. All forms of the fixation probability should satisfy $\lim_{s\to\infty} \phi(s) = 1$ (for an infinitely beneficial mutation) and $\lim_{s\to -1} \phi(s) = 0$ (for an infinitely deleterious mutation). They should also satisfy $\phi(0) = 1/N$ as required by symmetry of selectively-neutral genotypes. If $|s| \gg 1$ for all relevant mutations (fast selection), all beneficial mutations are essentially guaranteed to fix, while deleterious ones are guaranteed to be eliminated. Similar to zero-temperature Monte Carlo, the population can only undergo substitutions that increase fitness, and all allowed substitutions occur with the same rate Nu. When selection is slow compared to generation times but fast compared to genetic drift ($1 \gg |s| \gg N^{-1}$), $\phi(s) \approx s$ for $s > 0$ and $\phi(s) \approx 0$ for $s < 0$ [19]. Thus deleterious mutations always get eliminated as before, but beneficial mutations fix at the rate Nus [84].

1.4.2 The ensemble of evolutionary paths and the predictability of evolution

The simple dynamics presented above enable a precise definition of an evolutionary path. Since a monomorphic population has only one genotype at a time, we can define the **evolutionary path** $\varphi = \{\sigma_0, \sigma_1, \ldots\}$ as a time-ordered sequence of genotypes through which the population passes. Since the precise pathway taken by a population is stochastically determined, we are usually interested in an ensemble of paths. For example, we may consider the ensemble of paths from an ancestral state of the population to some final state (e.g., at a fitness maximum). There are many properties of this path ensemble relevant to evolutionary theory, including the distribution of path lengths, which indicates the number of substitutions in the genetic sequence, and the distribution of path times, which indicates the speed of evolution. We may also want to know how the paths are distributed across genotype or phenotype space. For example, paths may all bottleneck through a particular state, or be tightly constrained to pass through a particular set of genotypes. Underlying the importance of many of these quantities is the notion of evolutionary predictability.

The **predictability of evolution** is a question of paramount importance in biology [85]. If “life’s tape” could be replayed, would we see a completely different outcome because evolution is a largely stochastic phenomenon, or are accessible evolutionary paths so constrained that the outcome would have been the same or recognizably similar [14]? This question is not only of theoretical value in understanding the history of natural populations, but has practical relevance in predicting the future evolutionary outcomes of populations of endangered species and infectious diseases. Moreover, the growth of controlled, highly-parallelized evolution experiments on microbes enables quantitative tests of the concept in the laboratory [86–88].

There are two major aspects of predictability. The first aspect is the predictability of intermediate pathways between a fixed initial state and a fixed final state. Here the issue is assessing the diversity of the relevant path ensemble; if only a single path is available to a population, then obviously evolution is perfectly predictable, while the existence of many equivalent paths means evolution will be more unpredictable. One assessment of this diversity is simply counting the number of such paths. For example, Weinreich et al. [31] found that only a small fraction of all possible paths from wild-type *E. coli* to a strain resistant to antibiotics was accessible to adaptation. In another approach, Lobkovsky et al. [35] and Lobkovsky and Koonin [85] devised a measure called mean path divergence:

$$\mathcal{D} = \sum_{\varphi_1 \neq \varphi_2} \Delta(\varphi_1, \varphi_2) p(\varphi_1) p(\varphi_2), \tag{1.13}$$

where the sum is over all pairs of distinct paths φ_1 and φ_2 in an ensemble, $p(\varphi)$ is the probability of path φ, and $\Delta(\varphi_1, \varphi_2)$ is the path distance between φ_1 and φ_2. The path distance is defined as the average of the shortest Hamming distances between each genotype σ_1 on path φ_1 and all genotypes on path φ_2, and vice versa:

$$\Delta(\varphi_1, \varphi_2) = \frac{1}{\mathcal{L}[\varphi_1] + \mathcal{L}[\varphi_2]} \left(\sum_{\sigma_1 \in \varphi_1} \operatorname*{argmin}_{\sigma_2 \in \varphi_2} d(\sigma_1, \sigma_2) + \sum_{\sigma_2 \in \varphi_2} \operatorname*{argmin}_{\sigma_1 \in \varphi_1} d(\sigma_2, \sigma_1) \right), \tag{1.14}$$

where $\mathcal{L}[\varphi]$ is the length (number of substitutions) of path φ and $d(\sigma_1, \sigma_2)$ is the Hamming distance between σ_1 and σ_2. The divergence therefore captures not only how many paths are available, but weighs them by their proximity in genotype space.

Besides the diversity of intermediate pathways between fixed endpoints, the second aspect of predictability is the diversity of final states themselves. Can a population end up in very different genotypic or phenotypic states under repeated trials, regardless of the intermediate pathways? In the strong-selection limit, where deleterious and neutral substitutions never occur, local maxima on the fitness landscape serve as absorbing states for the population, so they serve as the only possible final states for the population over sufficiently long times. However, the availability of neutral and deleterious substitutions complicates this picture.

Although the concept of predictability has a long history in evolutionary theory, surprisingly little attention has been paid to it in the context of theoretical models, outside of the few examples cited here. In Chapters 4 and 5 we will present more systematic methods to quantify different aspects of evolutionary predictability, including the diversity of both intermediate paths and final states. Our approach is drawn from statistical physics and is easily applied to a wide range of evolutionary models.

Chapter 2

Time Reversibility and Evolutionary Steady State under Selection

In Sec. 1.4.1 we defined a simple model of evolutionary dynamics, that of monomorphic populations undergoing substitutions on a fitness landscape. In this chapter we explore properties of time reversibility and steady state in this model, especially in the presence of strong natural selection. This chapter largely reproduces Ref. 8.

Much theoretical work in population genetics has focused on gradual models of adaptation in which evolutionary change proceeds through selection of mutations with very small fitness advantage [89]. The idea of the extremely slow rate of phenotypic evolution was proposed by Darwin [13] and subsequently made popular by Fisher [90] in the context of the infinitesimal model. In more recent decades, experimental evidence such as the molecular clock and high levels of sequence variation in some proteins suggested that genetic drift, and not selection, was the key evolutionary driving force. This led to the neutral and nearly neutral theories of molecular evolution [15, 91, 92].

From the theoretical perspective, a key motivation for weak-selection models is their **universality**: many specific models are equivalent in the weak-selection, or diffusion, regime. This equivalence is observed for the simple Wright-Fisher [90, 93] and Moran [94] models, which share a diffusion limit with a variety of more elaborate models under the appropriate mapping of parameters (e.g., [21, 95–100]). Even though the simple Wright-Fisher model is undoubtedly a gross simplification of natural populations, this universality has driven the use of its diffusion limit [80, 101], and more generally, the use of exchangeable models [102] as plausible effective theories in a wide variety of applications.

However, there is mounting experimental evidence that stronger selection may be common in nature. Strongly deleterious mutations have long been known to exist, although they are typically eliminated by selection so efficiently that they play little role in evolutionary dynamics [15]. Mutations with strong selective advantage, on the other hand, may routinely occur in organisms faced with novel environments or environmental stresses such as high temperature [103–106], with early steps in adaptation typically exhibiting larger fitness gains than later ones. Furthermore, several QTL-mapping experiments have demonstrated that adaptive evolution frequently involves relatively few genetic changes with large fitness effects [89, 107, 108]. Using approaches developed in the weak-selection limit to predict the dynamics of strongly beneficial mutations (such as fixation times and the probability of fixation) may lead to significant errors [100, 109, 110].

Models attempting to include a wider range of selection strengths are often deterministic [111, 112] and therefore exclude populations with non-negligible genetic drift, while stochastic theories typically demonstrate model-dependent behavior when selection becomes too strong [113–115], which limits their application to natural systems. Thus there is a need to study universal properties of classes of stochastic models in which no *a priori* assumptions are made about the strength of selection.

In this chapter we investigate such properties, focusing on time reversibility (i.e., detailed balance) and the steady state of the substitution process. We restrict ourselves to asexual haploids for simplicity, which includes many populations of single-cell organisms [116–119]. For *any* time-reversible population model, such as the Moran process, we show that the substitution rates obey a simple scaling law. This result is exact in the monomorphic limit and requires no diffusion or weak-selection approximation. For irreversible models, we find that the scaling law is an accurate approximation for sufficiently weak selection, and in fact may hold for a large range of selection strengths beyond the classical diffusion limit, as we show for the simple Wright-Fisher model and its extensions. Since this scaling behavior is equivalent to time reversibility, this contradicts the belief that selection should break reversibility [120].

The scaling law also gives rise to a power-law formula for the steady-state distribution, which is exact for any reversible model. This generalizes the work of Sella and Hirsh [121], who obtain this result in the special case of the Moran model. Moreover, we find that strong selection plays little role in steady state, which is dominated by genetic drift and weak selection. Since evolutionary behavior in this regime is known to be universal through established results based on the diffusion approximation, the steady-state formula is accurate within a sizable range of selection strengths for a large class of population models, including many irreversible ones. The wide range of applicability of the time-reversibility condition greatly simplifies computational studies of evolutionary dynamics in biological systems, such as probabilistic phylogenetic inference [122]. Finally, the simple power-law form of the steady-state distribution allows inference of fitness landscapes from genomic data in systems for which the steady state is believed to be a good approximation, such as TF binding sites in yeast [72] (explored further in Chapter 3).

2.1 The scaling law and steady state

We consider the evolution of a population in the monomorphic limit, as described in Sec. 1.4.1. The genotype σ under consideration defines either an entire genome, or more realistically, a single locus in a genome (such as a gene or small DNA regulatory site) that is unlinked to the rest of the genome (linkage equilibrium) by frequent recombination with rate r, which satisfies $r \gg LNu$ [123]; here, recombination also includes homologous DNA transfer such as observed in bacteria. Therefore we can consider the evolution of the locus independently from the rest of the genome. We assume that the locus is short enough that recombination does not occur within the locus itself. In general, we are interested in loci with $< 10^3$ nucleotides, which easily meet these conditions. Such loci include short regulatory sequences of nucleotides such as TF binding sites as well as coding regions. Viruses or loci with mutation or recombination hotspots are outside the scope of this model. Note that while the locus of interest is unlinked to other genomic sites, there may be epistasis among the nucleotides or amino acids constituting the locus itself.

The probability distribution for the population's genotype is defined by a state vector $|\pi(t)\rangle$ that satisfies the master equation in Eq. 1.8. This Markov process is finite and irreducible, since there is a nonzero probability of reaching any genotype from any other genotype in finite time. Hence it has a unique steady-state distribution $|\pi\rangle$ satisfying [124]

$$\mathbf{W}|\pi\rangle = 0, \tag{2.1}$$

or in terms of the components $\langle\sigma|\pi\rangle = \pi(\sigma)$,

$$\sum_{\sigma}[\langle\sigma'|\mathbf{W}|\sigma\rangle\ \pi(\sigma) - \langle\sigma|\mathbf{W}|\sigma'\rangle\ \pi(\sigma')] = 0. \tag{2.2}$$

For simplicity we drop the time dependence to indicate the steady state distribution.

The form of this steady-state distribution depends on the underlying population genetics model that gives the fixation probability ϕ and hence the substitution rates in Eq. 1.5. Following the discussion in Sec. 1.4.1, we will assume the fixation probability depends only on the relative fitness between the mutant and wild-type genotypes and the population size N. For convenience, here we will parameterize $\phi = \phi(r)$ in terms of the fitness ratio r rather than the selection coefficient s (cf. Eq. 1.10):

$$r = \frac{\mathcal{F}(\sigma')}{\mathcal{F}(\sigma)} = 1 + s. \tag{2.3}$$

We aim to use arguments from population genetics to study time reversibility (or simply "reversibility"), which in turn determines the form of the steady state. **Time reversibility** is equivalent to **detailed balance**, a sufficient but not necessary condition for steady state:

$$\langle\sigma'|\mathbf{W}|\sigma\rangle\ \pi(\sigma) = \langle\sigma|\mathbf{W}|\sigma'\rangle\ \pi(\sigma'). \tag{2.4}$$

The left- and right-hand sides of this equation are the steady-state probability currents $\sigma \to \sigma'$ and $\sigma' \to \sigma$, respectively. Equation 2.4 means that these currents are exactly balanced for each pair of genotypes σ and σ', and hence there are no net currents, consistent with the notion that it is impossible to distinguish the forward and backward flow of time in steady state.

Throughout this work, we will assume that neutral evolution — when all genotypes are selectively neutral relative to each other — is reversible. In the neutral model, the fixation probability $\phi(\sigma'|\sigma) = 1/N$ for all σ and σ', and hence Eq. 1.5 shows that the neutral substitution rates are just the mutation rates [15]: $\langle\sigma'|\mathbf{W}|\sigma\rangle = u(\sigma'|\sigma)$. Let the steady-state distribution of the neutral substitution process be $\pi_0(\sigma)$. Then reversibility of the neutral model is expressed by

$$u(\sigma'|\sigma)\ \pi_0(\sigma) = u(\sigma|\sigma')\ \pi_0(\sigma'). \tag{2.5}$$

Many popular neutral models are reversible (see Ref. 122 for a summary), but this is not guaranteed. This issue will be discussed further in Sec. 2.3.

We now consider the reversibility of the substitution rates under selection. Let us first define the function

$$\psi(r) \equiv \frac{\phi(r)}{\phi(1/r)}. \tag{2.6}$$

Hence the ratio of the forward and backward substitution rates between σ and σ' is

$$\frac{\langle\sigma'|\mathbf{W}|\sigma\rangle}{\langle\sigma|\mathbf{W}|\sigma'\rangle} = \frac{u(\sigma'|\sigma)}{u(\sigma|\sigma')} \cdot \frac{\phi\left(\frac{\mathcal{F}(\sigma')}{\mathcal{F}(\sigma)}\right)}{\phi\left(\frac{\mathcal{F}(\sigma)}{\mathcal{F}(\sigma')}\right)} = \frac{\pi_0(\sigma')}{\pi_0(\sigma)} \cdot \psi\left(\frac{\mathcal{F}(\sigma')}{\mathcal{F}(\sigma)}\right), \tag{2.7}$$

where we have invoked the reversibility of the neutral rates (Eq. 2.5). Studying the properties of the ψ function is the main focus of this chapter: it will determine the existence of reversibility under selection and the form of the steady-state distribution. We will investigate both its general properties and its form for specific models.

We will first *assume* that the substitution rates $\langle\sigma'|\mathbf{W}|\sigma\rangle$ under selection are reversible, which will completely constrain the form of ψ and the steady state $\pi(\sigma)$ under selection. In this case, $\langle\sigma'|\mathbf{W}|\sigma\rangle\ \pi(\sigma) = \langle\sigma|\mathbf{W}|\sigma'\rangle\ \pi(\sigma')$, and hence

$$\frac{\pi(\sigma')}{\pi(\sigma)} = \frac{\langle\sigma'|\mathbf{W}|\sigma\rangle}{\langle\sigma|\mathbf{W}|\sigma'\rangle} = \frac{\pi_0(\sigma')}{\pi_0(\sigma)} \cdot \psi\left(\frac{\mathcal{F}(\sigma')}{\mathcal{F}(\sigma)}\right). \tag{2.8}$$

It follows that

$$\psi\left(\frac{\mathcal{F}(\sigma'')}{\mathcal{F}(\sigma')}\right) \cdot \psi\left(\frac{\mathcal{F}(\sigma')}{\mathcal{F}(\sigma)}\right) = \psi\left(\frac{\mathcal{F}(\sigma'')}{\mathcal{F}(\sigma)}\right), \tag{2.9}$$

that is, ψ generally satisfies $\psi(r_1)\psi(r_2) = \psi(r_1 r_2)$. Therefore $\psi(r)$ must be a simple power law:

$$\psi(r) = r^{\nu}, \tag{2.10}$$

for some constant ν [125]. The constant ν can only depend on the population size N, since this is the only other parameter in our population model. We will refer to Eq. 2.10 as the **scaling law** for ψ. Using the definition of $\psi(r)$ (Eq. 2.6), one can show that

$$\nu = \frac{2\phi'(1)}{\phi(1)} = 2N\phi'(1), \tag{2.11}$$

where $\phi'(1) = d\phi(r)/dr|_{r=1}$ and $\phi(1) = 1/N$ is the neutral fixation probability.

Now rewriting Eq. 2.8 with our explicit form of ψ,

$$\frac{\pi(\sigma')}{\pi(\sigma)} = \frac{\pi_0(\sigma')}{\pi_0(\sigma)} \left(\frac{\mathcal{F}(\sigma')}{\mathcal{F}(\sigma)}\right)^{\nu}, \tag{2.12}$$

we can deduce the steady state:

$$\pi(\sigma) = \frac{1}{Z}\ \pi_0(\sigma)\ (\mathcal{F}(\sigma))^{\nu}, \tag{2.13}$$

where Z is a normalization constant. Note that Eq. 2.13 can be rewritten in the form of a Boltzmann distribution, with energy replaced by the negative logarithm of fitness:

$$\pi(\sigma) = \frac{1}{Z}\ \pi_0(\sigma)\ e^{\nu \log \mathcal{F}(\sigma)}. \tag{2.14}$$

The Boltzmann form in Eq. 2.14 suggests a straightforward analogy with statistical mechanics [121, 126]. One may think of the evolutionary model defined by Eqs. 1.5 and 1.6 as describing an ensemble of monomorphic populations taking random walks on a fitness landscape. The ensemble of walkers eventually reaches steady state in genotype space, which is given by Eq. 2.13 or 2.14. Populations will be driven toward the peaks of the landscape by selection, which manifests itself as the $\mathcal{F}^{\nu}$ factor in the steady state; this effect becomes exponentially stronger as ν increases. This is analogous to energy minimization in statistical mechanics. However, as in statistical mechanics, we also expect the entropy of states to affect the steady-state distribution, since typically there are few states with optimal or near-optimal fitness and many states with low fitness. This density of states is given by the neutral distribution π_0. The corresponding entropy (defined as $\log \pi_0$) competes with selection the same way energy and entropy compete in statistical mechanics: selection favors high fitness states while entropy favors low fitness states since there are usually many more of them. These competing forces reach some balance in the form of a "free fitness" function that is maximized in the steady state, as described in Refs. 121, 126.

This steady-state formula was derived in the special case of the Moran model by Sella and Hirsh [121]. We generalize this earlier result by showing that *any* reversible substitution process leads to the scaling law for ψ and the steady-state formula of Eq. 2.13. Note that this conclusion, obtained in the monomorphic limit, requires no additional assumptions, such as the weak-selection diffusion approximation.

We have shown that reversibility implies this scaling law. We now show that the scaling law implies reversibility: we now assume Eq. 2.10 without assuming reversibility. Then

$$\frac{\langle\sigma'|\mathbf{W}|\sigma\rangle}{\langle\sigma|\mathbf{W}|\sigma'\rangle} = \frac{\pi_0(\sigma')}{\pi_0(\sigma)} \left(\frac{\mathcal{F}(\sigma')}{\mathcal{F}(\sigma)}\right)^{\nu}. \tag{2.15}$$

We can combine this with the steady-state condition (Eq. 2.2) to show that

$$\begin{aligned} 0 &= \sum_{\sigma}[\langle\sigma'|\mathbf{W}|\sigma\rangle\ \pi(\sigma) - \langle\sigma|\mathbf{W}|\sigma'\rangle\ \pi(\sigma')] \\ &= \sum_{\sigma}\langle\sigma|\mathbf{W}|\sigma'\rangle \left[\frac{\pi_0(\sigma')}{\pi_0(\sigma)} \left(\frac{\mathcal{F}(\sigma')}{\mathcal{F}(\sigma)}\right)^{\nu}\ \pi(\sigma) - \pi(\sigma')\right]. \end{aligned} \tag{2.16}$$

Clearly the distribution in Eq. 2.13 satisfies this condition, so it must be the unique steady state. The reversibility condition (Eq. 2.4) is satisfied as well, and thus the scaling law implies reversibility.

Therefore, time reversibility and the scaling behavior of ψ are mathematically equivalent, and both lead to the steady-state formula of Eq. 2.13. This means that we can concentrate our attention on determining the form of ψ, since its scaling behavior tells us the extent to which reversibility and Eq. 2.13 hold. Obviously not all models are reversible, so the scaling law will not hold exactly in those cases. However, we demonstrate below that the scaling behavior of ψ is at least an approximate feature of a large class of models, and therefore reversibility and the steady-state formula (Eq. 2.13) provide a good approximation within a sizable range of selection strengths.

Since it will be more convenient to describe the scaling behavior of ψ on logarithmic scales, we expand $\log\psi(r)$ in a power series in $\log r$ around the neutral limit ($\log r = 0$):

$$\log\psi(r) = \sum_{j=0}^{\infty} \frac{c_{2j+1}}{(2j+1)!}\ (\log r)^{2j+1} = c_1(\log r)\left[1 + \frac{1}{c_1}\sum_{j=1}^{\infty}\frac{c_{2j+1}}{(2j+1)!}\ (\log r)^{2j}\right], \tag{2.17}$$

where

$$c_i = \left(\frac{d^i}{d(\log r)^i} \log \psi(r) \right) \Bigg|_{r=1} . \tag{2.18}$$

Note that $\log \psi(r)$ is an odd function in $\log r$ (deduced from the definition in Eq. 2.6), and hence there are only odd powers in the expansion. Since $c_1 = 2\phi'(1)/\phi(1) = \nu$, we can write

$$\log \psi(r) = \nu(\log r) \left[1 + \frac{1}{\nu} \sum_{j=1}^{\infty} \frac{c_{2j+1}}{(2j+1)!} (\log r)^{2j} \right] . \tag{2.19}$$

The scaling behavior of ψ is captured by the first-order term in this expansion. As long as ν is nonzero, there will always be some neighborhood of selection strengths around the neutral limit, $r = 1$, in which the scaling law holds. We give an argument that $\nu \neq 0$ in Appendix A.1. The argument relies on the universal nature of the diffusion approximation to a population model. That is, discrete population models can be approximated by a continuous diffusion equation, and it is known that a large class of population models are equivalent under this approximation (e.g., [21, 95–100]). The diffusion approximation is valid for weak-selection strengths: $r - 1 = s \sim \mathcal{O}(N^{-1})$ [20]. Since the scaling behavior of ψ appears in the diffusion regime, it is shared by a large class of models.

The diffusion argument in Appendix A.1 also gives us insight into the interpretation of $\nu = 2N\phi'(1)$: it suggests that $\phi'(1) \sim \mathcal{O}(N^0)$ and therefore $\nu \sim \mathcal{O}(N)$. Thus we can interpret ν as a "scaling" effective population size that is of the same order as the census population size for fixed-size models or the variance effective population size for more general models. This is sensible in light of the Boltzmann form of the steady state (Eq. 2.14), which suggests that $1/\nu$ plays the role of temperature, i.e., the scale of stochastic fluctuations.

There is a range of selection strengths in which the scaling law is approximately valid. Specifically, we wish to find the range of fitness ratios r, which we will denote as (r_0^{-1}, r_0) with $r_0 > 1$, such that

$$\nu(1 \mp \epsilon)\log r < \log\psi(r) < \nu(1 \pm \epsilon)\log r, \tag{2.20}$$

where the upper signs are valid for $r > 1$, the lower signs are valid for $r < 1$, and $\epsilon > 0$ is a small number that we choose to control the accuracy of the power law approximation. This range is determined by the next coefficient in the expansion of Eq. 2.19,

$$\frac{c_3}{6\nu} = \frac{\nu^3 - 3\nu^2 N^2 + 2\nu N^3\left(N - 3\phi''(1)\right) + 4N^5\left(3\phi''(1) + \phi^{(3)}(1)\right)}{12\nu N^6}, \tag{2.21}$$

where we have evaluated the derivative of $\log\psi(r)$ in terms of $\phi(r)$ and substituted $\phi(1) = 1/N$ and $\nu = 2N\phi'(1)$. For small ϵ,

$$\frac{|c_3|}{6\nu}(\log r_0)^2 = \epsilon \quad \Longrightarrow \quad r_0 = \exp\left(\sqrt{\frac{6\nu\epsilon}{|c_3|}}\right). \tag{2.22}$$

For any particular model, we need only calculate ν and c_3 to obtain the range of selection strengths (r_0^{-1}, r_0) for which the scaling law is a good approximation.

Even outside of this range, however, deviations from the power law likely lead to negligible errors in estimating the probabilities of extremely unfit genotypes. This is a situation encountered when the monomorphic population is in steady state on the fitness landscape, with the majority of time spent in locally optimal high-fitness states from which many strongly deleterious but no strongly beneficial substitutions can be made. Specifically, assume that the range of fitness ratios for which the scaling-law approximation is valid, calculated from Eq. 2.22, is (r_0^{-1}, r_0). Suppose that genotype σ_1 has fitness $\mathcal{F}_1$ and genotype σ_2 has fitness less than $\mathcal{F}_1/r_0$ ($r_0 > 1$), and also assume that they are separated by a single mutation. By construction, the substitution from σ_1 to σ_2 is outside the range for which the power law is a valid approximation. Now suppose that there is a third genotype σ_3 (also separated by a single mutation from σ_1) with fitness of exactly $\mathcal{F}_1/r_0$, so that its probability is given by Eq. 2.13. Since ψ must be monotonically increasing, the probability of the unfit σ_2 is bounded from above by the probability of σ_3:

$$\pi(\sigma_2) < \frac{1}{Z}\pi_0(\sigma_3)\ r_0^{-\nu}\mathcal{F}_1^{\nu}. \tag{2.23}$$

Then the ratio of $\pi(\sigma_2)$ to $\pi(\sigma_1)$ has an upper bound as well:

$$\frac{\pi(\sigma_2)}{\pi(\sigma_1)} < \frac{\pi_0(\sigma_3)\ r_0^{-\nu}\mathcal{F}_1^{\nu}}{\pi_0(\sigma_1)\ \mathcal{F}_1^{\nu}} \simeq r_0^{-\nu}, \tag{2.24}$$

where the last relation holds because the neutral probabilities $\pi_0(\sigma_1)$ and $\pi_0(\sigma_3)$ are of the same order of magnitude (under the reasonable assumption that mutation rates within the locus are all of the same order). Since ν is proportional to the population size, the maximum fitness ratio r_0 in the scaling region need not be very large to generate an enormous suppression of the unfit genotype in steady state. Thus inaccuracies in the probabilities of unfit genotypes caused by deviations from the scaling law will be negligible for all practical purposes.

Furthermore, we can explicitly show that the selection strengths of the dominant substitutions in steady state are precisely those described by the diffusion approximation. In steady state, it is sufficient to consider genotypes that have relative probabilities, with respect to the most fit genotype, of at least $\delta > 0$. Then the relevant fitness ratios r are constrained by $r^{-\nu} > \delta$ or $r < \delta^{-1/\nu}$. Since $\nu \sim \mathcal{O}(N)$, we expand in powers of $1/\nu$ to obtain

$$r < 1 - \frac{1}{\nu}\log\delta + \mathcal{O}(\nu^{-2}). \tag{2.25}$$

In terms of $s = r - 1$, this implies $s \sim \mathcal{O}(\nu^{-1}) \sim \mathcal{O}(N^{-1})$, which is the selection strength for which the diffusion approximation is valid [20]. Therefore the steady state of substitutions is adequately described by the diffusion approximation and thus by the scaling law (Eqs. 2.10 and 2.13). As a result, only the optimal genotype and slightly less fit neighboring states have non-negligible probabilities in steady state.

The steady-state distribution of Eq. 2.13 was previously derived for the special cases of the Moran process by Sella and Hirsh [121] and for the diffusion limit of the Wright-Fisher model by Sella and Hirsh [121], Lässig [71], and Li [127], among others. Indeed, some form of this formula can even be found in work by Wright [93]. We have generalized these results by showing that the steady-state formula holds exactly for *any* reversible model, not just the Moran process, without requiring any diffusion approximation. For irreversible models, we have shown how this result arises as an approximation, and determined its range of validity. Surprisingly, weak selection dominates steady-state behavior in a wide class of population models, justifying application of the steady-state formula to systems which may include mutations with large fitness effects.

2.2 Scaling for specific population models

We now verify the general results of the previous section for specific models, computing the scaling effective population size ν and the range of selection strengths for which the scaling law is a good approximation.

2.2.1 The Moran model

Consider a haploid population of fixed size N with two genotypes, A and B, and let n denote the number of organisms with genotype B. The single time-step transition probabilities of the Moran model are then [20, 94]

$$\begin{aligned}\langle n+1|\mathbf{P}|n\rangle &= \frac{\mathcal{F}_{\mathsf{B}}}{\bar{\mathcal{F}}}\frac{n}{N}\left(1-\frac{n}{N}\right)\\ \langle n-1|\mathbf{P}|n\rangle &= \frac{\mathcal{F}_{\mathsf{A}}}{\bar{\mathcal{F}}}\frac{n}{N}\left(1-\frac{n}{N}\right)\\ \langle n|\mathbf{P}|n\rangle &= 1-\langle n+1|\mathbf{P}|n\rangle-\langle n-1|\mathbf{P}|n\rangle,\end{aligned} \tag{2.26}$$

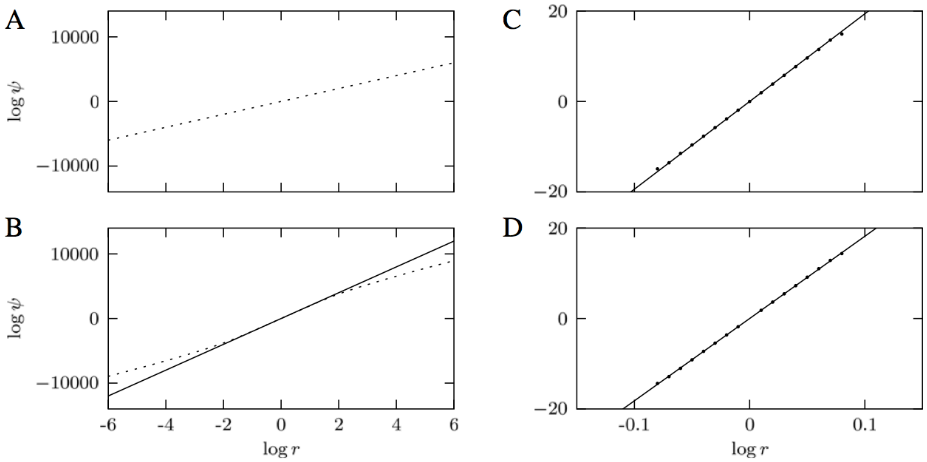


Figure 2.1: **The scaling law for different population models.** We show $\log \psi(r)$ as a function of $\log r$ for four population models. The scaling law appears as the straight line $\log \psi(r) = \nu \log r$. (A) The Moran model with $N = 1000$. Here the scaling law is exact with $\nu = N - 1$. (B) The simple Wright-Fisher model for $N = 1000$, calculated using the numerical procedure from Appendix A.2. The numerical calculation is the dashed line and the scaling-law prediction is the solid line. Here the scaling law is not exact but holds as a good approximation for a large range of selection strengths. The scaling effective population size is $\nu = 2(N - 1)$. (C) A modified Wright-Fisher model with population size N that varies sinusoidally as in Eq. 2.33, with $N_0 = 100$, $\alpha = 20$ and $T = 20$ generations. Simulation results are shown as dots and the scaling law as a solid line. The scaling law is an accurate approximation with $\nu = 2(N_e - 1)$, where $N_e = \sqrt{N_0^2 - \alpha^2}$ is the harmonic mean of the census population sizes. Because explicit simulations are required (as opposed to the numerical procedure used for the simple Wright-Fisher model), poor statistics on deleterious fixations and beneficial extinctions restricts us to considering smaller population sizes and range of selection strengths. (D) A model based on Ref. 128, where the mutant and wild-type may have different variances in offspring number in addition to different means. Here fitness is defined as $\mu - \sigma^2/N$, where μ is the average number of offspring and σ^2 is the variance. As in (C), we use $N = 100$ for numerical reasons. The scaling law is deduced by a linear fit.

where $\mathcal{F}_\mathsf{A}$ and $\mathcal{F}_\mathsf{B}$ are fitnesses of alleles A and B and $\bar{\mathcal{F}} = (n/N)\mathcal{F}_\mathsf{B} + (1 - n/N)\mathcal{F}_\mathsf{A}$ is the average fitness. In this case the probability of fixing a single mutant is [20]

$$\phi(r) = \frac{1 - r^{-1}}{1 - r^{-N}}, \tag{2.27}$$

where $r = \mathcal{F}_\mathsf{B}/\mathcal{F}_\mathsf{A}$. (This is equivalent to Eq. 1.12.) A straightforward calculation shows that $\psi(r) = \phi(r)/\phi(1/r) = r^{N-1}$ [121]. Hence $\nu = N - 1$ for the Moran model, and the scaling law holds exactly if the neutral substitution rates are reversible (Fig. 2.1A).

2.2.2 The Wright-Fisher model

We define the simple Wright-Fisher model [90, 93] analogously with the Moran case. Given that there are n organisms of genotype B in the current generation, the probability of having n' B organisms in the next generation is [20, 129]

$$\langle n'|\mathbf{P}|n\rangle = \binom{N}{n'} q^{n'}(1-q)^{N-n'}, \quad \text{where} \quad q \equiv \frac{n}{N}\frac{\mathcal{F}_{\mathsf{B}}}{\bar{\mathcal{F}}}. \tag{2.28}$$

Unlike the Moran model, the Wright-Fisher model is ill-suited to exact treatment, and hence the traditional approach to it has been the diffusion approximation. The diffusion theory yields many results in the neutral and weak-selection regimes [80, 101, 130], such as the formula for the fixation probability:

$$\phi(r) = \frac{1-e^{2(1-r)}}{1-e^{2N(1-r)}}, \tag{2.29}$$

where $r = \mathcal{F}_{\mathsf{B}}/\mathcal{F}_{\mathsf{A}}$ (equivalent to Eq. 1.11). However, there are two problems with the classical diffusion approach. The first is that the jump moment functions [131] $M(x, r)$ and $V(x, r)$ are typically expanded to the lowest order in $r - 1$ for the weak-selection regime (as in Appendix A.1), and so all subsequent calculations, including those leading to the fixation probability in Eq. 2.29, are not strictly valid for selection strengths beyond $s = r - 1 \sim \mathcal{O}(N^{-1})$. This expansion in selection strength, however, is not necessary, as it is possible to carry out the diffusion approximation using the exact moments derived from Eq. 2.28. This approach yields accurate results in the polymorphic limit, but fails to give an accurate formula for the fixation probability. This is due to the inherent breakdown of diffusion when the underlying discrete nature of the model becomes important, which is especially pronounced when selection effects are strong.

Since the diffusion approach is unsuitable to describe fixation outside of a fairly narrow range of selection strengths, we take a more accurate but numerical approach: calculating fixation probabilities directly from the discrete Markov chain defined in Eq. 2.28 (see

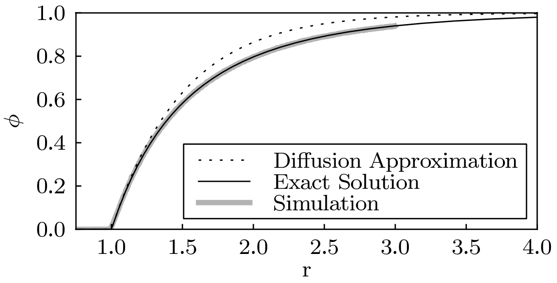


Figure 2.2: **Wright-Fisher fixation probability.** Plot of $\phi(r)$, the probability that a single mutant fixes as a function of its fitness ratio with the wild-type. For $N = 1000$, we compare an explicit simulation of the Wright-Fisher model with our discrete Markov chain approach (Eq. A.14) and Kimura's diffusion approximation (Eq. 2.29). The agreement between the discrete Markov chain and the simulation is excellent, in contrast with the noticeable disagreement between the simulation and the diffusion approximation at larger selection strengths.

Appendix A.2 for details). The end result is an efficient numerical procedure for accurate calculation of the fixation probability, and hence the ψ function, for any N and r. Figure 2.2 compares a simulation of $\phi(r)$ with this numerical approach along with the diffusion approximation (Eq. 2.29). The numerical calculation and the simulation match very well for all selection strengths, but there is noticeable disagreement with the diffusion result beyond the weak-selection regime.

Now we consider the expansion of $\psi(r)$ for the simple Wright-Fisher model. We know from diffusion theory that $\nu = 2N\phi'(1) = 2(N-1)$ [80]. Hence the expansion of $\psi(r)$ has the form

$$\log \psi(r) = 2(N-1)\log r + \mathcal{O}((\log r)^3). \tag{2.30}$$

Thus the power law and the steady state in Eq. 2.13 hold approximately with $\nu = 2(N-1)$. As Appendix A.2 shows, the form of the exact fixation probability is too complex to be useful for analytical calculations, such as computing c_3 in Eq. 2.21 to determine the range of selection strengths for which the power-law approximation is approximately valid. However, we can numerically compute this next-order coefficient for a range of N using the method in Appendix A.2 to obtain derivatives of fixation probabilities for Eq. 2.21. Figure 2.3 shows,

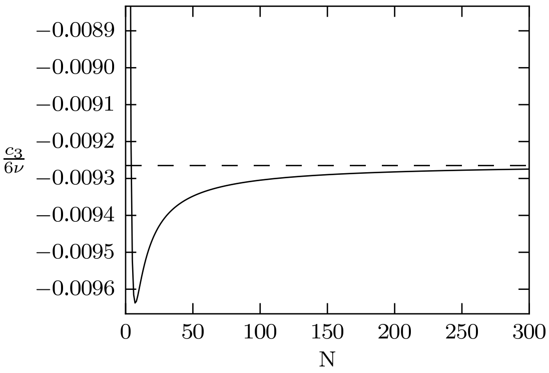


Figure 2.3: **Plot of ψ expansion coefficient.** Plot of $c_3/6\nu$ as a function of N for the simple Wright-Fisher model, obtained numerically from $\phi(r)$ using the procedure described in Appendix A.2. For realistic N values it rapidly converges to the constant ≈ -0.0093. This small value means that the scaling-law approximation is valid for a large range of selection strengths, and its N-independence means that this range does not shrink as N grows, contrary to the prediction of diffusion theory.

remarkably, that the next-order correction is independent of N for large N. Indeed, as N increases to realistic values the next-order coefficient rapidly converges to a small value of

$$\frac{c_3}{6\nu} \approx -0.0093. \tag{2.31}$$

Its smallness means that the scaling law is valid for a large range of selection strengths in the simple Wright-Fisher model. Indeed, for deviations from the power law of at most 5%, we set $\epsilon = 0.05$ in Eq. 2.22 and find that the fitness ratio r is constrained to be between 0.098 and 10.2. This corresponds to a selection coefficient s between -0.9 and 9.2, well beyond the typical weak-selection limits of $\pm\mathcal{O}(N^{-1})$. A numerical calculation of ψ confirms this large scaling region (Fig. 2.1B). Indeed, using the argument leading to Eq. 2.24, unfit genotypes that might exhibit deviations from the scaling law will be suppressed by at least a factor of $r_0^{-\nu}$, where (r_0^{-1}, r_0) is the range of fitness ratios for which the scaling law approximately holds. If we let $r_0 \approx 10.2$, even a very conservative $N = 200$ means that these unfit genotypes are suppressed by more than 10^{-402} relative to the most fit genotype.

The N-independence of $c_3/6\nu$ means that the size of the scaling region does not change with N. The standard diffusion approach implies a degeneracy of N and s: $Ns \sim \mathcal{O}(1)$, so that as N increases, the range of selection strengths that are considered weak shrinks. This is not intrinsic to the Wright-Fisher model, but is merely an emergent property in the diffusion limit [21]. Our result, however, shows that the scaling law is valid well beyond diffusion. In contrast, $c_3/6\nu$ calculated using Kimura's diffusion approximation (Eq. 2.29) is given by:

$$\frac{c_3}{6\nu} = -\frac{1}{6}N. \tag{2.32}$$

Since this coefficient grows with N, the scaling region for r shrinks as N increases. This is consistent with the selection-drift degeneracy predicted by diffusion, but it is clearly misleading in light of our analysis of the full Wright-Fisher model, since it would erroneously imply that the scaling law and reversibility hold for an extremely small range of selection strengths. This provides an example of the danger posed by extrapolating diffusion results to arbitrary regions of parameter space: the universality of the scaling law is much stronger than diffusion could predict. While this turns out to be unimportant for steady state, which is dominated by weak selection, the fact that reversibility approximately holds in systems with strong selection affects dynamical properties as well.

2.2.3 Other models

Models that share the diffusion limit with the Moran and Wright-Fisher models will also share the scaling law. This encompasses a wide class of exchangeable models [98, 99, 102]. For instance, many generalizations of the Wright-Fisher model with varying N are known to have properties equivalent to the simple Wright-Fisher model with some effective population size N_e [95, 97, 132]. Other generalizations, such as incorporating the effects of subdivided populations, also lead to equivalencies [96, 100].

As an example we consider the case when N varies periodically. For periods of oscillation smaller than fixation times, it is known that the Wright-Fisher diffusion results carry over with an effective population size N_e equal to the harmonic mean of the census population sizes [95, 97]. Let the transition probabilities be of the Wright-Fisher form (Eq. 2.28), with N changing over time according to

$$N(t) = N_0 + \alpha \sin\left(\frac{2\pi t}{T}\right), \tag{2.33}$$

where N_0 is the average size and T is the period of oscillation. The harmonic mean can be shown to be $N_e = \sqrt{N_0^2 - \alpha^2}$. In Fig. 2.1C, we use explicit simulations to compute $\psi(r)$, and we indeed find scaling behavior with $\nu = 2(N_e - 1)$. This slope, predicted through mapping to the simple Wright-Fisher model, is also obtained by a linear fit to the explicit simulation. Thus the scaling law still holds. For this model we do not have a numerical technique for fixation probabilities like the one used for the simple Wright-Fisher model (Appendix A.2), and explicit simulations prevent accurate statistics on fixation of very deleterious and extinction of very beneficial mutations, limiting us to a smaller range of selection strengths. Nevertheless, deviations beyond this smaller range can still be shown to be negligible in steady state. As Fig. 2.1C shows, the scaling region extends to at least $r_0 \approx 1.08$. Therefore any unfit genotypes leading to deviations must be suppressed by at least a factor of $r_0^{-\nu}$: even for $N_e = 200$, this is a suppression of 10^{-14}.

Other models beyond the paradigms of exchangeable and Wright-Fisher-type models may also demonstrate the scaling behavior. For instance, whereas Wright-Fisher and Moran models typically incorporate selective advantage as a difference in the mean number of offspring between genotypes, Gillespie proposed to incorporate stochasticity at the level of selection by allowing for different variances in offspring number [128, 133, 134]. In these models fitness is characterized by $\mu - \sigma^2/N$, where μ is the mean and σ^2 is the variance of the offspring number for a given allele. Other authors have extended models of this type to describe spatial variation, age structure, and demographic stochasticity, which may be important for small populations or populations subdivided into small demes [113–115].

Here, we simulate a model described by Gillespie [128]. Consider a haploid population of two genotypes, A and B. Each generation, every individual i produces a number of offspring $1 + X_i$, where X_i is a binomially-distributed random variable. This variable has mean μ_{A} and variance σ^2_{A} if i is of type A, or μ_{B} and σ^2_{B} if i is of type B. Adding 1 to X_i simply guarantees that there are at least N total offspring. These offspring are then culled by sampling without replacement until there is a new generation of exactly N organisms. We simulate this process to obtain the ψ function (Fig. 2.1D). Fitness ratios r are defined using the fitness definition $f_i = \mu_i - \sigma_i^2/N$. By repeating the simulation for several population sizes, we observe that ν is proportional to N (for each N, ν is obtained by a linear fit, one of which is shown in Fig. 2.1D).

2.3 Discussion

2.3.1 Universality

The notion of universality has been key to the success of population genetics. The remarkable fact that many population models with varying degrees of complexity share the same diffusion limit when selection is weak has proven to be a strong justification of their use as effective phenomenological theories [21, 115]. However, in light of the growing body of evidence that strong or at least intermediate selection may be important in some systems, it is desirable to pursue models that make no *a priori* assumptions about the strength of selection, and in particular, to find universal properties of such models. Our study shows that strong-selection effects are negligible in the steady state of the substitution process, so that the universality of the diffusion limit gives rise to a universal scaling law (Eq. 2.10) which determines the steady-state distribution (Eq. 2.13). Furthermore, the scaling law is proven to hold exactly for *any* reversible process (such as the Moran model) and holds approximately within a sizable range of selection strengths even for irreversible models. In some cases, such as the simple Wright-Fisher model, this range is so large that deviations from it are not practically important. This finding significantly generalizes previous work of Refs. 71, 121, 127, among others.

2.3.2 Theoretical significance of time reversibility

The existence of reversibility in the weak-selection limit is not surprising in light of diffusion theory. Indeed, diffusion models are essentially always reversible [20, 135, 136], and diffusion is known to adequately capture weak-selection behavior [137]. The fact that reversibility is broken by some models and not others when selection is strong is also clear. The Moran process, for instance, is well-known to be exactly reversible in all regimes, as are all models with tridiagonal transition matrices [20]. The Wright-Fisher model is not exactly reversible, and indeed we see that reversibility becomes significantly broken beyond a certain selection strength. In general, we find that the scaling behavior of the ψ function (Eq. 2.6) indicates the extent to which a model is time reversible.

But besides being a technical convenience, what is the deeper significance of reversibility? In modern studies of population genetics and evolution, reversibility plays a crucial role in linking the prospective and retrospective paradigms [138]. Traditional population models are prospective: the interest is in calculating future properties given the current ones. However, more recent approaches, especially due to the emergence of large-scale molecular data, have led to the wide use of the retrospective paradigm, which looks backward in time from the present. This is the essence of coalescent theory and phylogenetics [122, 139]. Time reversibility links the prospective and retrospective paradigms and thus has been exploited, for instance, in studies of age properties [20, 135, 140] and in phylogenetic methods [122].

An additional consequence of reversibility is the nonexistence of net probability currents in steady state, as guaranteed by Eq. 2.4. That is, reversible Markov models will have no net probability currents through any cycle of states, since such a current would distinguish between forward and backward directions in time. What does this mean for evolutionary models? Consider, for instance, a monomorphic substitution model with three genotypes, A, B, and C, in order of decreasing fitness. If the substitution process is irreversible, there would be a net current around the loop $\mathsf{C} \to \mathsf{B} \to \mathsf{A} \to \mathsf{C}$. The net currents $\mathsf{C} \to \mathsf{B}$ and $\mathsf{B} \to \mathsf{A}$ flow from less fit to more fit alleles, but to complete the cycle, there is also a current $\mathsf{A} \to \mathsf{C}$ from a more fit allele to a less fit allele. This current must exist in *any* irreversible substitution model with selection, a strange consequence of evolutionary irreversibility.

2.3.3 Applications

Models of monomorphic populations evolving through successive substitutions on a fitness landscape have important applications to molecular data, since loci in many asexual populations are believed to be well-approximated as monomorphic [116–119]. In particular, population genetics-based approaches allow for inference of biologically meaningful parameters, such as selection coefficients, as opposed to merely inferring overall substitution rates [120]. A precise form of the steady-state distribution is important in these applications; for example, it can be used to weigh ancestral nodes in phylogenetic inference calculations.

Several recent studies of codon usage bias have employed population genetics-based models of substitution with selection (e.g., [120, 127, 141–145]). Results for the steady-state distribution using the standard Wright-Fisher diffusion approximation (Eq. 2.29) for individual codons have been reported that are consistent with Eq. 2.13 in the limit of weak selection. However, there is growing experimental evidence that big-benefit single mutations may occur more often than previously thought. Studies on bacteriophages adjusting to new environmental conditions reported fitness ratios of nearly 4 [103–106], clearly beyond the diffusion regime. Thus, it is necessary to understand the role of these mutations in steady state and whether the steady-state distribution predicted from weak-selection must be modified in such systems. Our theoretical framework has enabled us to show that mutations with large fitness ratios are negligible in steady state.

Throughout this work we have assumed reversibility of the underlying mutation process. Reversible models are much more suitable to analytic and computational treatment, and thus reversibility is a key feature of many widely-used nucleotide and amino acid mutation models (e.g., [122, 146–150]). Moreover, Rodríguez et al. [151] have shown that it is not even possible to make self-consistent estimates of substitution rates from pairwise sequence alignments without assuming reversibility, although some work has been done to treat this type of molecular data with irreversible models [152]. Nevertheless, mutation rates are determined by complex biochemical factors (such as replication and error-correcting machinery), so there is no obvious reason to believe that reversibility must hold.

Our approach can be used to describe arbitrary fitness landscapes for the locus under consideration, including those with a fitness function that depends on the state of the entire DNA or protein sequence at the locus. Standard models of sequence evolution typically assume that all nucleotides or amino acids evolve independently of each other [122]. This approximation excludes correlations among sites within a locus and the corresponding epistatic effects, whose importance is being increasingly emphasized [31, 32, 38, 55].

One application of particular interest is the ability to infer an arbitrary fitness landscape from sequence data under the assumption of steady state. Indeed, Eq. 2.13 can be inverted to obtain the fitness function in terms of the neutral distribution and the steady-state distribution under selection [71, 72]:

$$\log\left(\frac{\pi(\sigma)}{\pi_0(\sigma)}\right) = \nu \log \mathcal{F}(\sigma) - \log Z. \tag{2.34}$$

Here the left-hand side depends only on genotype distributions that can, in principle, be obtained from sequence data. Since the scaling effective population size ν and normalization Z are unknown in real systems, Eq. 2.34 gives logarithmic fitness up to an overall scaling and shift.

The application of Eq. 2.34 requires an ensemble of loci that have reached evolutionary steady state. To assess this assumption, we estimate the time required to reach steady state in our substitution model. As discussed in Sec. 1.4.1, the monomorphic limit requires

$u \ll (LN \log N)^{-1}$ for neutral evolution [81, 82]. Assuming that deleterious substitutions do not affect equilibration towards steady state (due to exponential suppression of their substitution rates), equilibration times will be dominated by neutral evolution. Equation 1.5 then implies that the neutral substitution rate is equal to the mutation rate.

For sequences consisting of L nucleotides, we can model the locus genotype space as the vertices of a hypercube in $2L$ dimensions, since two bits encode a single nucleotide. A random walk on a hypercube of dimension d with standard connectivity reaches steady state on the order of $d \log d$ steps [153]. However, since the nucleotide sequence space hypercube is more connected, we may take $2L \log(2L)$ as an upper bound on the required number of steps. Combining this with the minimum average time to make a single neutral substitution step, $LN \log N$, we estimate that evolutionary steady state will be reached in the order of

$$(LN \log N) \times (2L \log(2L)) \text{ generations.} \tag{2.35}$$

For small genomic loci ($L = \mathcal{O}(10)$ nucleotides) in microbial organisms with generation times of approximately 10^{-4} years, an effective population size $N \sim 10^6$ yields an estimated time to reach steady state of about a million years, a reasonable value on evolutionary timescales. Moreover, the presence of selection, the additional connectivity of genotype space compared to a standard hypercube, and a smaller effective population size N will further shorten this timescale.

Moreover, the genotype space may be projected onto a lower-dimensional subspace. Previous work has described models of TF binding site evolution in *S. cerevisiae* in which the distribution of binding sites has been projected onto free energies of TF-DNA binding [69–72]. The steady state is expected to be reached more quickly in the one-dimensional energy space than in the high-dimensional genotype space [72]. Mustonen et al. [72] also find that energy distributions of binding sites for the same TF in different yeast species are remarkably similar despite significantly different divergence times, suggesting that these distributions have indeed reached evolutionary steady state.

This previous work, however, has relied purely on the diffusion approximation of the Wright-Fisher model. Such an approximation is not obviously valid in this application, since strong-selection effects are expected from binding site biophysics: single base pair mutations may be sufficient to completely inhibit TF binding [154, 155], potentially causing misregulation of an essential gene. We have demonstrated in this work that strong selection does not affect the steady state. The universality of the steady-state distribution then justifies application of Eq. 2.34 to genomic data such as collections of TF binding sites. In Chapter 3, we will apply these results to evolution of regulatory sites in yeast, exploring the biophysical origins of the underlying fitness landscapes.

Chapter 3

Fitness Landscapes of Transcription Factor Binding Sites

Here we apply the results of the preceding chapter to a specific biological system. This chapter reproduces Ref. 9. As described in Chapter 1, most traditional studies of molecular evolution rely on simplified models of fitness landscapes [24, 34, 41, 42] or reconstruct the landscapes empirically based on limited experimental data [24, 31, 32, 36, 37]. However, fitness landscapes are fundamentally shaped by complex interactions involving DNA, RNA, proteins, and other molecular species present in the cell. Thus we should be able to cast these landscapes in terms of biophysical properties such as binding affinities, molecular stabilites, and degradation rates. The increasing availability of quantitative high-throughput data on molecular interactions in the cell has led to growing efforts aimed at developing models of evolution that explicitly incorporate the underlying biophysics [10, 11, 27, 55, 56, 58–60, 67–72, 156]. These models combine evolutionary theory with physical models of molecular systems, for example focusing on how protein folding stability or specificity of intermolecular interactions shapes the ensemble of accessible evolutionary pathways and steady-state distributions of biophysical phenotypes.

Evolution of gene regulation is particularly well-suited to this type of analysis. Gene activation and repression are mediated by binding of **transcription factor** proteins (TFs) to their cognate genomic sites. These **binding sites** are short nucleotide sequences, typically 5–25 bp in length, in gene promoters that interact specifically with DNA-binding domains on the TFs [157]. In eukaryotes, a given TF can have numerous binding sites in the genome, and many genes are regulated by several TFs [157, 158]. Understanding TF-mediated regulation is key to understanding complex regulatory networks within eukaryotic cells — one of the main challenges facing molecular biology. Moreover, the availability of high-throughput datasets on the genomic locations of TF binding sites [159–162] and on TF-DNA energetics [163–166] make it possible to develop biophysical models of evolution of gene regulation.

In this chapter we consider evolution of TF binding sites in the yeast *Saccharomyces cerevisiae* [9]. We study how energetics of protein-DNA interactions affect the structure of the binding site fitness landscape. In a significant extension of previous work which analyzed a single yeast TF [72], we consider a collection of 25 *S. cerevisiae* TFs for which models of TF binding energetics were built using high-throughput *in vitro* measurements of TF-DNA interactions [166]. We focus on 12 TFs for which sufficient data on genomic sites [162] are also available. We use the model of monomorphic populations undergoing consecutive substitutions defined in Chapters 1 and 2 to infer fitness landscapes, as a function of TF binding energy, from observed distributions of TF binding sites in the yeast genome [72]. In contrast to the previous work [72], we rationalize these fitness landscapes in terms of a simple parametric model based on thermodynamics of TF-DNA binding, obtaining explicit values of effective evolutionary parameters. Our analysis sheds light on the genome-wide importance of TF-DNA interactions in regulatory site evolution.

Moreover, we investigate the hypothesis that universal biophysical constraints, rather than site-specific selective pressures, dominate evolution of regulatory sites. We test the relationship between TF binding energies and various biological properties, such as the essentiality of the corresponding gene [167]. We find no clear relationship between physical and biological properties of TF sites, which indicates that evolution of site energetics is largely insensitive to site-specific biological functions and is therefore driven by global biophysical constraints.

3.1 Biophysical model of TF binding site evolution

3.1.1 Energetics of TF-DNA binding

The probability of a binding site to be bound by a TF is given by the Fermi-Dirac function of the free energy E of TF-DNA interaction [168]:

$$p_{\mathrm{bound}}(E) = \frac{1}{1 + e^{\beta_{\mathrm{phys}}(E - \mu_{\mathrm{phys}})}}, \tag{3.1}$$

where β_{phys} is the physical inverse temperature (≈ 1.69 (kcal/mol)$^{-1}$ at room temperature), and μ_{phys} is the physical chemical potential, a function of the TF concentration. The binding energy $E = E(\sigma)$ of a site is a function of its nucleotide sequence, $\sigma = (\sigma_1, \ldots, \sigma_L)$, where L is the length of the site and $\sigma_i \in \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$. Note that $p_{\mathrm{bound}}(E) \approx e^{-\beta_{\mathrm{phys}}(E - \mu_{\mathrm{phys}})}$ if $E \gg \mu_{\mathrm{phys}}$, resulting in a Boltzmann-like exponential distribution. In the mean-field approximation, each nucleotide makes an additive contribution to the total energy of the site [163]. These contributions are parameterized by an energy matrix, whose entries $\epsilon_i^{\sigma_i}$ give the contribution to the total energy from the nucleotide σ_i at position i:

$$E(\sigma) = \sum_{i=1}^{L} \epsilon_i^{\sigma_i}. \tag{3.2}$$

Energy matrices can be readily generalized to more complex models of sequence-dependent energetics, such as those with contributions from dinucleotides, although here for simplicity we focus on the additive model.

3.1.2 Biophysical model of binding site evolution

We will suppose a binding site evolves according to the monomorphic substitution model of Chapters 1 and 2, defined by Eqs. 1.5 and 1.6 with steady state given by Eq. 2.13. Since we are primarily interested in the biophysical aspects of binding site evolution, it is more convenient to consider evolution in the phenotype space of binding energies by projecting

Eq. 2.13 via the genotype-phenotype map of Eq. 3.2:

$$\pi(E) = \frac{1}{Z}\pi_0(E)\mathcal{F}(E)^{\nu}. \tag{3.3}$$

Here the binding site fitness $\mathcal{F}(E)$ depends only on the binding energy E. As a general ansatz, we will assume that the fitness depends on the binding energy through the physical binding probability p_{bound}: $\mathcal{F}(E) = \mathcal{F}(p_{\text{bound}}(E))$. Further, we assume that an organism with an always-bound site ($p_{\text{bound}} = 1$, $E \to -\infty$) has fitness 1, while an organism with a site that never binds ($p_{\text{bound}} = 0$, $E \to +\infty$) has fitness $f_0 < 1$. Since real sites are somewhere in between these extremes, a simple hypothesis for the fitness function is an average of these two fitness values weighted by the thermodynamic probabilities of the site being bound or unbound:

$$\mathcal{F}(E) = p_{\text{bound}}(E) + f_0(1 - p_{\text{bound}}(E)) = \frac{1 + f_0 e^{\beta_{\text{phys}}(E-\mu_{\text{phys}})}}{1 + e^{\beta_{\text{phys}}(E-\mu_{\text{phys}})}}. \tag{3.4}$$

Equation 3.4 assumes that the fitness function depends linearly on the TF binding probability p_{bound}, which equals the site's average occupancy. However, this linear dependence may be too restrictive. For example, it does not account for the scenario in which a cell only requires p_{bound} to be above some minimum threshold p_{min}, such that the fitness equals 1 when $p_{\text{bound}} > p_{\text{min}}$ and zero otherwise. To include a wider range of fitness functions, we extend our model in Eq. 3.4 by treating β and μ as effective fitting parameters (β_{eff} and μ_{eff}) that may deviate from their physical counterparts:

$$\mathcal{F}(E) = \frac{1 + f_0 e^{\beta_{\text{eff}}(E-\mu_{\text{eff}})}}{1 + e^{\beta_{\text{eff}}(E-\mu_{\text{eff}})}}. \tag{3.5}$$

When $\beta_{\text{eff}} = \beta_{\text{phys}}$ and $\mu_{\text{eff}} = \mu_{\text{phys}}$, Eq. 3.5 is equivalent to Eq. 3.4 and fitness is linearly proportional to p_{bound}, but deviations between these effective and physical parameters introduce nonlinear dependence of fitness on p_{bound}. For example, the case in which p_{bound}

must only exceed a minimum threshold $p_{\rm min}$ is equivalent to Eq. 3.5 with $f_0 = 0$, $\beta_{\rm eff} \to \infty$, and $\mu_{\rm eff} = \mu_{\rm phys} + \beta_{\rm phys}^{-1} \log((1 - p_{\rm min})/p_{\rm min})$. For the remainder of the paper, we will focus on the effective fitness function of Eq. 3.5, which we will use to infer the effective parameters from data. For simplicity we will drop the explicit "eff" labels on β and μ.

An important feature of Eq. 2.13 is that we may invert it to obtain the fitness function in terms of the observed steady-state distributions $\pi(\sigma)$ and $\pi_0(\sigma)$, or $\pi(E)$ and $\pi_0(E)$ in energy space [71]:

$$\log\left(\frac{\pi(\sigma)}{\pi_0(\sigma)}\right) = \nu \log \mathcal{F}(\sigma) - \log Z \quad \Longrightarrow \quad \log\left(\frac{\pi(E)}{\pi_0(E)}\right) = \nu \log \mathcal{F}(E) - \log Z. \tag{3.6}$$

Thus, given a distribution of evolved binding site sequences π and a neutral distribution π_0, we can use Eq. 3.6 to infer the logarithm of the fitness landscape up to an overall scale and shift. This can be done without any *a priori* knowledge of the shape of the fitness function. Moreover, given a specific functional form of $\mathcal{F}(E)$, such as the effective Fermi-Dirac fitness in Eq. 3.5, we can perform a maximum likelihood fit of the observed sequence distribution to infer values of parameters β, μ, ν, and f_0. The resulting fitted function can be evaluated by comparison to the general inference in Eq. 3.6.

When $1 - f_0 \ll 1$, $\mathcal{F}(E)^\nu$ contains an approximate degeneracy in terms of $\nu(1 - f_0) \equiv \gamma$, i.e., all fitness functions with constant γ are approximately equivalent. Indeed, the steady-state distribution in Eq. 3.3 depends on the quantity $\mathcal{F}(E)^\nu$, which can be written as:

$$\mathcal{F}(E)^\nu = \left(1 - \frac{\gamma}{\nu}(1 + e^{-\beta(E-\mu)})^{-1}\right)^\nu \approx e^{-\gamma(1+e^{-\beta(E-\mu)})^{-1}} \tag{3.7}$$

if $\gamma(1 + e^{-\beta(E-\mu)})^{-1} \ll \nu$ or, since $0 \leq (1 + e^{-\beta(E-\mu)})^{-1} \leq 1$, if $1 - f_0 \ll 1$. Therefore in this limit, the steady-state distribution $\pi(\sigma)$ depends only on the parameter γ and not on f_0 and ν separately.

This degeneracy in the steady-state distribution is not surprising in light of the underlying population genetics, which also provides an interpretation of γ. The quantity $1 - f_0$ is the selection coefficient s between the two phenotypes of the system, e.g., the bound and unbound states of the TF binding site. As discussed above, the quantity ν is an effective population size, which sets the strength $1/\nu$ of genetic drift. When $s \ll 1$ and $\nu \gg 1$, steady-state properties of the population (e.g., allele frequency distribution, fixation probability) are described by the diffusion limit and mathematically depend only on Ns, or in our model, on $\nu(1 - f_0) = \gamma$ [20, 21], which quantifies the strength of selection relative to the strength of drift. When $\gamma > 1$, selection is strong relative to drift, while $\gamma < 1$ indicates that selection is relatively weak. Note that only the absolute magnitude of the selection coefficient $s = 1 - f_0$ is required to be small for this degeneracy to hold; the selection strength relative to drift, quantified by γ, may still be large.

Two regions of parameter space also exhibit a degeneracy between μ and γ. If $\mu \gg E$ for all site energies E, all of the observed sites are predicted to be highly occupied and $p_{\text{bound}}(E) \approx 1 - e^{\beta(E-\mu)}$. We may thus approximate

$$\begin{aligned}\mathcal{F}(E)^\nu &\approx \left(1 - (1 - f_0)e^{\beta(E-\mu)}\right)^\nu \\ &\approx 1 - \nu(1 - f_0)e^{\beta(E-\mu)} = 1 - (\gamma e^{-\beta\mu})e^{\beta E},\end{aligned} \tag{3.8}$$

and thus all fitness functions with constant $A_1 = \gamma e^{-\beta\mu}$ are approximately equivalent. One can thus make μ arbitrarily large (while holding A_1 fixed by varying γ) without breaking the degeneracy. If μ is decreased the degeneracy will eventually break as $\mu \gg E$ is violated. A similar degeneracy appears when $\mu \ll E$, as then $p_{\text{bound}}(E) \approx e^{-\beta(E-\mu)}$; if additionally $f_0 \approx 1$, then

$$\begin{aligned}\mathcal{F}(E)^\nu &\approx \Big(f_0 + (1-f_0)(e^{-\beta(E-\mu)})\Big)^\nu \\ &\approx 1 - \nu(1-f_0)e^{-\beta(E-\mu)} = 1 - (\gamma e^{\beta\mu})e^{-\beta E}.\end{aligned} \tag{3.9}$$

(We can remove an overall factor of f_0^ν because the distribution $\pi(E)$ in Eq. 3.3 is invariant under an overall rescaling of fitness.) Therefore all fitness functions with $A_2 = \gamma e^{\beta\mu}$ are approximately equivalent in this case. Here, μ can be made arbitrarily negative without breaking the degeneracy.

Thus, parameter fits fall into three cases for different TFs: If $\mu \ll E$, TF-DNA binding energies fit to the right (exponential) end of the Fermi-Dirac function, and we cannot infer a unique μ. Similarly, if $\mu \gg E$, TF-DNA binding energies fit to the left (high occupancy) side of the Fermi-Dirac function, and we again cannot infer μ precisely. However, if $\mu \approx E$, neither degeneracy holds and a unique μ can be inferred. Despite the fact that μ cannot always be predicted, we can unambiguously classify each fit into one of these three cases.

3.1.3 Selection strength and its dependence on biophysical parameters

We now consider how changes to biophysical parameters of the model affect the strength of selection on binding sites. The selection coefficient for a mutation with small change in energy ΔE is

$$s(E) = \frac{\mathcal{F}(E+\Delta E)}{\mathcal{F}(E)} - 1 \approx \frac{d\log\mathcal{F}}{dE}\Delta E. \tag{3.10}$$

Therefore we can characterize local variations in the strength of selection by considering $\tilde{s}(E) = |d\log\mathcal{F}/dE|$, the per-unit-energy local selection coefficient. For the Fermi-Dirac landscape, we obtain

$$\tilde{s}(E) = \left|\frac{d}{dE}\log\mathcal{F}(E)\right| = \frac{\beta(1-f_0)z}{(1+z)(1+f_0 z)}, \quad \text{where} \quad z = e^{\beta(E-\mu)}. \tag{3.11}$$

We use the absolute value here since the sign of the selection coefficient is always unambiguous, as the Fermi-Dirac function decreases monotonically with energy.

We can also ask how variations in β affect the local strength of selection. Variation of $\tilde{s}(E)$ with β depends qualitatively on both $E - \mu$ and whether f_0 is zero or nonzero. In Fig. 3.1 we show $\log \mathcal{F}(E)$, $\tilde{s}(E)$, and the derivative

$$\frac{\partial \tilde{s}}{\partial \beta} = \frac{z(1-f_0)}{(1+z)^2(1+f_0 z)^2}[(1-f_0 z^2)\log z + (1+z)(1+f_0 z)]. \tag{3.12}$$

For $f_0 = 0$ (Fig. 3.1A–C), increasing β increases selection strength for $E - \mu \geq 0$. Here the fitness function drops to zero exponentially, and increasing β steepens the exponential drop. However, for $E - \mu < 0$, the effect of changing β depends on the value of β relative to $E - \mu$. For large β, increasing β actually decreases selection strength; this is because β sets the rate at which the Fermi-Dirac function converges to unity, and hence increasing β flattens the landscape in that region. However, for sufficiently small β, the threshold region is large enough that increasing β still increases selection. The boundary between positive and negative values of $\partial\tilde{s}/\partial\beta$ are the solutions of the equation $\partial\tilde{s}/\partial\beta = 0$: $\beta(E-\mu) = \log W(e^{-1}) \approx -1.278$, where W is the Lambert W-function (Fig. 3.1C).

This situation changes qualitatively in the regime $E - \mu > 0$ when $f_0 \neq 0$ (Fig. 3.1D-F). In this case, for sufficiently large β, increasing β weakens selection. This is different in the case of nonzero f_0 because on the high-energy tail, the fitness is converging to a nonzero number f_0, and thus selection becomes asymptotically neutral. Hence, when $f_0 \neq 0$, increasing β only strengthens selection very close to $E - \mu = 0$. Using Eq. 3.12, the boundaries in Fig. 3.1F are given by the solutions of $(f_0 z^2 - 1)\log z = (1+z)(1+f_0 z)$. This equation can be solved numerically to obtain two solutions, $z_1^* < 1$ and $z_2^* > 1$. The boundaries in Fig. 3.1F are thus given by the curves $\beta(E-\mu) = \log z_1^*$ for $E - \mu < 0$ and $\beta(E-\mu) = \log z_2^*$ for $E - \mu > 0$.

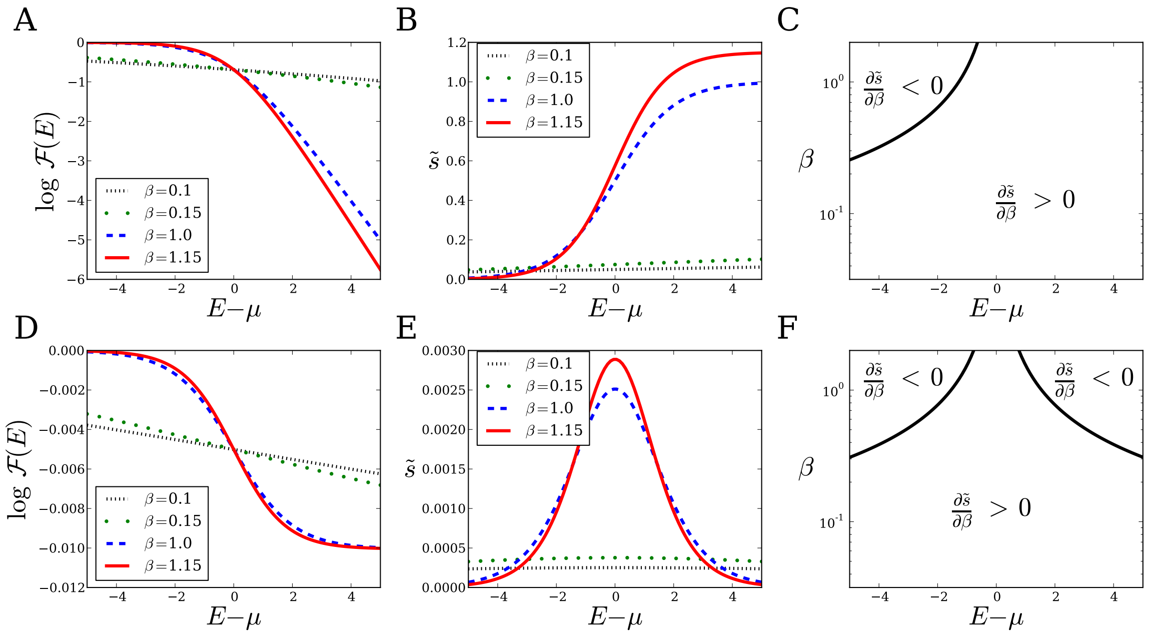


Figure 3.1: **Fitness and selection strength as functions of energy $E-\mu$ and inverse temperature β.** Energy is measured with respect to the chemical potential μ. Top row uses $f_0 = 0$; bottom row uses $f_0 = 0.99$. (A,D) Logarithm of Fermi-Dirac fitness versus energy for several values of β; note that the high-energy tail looks distinctly different when f_0 is nonzero. (B,E) Per-unit-energy selection strength $\tilde{s}$ versus energy for several values of β; note that the relative ordering of selection strength curves depends on the value of $E-\mu$. (C,F) Sign of derivative of selection strength with respect to β, as a function of $E-\mu$ and β. Black boundary in (C) is the curve $\beta(E-\mu) = \log W(e^{-1}) \approx -1.278$, where W is the Lambert W-function; the boundaries in (F) are the curves $\beta(E-\mu) = \log z_1^* \approx -1.541$ and $\beta(E-\mu) = \log z_2^* \approx 1.545$, where z_1^*, z_2^* are the solutions to $\partial\tilde{s}/\partial\beta = 0$ (Eq. 3.12) with $f_0 = 0.99$.

3.2 Assessment of model assumptions

Two main assumptions inherent in our evolutionary model are monomorphism and steady state. Here, we assess how violating these assumptions affects inference of evolutionary parameters β, μ, ν, and f_0. To test this, we generate simulated data sets of binding site sequences evolving under a haploid asexual Wright-Fisher model with the Fermi-Dirac fitness function (Eq. 3.5; see Appendices B.1 and B.2 for details).

3.2.1 Deviations from the monomorphic limit

To test the effects of polymorphism on the accuracy of our predictions, we perform a set of simulations for a range of mutation rates u. Each simulation in the set follows the Wright-Fisher process to the steady state. We construct the observed distribution π_{obs} by randomly choosing a single sequence from the final population of each simulation, which may not be monomorphic for larger u (Fig. 3.2A). From π_{obs}, we carry out maximum-likelihood inference of the fitness landscape as a function of energy using Eq. 3.3 (Fig. 3.2B), as described in Appendix B.1.

Additionally, for each u we record the average number of unique sequences present in the population in steady state, and compute the total variation distance (TVD; Eq. B.2 in Appendix B.2) between $\pi_{\text{obs}}(E)$ and the monomorphic prediction $\pi(E)$ using Eq. 3.3 (Fig. 3.2C). The TVD ranges from zero for identical distributions to unity for completely non-overlapping distributions. As expected, at low mutation rates the steady-state distribution and the fitness function match monomorphic predictions well. At higher mutation rates, the TVD starts to increase and Eq. 2.13 overestimates the fitness of low-affinity sites. The population becomes polymorphic in this limit. With very high mutation rates, π_{obs} approaches the neutral distribution π_0 since the population is largely composed of newly generated mutants which have not experienced selection. A condition for monomorphism in a neutrally evolving population is $u \ll (LN \log N)^{-1}$ (Sec. 1.4.1). Using $N = 1000$ and $L = 10$ as in our simulations yields $u \ll 1.4 \times 10^{-5}$ in the monomorphic limit, consistent

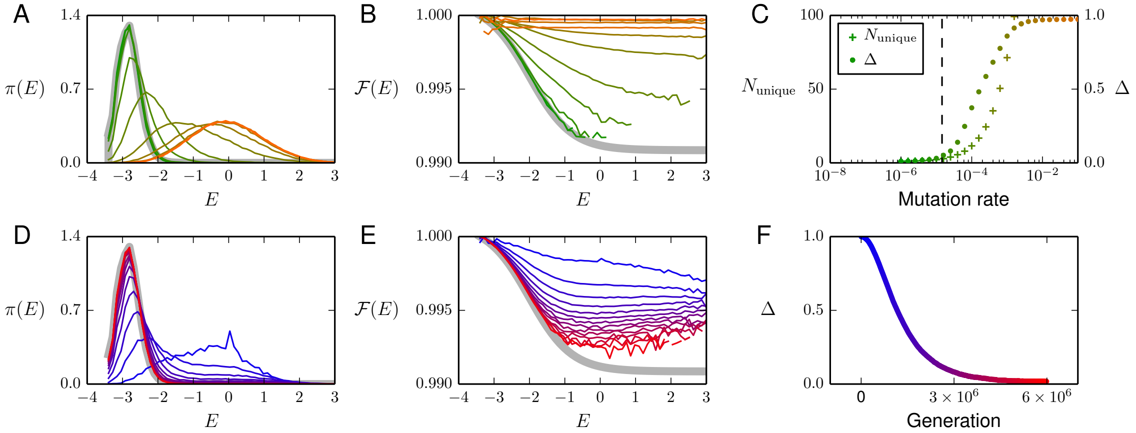


Figure 3.2: **The monomorphic limit and steady state of a Wright-Fisher model of population genetics.** In (A)–(C) we show results from simulations at various mutation rates, using a fitness function with $f_0 = 0.99$, $\beta = 1.69$ (kcal/mol)$^{-1}$, and $\mu = -2$ kcal/mol. Each mutation rate data point is an average over 10^5 independent runs, as described in Appendix B.1. Colors from green to orange correspond to increasing mutation rates. (A) Observed steady-state distributions $\pi_{\mathrm{obs}}(E)$ for various mutation rates. The steady state $\pi(E)$ predicted using Eq. 2.13 is shown in gray. (B) Fitness functions $\mathcal{F}(E)$ predicted using observed distributions $\pi_{\mathrm{obs}}(E)$ in Eq. 3.6. The exact fitness function is shown in gray. Inferred fitness functions are matched to the exact one by using the known population size N, and setting the maximum fitness to 1.0 for each curve. (C) For each mutation rate, the total variation distance (TVD) Δ between $\pi_{\mathrm{obs}}(E)$ and $\pi(E)$, and the average number of unique sequences in the population N_{unique} (the degree of polymorphism) are shown. The predicted bound $(NL \log N)^{-1}$ on mutation rate required for monomorphism is shown as a dashed line. In (D)–(F) we show simulations in the monomorphic regime which have not reached equilibrium, with the same parameters as in (A)–(C) and $u = 10^{-6}$. Colors from blue to red correspond to the increasing number of generations. In (F), TVD Δ is calculated in energy space as described in Appendix B.2.

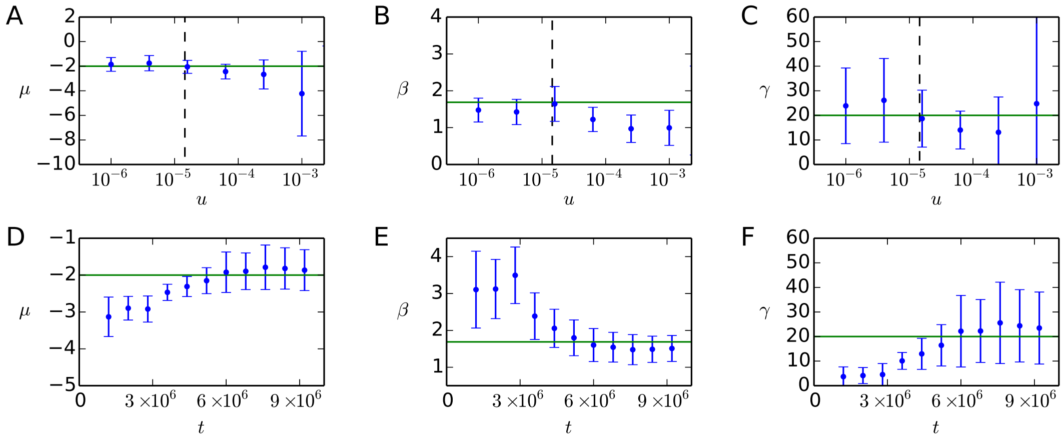


Figure 3.3: **Fitted parameters of the Fermi-Dirac function from Wright-Fisher simulations.** In (A)–(C) the fitted values of μ, β and $\gamma = \nu(1 - f_0)$ are shown as functions of mutation rate u. For each mutation rate, we generate 200 random samples of 500 sequences from the 10^5 sequences generated in simulations used in Fig. 3.2A–C. We fit the parameters of the fitness function on each sample separately by maximum likelihood (see Appendix B.1). Shown are the averages (points) and standard deviations (error bars) over 200 samples at each mutation rate. The exact values used in the simulation are represented by horizontal green lines. The predicted bound $(LN \log N)^{-1}$ on mutation rates required for monomorphism is shown as a vertical dashed line. In (D)–(F) the fitted values of μ, β, and γ are shown as functions of the number of generations t, for the non-steady state simulations used in Fig. 3.2D–F. The sampling procedure, the maximum likelihood fit, and the representation of parameter predictions are the same as in (A)–(C).

with the results in Fig. 3.2C.

We also infer parameters β, μ and γ with a maximum likelihood fit. As expected, all parameters converge to the exact values in the monomorphic limit (Fig. 3.3A–C). When the population is not truly monomorphic, μ and β tend to be underestimated on average, with larger variation in inferred values (larger error bars in Fig. 3.3A,B). For γ, polymorphism has no clear bias on the average inferred value, although it also appears to increase the variation.

3.2.2 Deviations from evolutionary steady state

We perform another set of simulations to test the accuracy of our predictions in a population that has not yet reached steady state. We use the same fitness landscape and population size, but fix u to 10^{-6}, within the monomorphic limit. At each point in time (measured as the number of generations), we construct π_{obs} as described in Appendix B.2 (Fig. 3.2D), and infer the fitness function (Fig. 3.2E). We also compute the TVD between the observed distribution π_{obs} and the steady-state prediction (Fig. 3.2F). With time, π_{obs} converges to the steady state (Eq. 2.13) and the TVD decays to zero, enabling accurate reconstruction of the fitness function in the region $E - \mu \approx 0$ (although it still diverges from the exact function in the high-energy tail, where few sequences are available at steady state). The equilibration time is expected to be proportional to u^{-1}, or 10^6 generations; indeed, Fig. 3.2F places the equilibration timescale at about 4×10^6 generations. As the population equilibrates, accurate inference of the fitness function parameters becomes possible (Fig. 3.3D–F). We see that parameters inferred from a population out of steady state tend to underestimate μ and γ, and overestimate β.

3.3 Transcription factor binding sites in yeast

We now turn to considering the evolution of TF binding sites in *S. cerevisiae*. How well does *S. cerevisiae* satisfy the assumptions of our evolutionary model? *S. cerevisiae* is not a purely haploid organism but rather goes through both haploid and diploid stages. In *S. paradoxus*, most of the reproduction is haploid and asexual with 1000 generations spent in the haploid stage for each generation in the diploid stage, and heterozygosity is low [169]. Based on the analysis of yeast genomes, wild yeast populations show extremely limited outcrossing and recombination and are geographically distinct [170]. Thus, *S. cerevisiae* may be regarded as haploid to a reasonable approximation, with recombination during the diploid stages unlinking TF binding sites. This is consistent with our model, which assumes a haploid population and independent evolution of binding sites.

We next consider whether natural populations of *S. cerevisiae* are within the mutation rate limits required for monomorphism. The mutation rate for *S. cerevisiae* has been estimated to be 0.22×10^{-9} mutations per bp per cell division [169]. Assuming binding site loci of length $L = 10$, this sets a bound on the effective population size N of 2.7×10^7, below which the population will be monomorphic. This is roughly equal to the estimated effective population size of *S. cerevisiae* of $\approx 10^7$ individuals [169], based on the analysis of neutral regions in the yeast genome. Thus it is plausible that *S. cerevisiae* population sizes are below or near the limit for monomorphism, justifying the use of our model result in Eq. 2.13. Furthermore, in *S. cerevisiae* and *S. paradoxus* the proportion of polymorphic sites in a population has been found to be about 0.001 [169, 171, 172], generally with no more than two alleles segregating at any one site [169]. According to this estimate, we expect about 1% of binding sites of length 10 bp to be polymorphic, corresponding to an average polymorphism of 1.01 in Fig. 3.2C.

For *S. cerevisiae*, the equilibration time estimate is $u^{-1} \approx 5 \times 10^9$ generations, or about 2×10^6 years for an estimated 8 generations per day [173]. This is several times less than the 5–10 million years of divergence time for the most recent speciation event with *S. paradoxus* [174]. Thus steady state may plausibly be reached for a fast-reproducing organism like *S. cerevisiae* over evolutionary times scales.

3.3.1 Site-specific selection

We obtain curated binding site locations in *S. cerevisiae* from Ref. 162, and energy matrices (EMs) from Ref. 166, as described in Appendix B.3. Besides the assumptions of monomorphism and steady state, we also require a set of binding sites evolving under universal selection constraints if we are to infer the fitness landscape using Eq. 3.3. A collection of sites binding to the same TF is an obvious candidate, since these sites all experience the same physical interactions with the TF. However, it is possible that selection is site-specific: rather than evolving on the same fitness landscape, different sites for the same TF may be under different selection pressures depending on which genes they regulate, their position on the chromosome, etc. For example, genes under strong selection might require very reliable regulation, so that their upstream binding sites are selected for tight binding to TFs. In less essential genes, the requirement of high-affinity binding might be relaxed. Before directly applying the evolutionary model, we investigate several of these site-specific scenarios to determine if any are supported by the data. We perform several direct tests of site-specific selection by searching for correlations between site TF-binding energies and other properties of the site or the gene it regulates.

We classify fitness effects of genes using knockout lethality, which is available in the Yeast Deletion Database [167, 175]. This database classifies genes as either essential or nonessential based on the effects of gene knockout, and provides growth rates for nonessential gene knockouts under a variety of experimental conditions. We divide binding sites of each TF in our data set into two groups: those regulating essential genes and those regulating nonessential genes.

In Fig. 3.4A we compare mean binding energies of sites regulating essential genes with those regulating nonessential genes for each TF. Using a null model as described in Appendix B.4, we find no significant difference (at $p = 0.05$ level) between the two groups of sites for any TF except RPN4, for which $p = 0.048$ and the difference in mean energies is 0.33 kcal/mol. The mean p-value of the null model over all TFs is 0.47. In Fig. 3.4B we compare the variance of the energy of the sites regulating essential and nonessential genes; sites regulating essential genes may be selected for more specific values of binding energy if precise regulation is required. We find no overall trend: for some TFs sites regulating essential genes have more energy variation than those regulating nonessential genes, but for other TFs the situation is reversed.

For the sites regulating nonessential genes, we also correlate the site binding energy with the growth rate of a strain in which the regulated gene was knocked out (Appendix Table B.1, column B). The Spearman rank correlation between each site's binding energy and the regulated gene's effect on growth rate produces a mean p-value of 0.56. We find no significant correlation for any TF at $p = 0.05$ level except MSN2, with $p = 0.046$.

It is also possible that regulation of highly-expressed genes may be more tightly controlled. Indeed, gene expression level is weakly, though significantly, correlated with gene

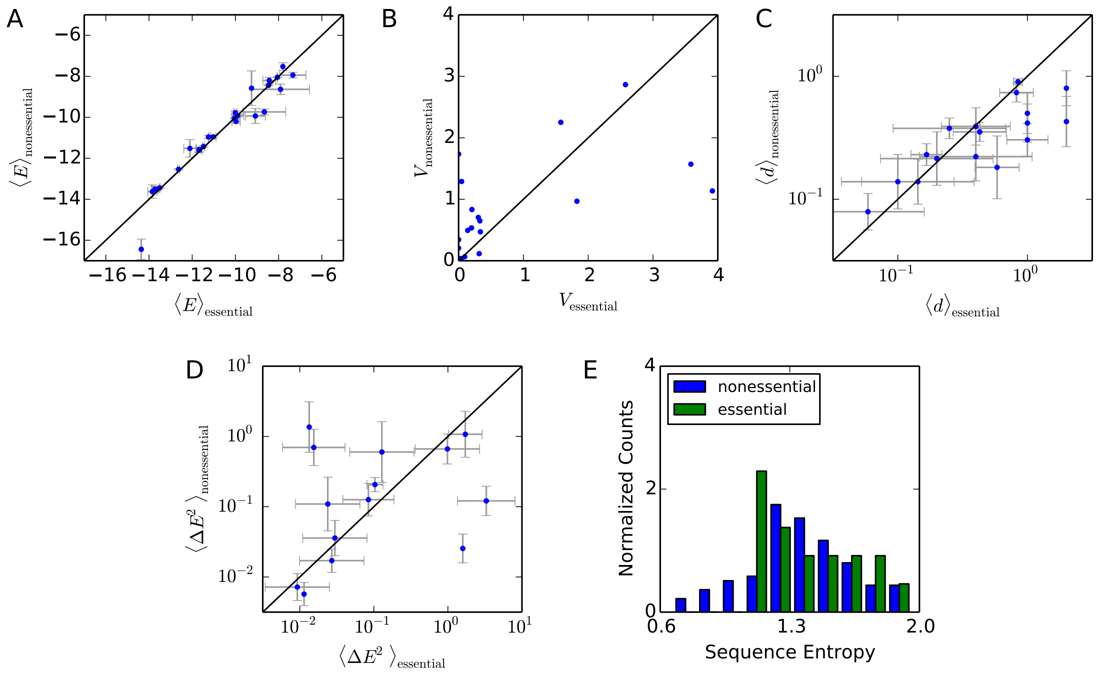


Figure 3.4: **Tests of site-specific selection.** We divide binding sites for each TF into two groups: those regulating essential genes and those regulating nonessential genes. (A) Comparison of mean binding energies of sites regulating essential ($\bar{E}_{\text{essential}}$) and nonessential genes ($\bar{E}_{\text{nonessential}}$) for each TF in the data set. (B) Comparison of variance in binding energies for sites regulating essential ($V_{\text{essential}}$) and nonessential ($V_{\text{nonessential}}$) genes. (C) Mean Hamming distance between corresponding sites in *S. cerevisiae* and *S. paradoxus* for sites regulating essential versus nonessential genes. (D) Mean squared difference in binding energy between corresponding sites in *S. cerevisiae* and *S. paradoxus* for sites regulating essential versus nonessential genes. In (A)–(D), 25 TFs were used; black diagonal lines have slope one. In (A),(C),(D), vertical and horizontal error bars show the standard error of the mean in each group. Points lacking error bars have only one sequence in that group. (E) Normalized histogram of TF binding site sequence entropies, divided into 16 essential and 109 nonessential TFs, for 125 TFs in Ref. 162.

essentiality [176]. We compare the binding energy of sites to the overall expression level of their regulated genes measured in mid-logphase yeast cells cultured in YPD [176] (Appendix Table B.1, column C), and again find no correlation using the Spearman rank correlation except for DAL80 ($p = 0.029$), with mean p-value of 0.53.

Another measure of the selection pressures on genes is their rate of evolution as measured by K_A/K_S, the ratio of nonsynonymous to synonymous mutations in a given gene between species. According to the neutral theory of evolution, genes which evolve slowly must be under higher selective pressure, and therefore the sites regulating them might likewise experience stronger selective pressures. As described in Appendix B.4, we measure the K_A/K_S ratio between *S. cerevisiae* and *S. paradoxus* protein coding sequences, and compare it to the binding energy of the sites regulating those genes (Appendix Table B.1, column D). We find very weak Spearman rank correlations for ATF2, RPN4, GAT1 and CAD1, all roughly with $p = 0.02$. We find no other significant correlation at the $p = 0.05$ level, with a mean p-value of 0.4.

Similarly, one might expect sites regulating essential genes to be more conserved. However, we find that the average Hamming distance between corresponding binding sites in *S. cerevisiae* and *S. paradoxus* [162] is no different for sites regulating essential genes than for those regulating nonessential genes, as shown in Fig. 3.4C. Using the null model described in Appendix B.4, most TFs are above $p = 0.05$ with the exceptions of YAP7 ($p = 0.04$) and PDR3 ($p = 0.009$), with an average p-value of 0.31. Similarly, there is no significant difference in the binding energies of these orthologous sites as determined from the EMs, as shown in Fig. 3.4D, except for PDR3 ($p = 0.01$), with mean p-value of 0.42.

We can also consider how the essentiality of the TFs themselves affects the sequences of their binding sites; for example, essential TFs may constrain their binding sites to a more conserved sequence motif. We divide 125 TFs from Ref. 162 which had 10 or more sequences and for which essentiality information was available into 16 essential and 109 nonessential TFs using the Yeast Deletion Database [167, 175], and calculate the sequence entropy of binding sites for each TF. The distribution of sequence entropies in Fig. 3.4E shows no significant difference between essential and nonessential TFs ($p = 0.9$ for the null model).

Finally, it is possible that sites experience different selection pressures depending on their distance to the transcription start site (TSS). Again, we find no significant correlations between binding energy and distance to the TSS: Spearman rank correlation yields mean p-value of 0.55 and all p-values above 0.05 (Appendix Table B.1, column E). Overall, we find no systematic evidence that site-specific properties of binding sites determine their binding energies. These findings are in broad agreement with a previous report [72], which suggested that site-specific selection can be ruled out because of the significant variation in binding affinity between orthologous sites of different species, which was found to be consistent with the variance predicted by a model including only drift and site-independent selection.

3.3.2 Inference of biophysical fitness landscapes

The above analysis indicates that the evolution of binding site energies does not depend significantly on site-specific effects, suggesting that more universal principles govern the observed distribution of sites binding a given TF. Thus, we can fit a single fitness function to a collection of TF-bound sites via Eqs. 2.13 and 3.6. Of the 25 TFs considered in the previous section, here we focus on 12 TFs with > 12 unique binding site sequences.

First we derive the neutral distribution $\pi_0(E)$ of site energies based on mono- and dinucleotide frequencies obtained from intergenic regions of the *S. cerevisiae* genome, as described in Appendix B.5. It has been suggested that L-mers not functioning as regulatory sites (e.g., located outside promoters) may be under evolutionary pressure not to bind TFs [177]; however, consistent with previous reports [72, 178], we find that sequences sampled from the intergenic regions of the genome are close to the neutral distribution expected from mono- and dinucleotide frequencies, except for the expected enrichment at low energies due to functional binding sites. This distribution is shown in Fig. 3.5A for REB1 and in Appendix Table B.2, column B for all other TFs.

Assuming the observed set of binding site energies for a TF adequately samples the distribution $\pi(E)$, we can use our estimate of the neutral distribution $\pi_0(E)$ in Eq. 3.6 to reconstruct the fitness landscapes as a function of TF binding energy up to an overall scale and shift (Fig. 3.6). Although the fitness functions may be noisy due to imperfect sampling of $\pi(E)$, they nevertheless provide important qualitative insights. In particular, in all cases fitness decreases monotonically as binding energy increases, indicating that stronger-binding sites are more fit. That is, we observe no fitness penalty for binding too strongly, at least within the range of energies spanned by $\pi(E)$.

Fitted Fermi-Dirac landscapes. For each TF we perform a maximum-likelihood fit of the binding site data to the distribution in Eq. 2.13 with the Fermi-Dirac landscape of Eq. 3.5 (Fig. 3.5 for the REB1 example, Table 3.1 and Appendix Table B.2 for all other TFs; see Appendix B.1 for details). The model of Eq. 3.5 has four fitting parameters: β, μ, ν, and f_0. However, as shown in Sec. 3.1.2, in the $1 - f_0 \ll 1$ limit the fitness function depends on $\gamma = \nu(1 - f_0)$ rather than f_0 and ν separately. Thus we also carry out constrained "non-lethal" Fermi-Dirac fits in which f_0 is fixed at 0.99. Note that due to the γ-μ degeneracy, in some cases μ effectively fits to the limiting cases $\mu \to \infty$ or $\mu \to -\infty$ rather than a specific value. Because we only fit in the range $-20 < \mu < 0$ (see Appendix B.1), a value of $\mu \approx -20$ shows that the fit is subject to the $\mu \ll E$ degeneracy, while $\mu \approx 0$ shows that it is subject to the $\mu \gg E$ degeneracy. As mentioned above, the input to each fit is a collection of genomic TF binding sites $\{\sigma\}$ [162] and the EM predicted on the basis of high-throughput *in vitro* TF-DNA binding assays [166]. The EM allows us to assign a binding energy $E(\sigma)$ to each site.

A summary of maximum-likelihood parameter values for all TFs is shown in Table 3.1 and Appendix Table B.2, column D. The variation of log-likelihood with fitting parameters is shown in Appendix Table B.2, columns G and H. Since for many TFs relatively few

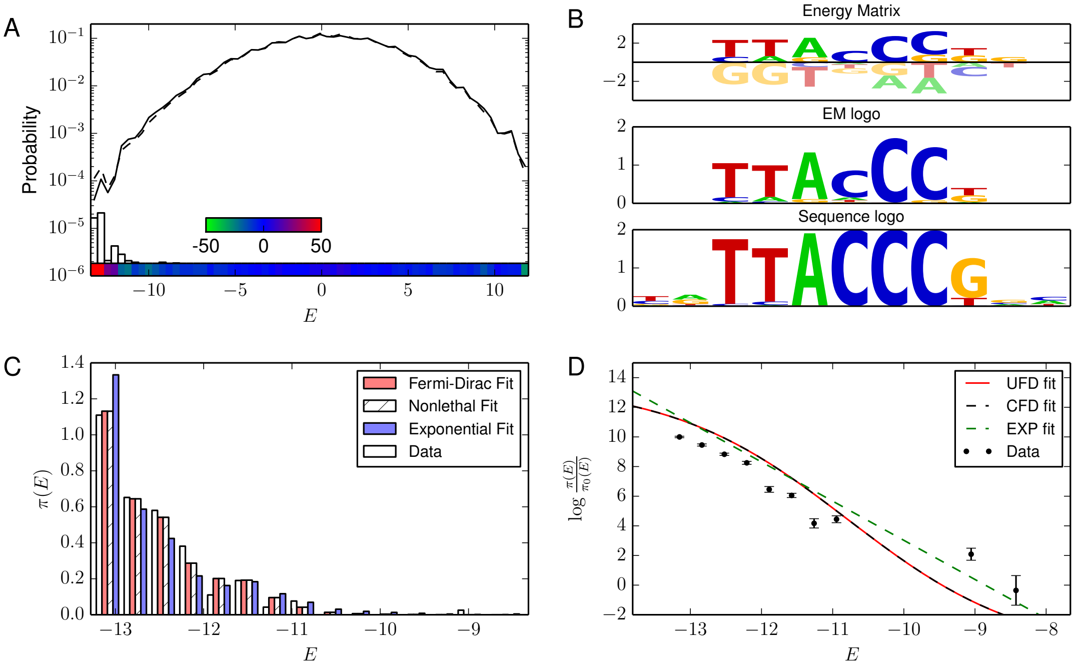


Figure 3.5: **Parametric inference of REB1 fitness landscape.** (A) Histogram of energies of intergenic sites calculated using the REB1 energy matrix (dashed line). The neutral distribution of sequence energies expected from the mono- and dinucleotide background model (solid line; see Appendix B.5 for details). The color bar on the bottom indicates the percent deviation between the two distributions (red is excess, green is depletion relative to the background model). The vertical bars show the distribution of functional sites [162], which correspond to the low-energy excess in the distribution of intergenic sites. (B) From top to bottom: REB1 energy matrix, the sequence logo obtained from the energy matrix by assuming a Boltzmann distribution at room temperature at each position in the binding site ($\pi^i(\sigma_i) = \pi_0^i(\sigma_i)e^{-\beta\epsilon_i^{\sigma_i}}/Z_i$), and the sequence logo based on the alignment of observed REB1 genomic sites. (C) Histogram of binding site energies and its prediction based on the three fits in (C) (Eq. 3.3). (D) Fitness function inference. Dots represent data points (as in Fig. 3.6); also shown are the unconstrained fit to the Fermi-Dirac function of Eq. 3.5 ("UFD"; solid red line), constrained fit to Eq. 3.5 with $f_0 = 0.99$ ("CFD"; dashed black line), and fit to an exponential fitness function ("EXP"; dashed green line). Error bars in (D) are calculated as in Fig. 3.6.

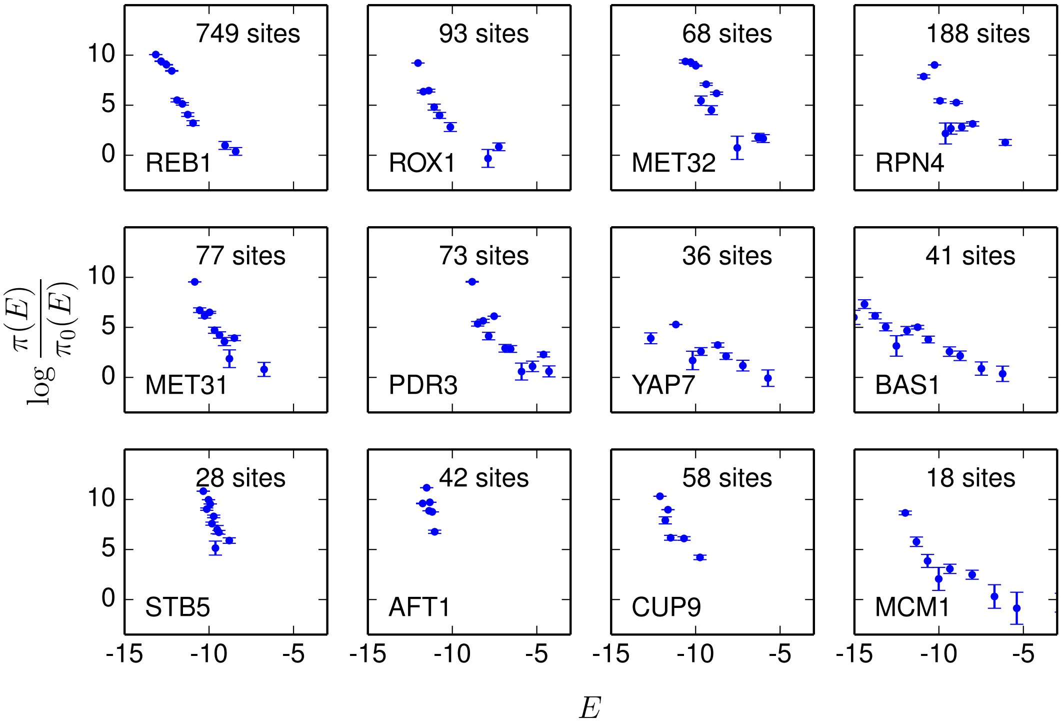


Figure 3.6: **Qualitative behavior of fitness landscapes.** Shown are plots of $\log(\pi(E)/\pi_0(E))$ for 12 TFs, which, according to Eq. 3.6, equals the logarithm of fitness up to an overall scale and shift. For each TF, sequences are grouped into 15 equal-size energy bins between the minimum and maximum energies allowed by the energy matrix. Shown also are the total number of binding sites for each TF. Error bars are calculated as $\sqrt{p(1-p)/n}$, where p is the fraction of sites falling in a given bin out of n total sites, as would be expected if the sequences were randomly distributed according to the observed distribution.

TF	f_0	$\gamma = \nu(1 - f_0)$	β (in (kcal/mol)$^{-1}$)	$E - \mu$
REB1	0.999	18.2	0.794	≈ 0
ROX1	0.993	335	0.398	> 0
MET32	0.973	133	0.251	> 0
RPN4	1.01×10^{-4}	1.58	2.0	≈ 0
MET31	4.54×10^{-5}	1.58	1.58	≈ 0
PDR3	0.988	242	0.251	> 0
YAP7	4.54×10^{-5}	1.58	1.0	≈ 0
BAS1	4.54×10^{-5}	2.51	0.501	≈ 0
STB5	0.401	150	0.316	< 0
AFT1	4.54×10^{-5}	3.98	5.01	≈ 0
CUP9	0.978	219	0.316	> 0
MCM1	0.998	83.1	0.251	> 0

Table 3.1: **Summary of unconstrained Fermi-Dirac landscape fits to TF binding site data.** Columns show maximum-likelihood value of f_0, $\gamma = \nu(1 - f_0)$, and β. The last column shows whether most binding site energies E are lower than the inferred chemical potential μ, near it, or above it (see Appendix Table B.2 for details).

binding sites are available, in Appendix Table B.3 we evaluate the goodness of fit using randomly chosen subsets of binding sites and Hessian analysis. Six of the TFs (REB1, ROX1, MET32, PDR3, CUP9, and MCM1) are in the $1 - f_0 \ll 1$ regime where only γ can be inferred unambiguously. Indeed, non-lethal Fermi-Dirac fits with $f_0 = 0.99$ yield very similar values of log-likelihood and γ (Appendix Table B.2, column D). In all of these cases, γ is considerably greater than 1, implying that selection is strong compared to drift, and the effective population size is large (the $s \ll 1$, $Ns \gg 1$ regime in population genetics).

Five TFs (RPN4, MET31, YAP7, BAS1, and AFT1) have very small values of f_0 (Table 3.1), indicating that on average, removing their binding sites is strongly deleterious to the cell. In these cases, the global maximum occurs in the vicinity of $f_0 = 0$, away from the degenerate region of parameter space (Appendix Table B.2, column H, insets). Note however that the likelihood surface is always degenerate in the region of parameter space with $1 - f_0 \ll 1$ and $\gamma =$ constant; this is true even when the global maximum likelihood does not occur in that region, as observed for these five TFs. Since $1 - f_0 \approx 1$, $\nu \approx \gamma$, which is a small value in four out of five cases (Table 3.1). Given the strength of selection, small effective population sizes (which indicate that genetic drift is strong) are necessary to reproduce the observed variation in binding site sequences. Finally, sites for STB5 have an intermediate value of $f_0 = 0.401$, which means they are under strong selection but are not necessarily essential.

The fits to the Fermi-Dirac fitness landscapes also provide estimates of the effective inverse temperature β (Table 3.1). The inferred values of β can be compared to the physical value at room temperature, $\beta_{\text{phys}} = 1.69$ (kcal/mol)$^{-1}$. Ten of the TFs (REB1, ROX1, MET32, MET31, PDR3, YAP7, BAS1, STB5, CUP9, MCM1) have β's lower than the physical value, while in the other two (RPN4, AFT1) $\beta > \beta_{\text{phys}}$. In most TFs the fitted inverse temperature β is far from its physical counterpart, although in several cases the likelihood function is fairly flat in the vicinity of the peak, indicating that a wider range of β values is admissible (Appendix Table B.2, column G).

The inferred value of μ relative to the distribution of energies E of the binding sites tells us in which qualitative regime of the Fermi-Dirac fitness landscape the sites lie. For five TFs (ROX1, MET32, PDR3, CUP9, MCM1), $E - \mu > 0$, and the sites reside on the exponential tail of the landscape and are subject to the $\mu \ll E$ degeneracy. For all of these TFs $1 - f_0 \ll 1$, as required by the degeneracy. For a group of six TFs (REB1, RPN4, MET31, YAP7, BAS1, AFT1), $E - \mu \approx 0$, so that the sites lie on the bound-unbound threshold. In this regime, changing the energy of the site through mutations may lead to a large change in fitness. Finally, $E - \mu < 0$ for STB5, and the sites lie on the high-fitness plateau and are subject to the $\mu \gg E$ degeneracy. The degeneracies in μ are also illustrated in Appendix Table B.2, column G.

What does $\beta \neq \beta_{\text{phys}}$ say about the nature and strength of selection? We address this question using the local selection coefficient, $\tilde{s}(E) = |d \log \mathcal{F}/dE|$ (Eq. 3.11). The magnitude of the selection coefficient depends qualitatively on both $E - \mu$ and whether f_0 is zero or nonzero (Fig. 3.1). For five of the TFs (ROX1, MET32, PDR3, CUP9, MCM1), $f_0 \neq 0$, $\beta < \beta_{\text{phys}}$, and $E - \mu > 0$. Thus these TFs are in a regime where decreasing

β strengthens selection (Fig. 3.1F). In other words, selection is stronger for these binding sites than expected from purely biophysical considerations. For RPN4 and AFT1, $f_0 \approx 0$, $\beta > \beta_{\text{phys}}$, and $E \approx \mu$. Hence $\partial\tilde{s}/\partial\beta > 0$, and selection is again stronger than expected. STB5 exhibits $\beta < \beta_{\text{phys}}$ and lies on the high fitness plateau ($E - \mu < 0$), and thus selection is also stronger than expected. In contrast, REB1, MET31, YAP7, and BAS1 exhibit $\beta < \beta_{\text{phys}}$ and lie on the threshold $E - \mu \approx 0$, and hence selection is weaker than expected in these three cases.

Fitness landscape model selection. Since the constrained Fermi-Dirac fits have one fewer adjustable parameter than the unconstrained fits, it is more consistent to do model selection on the basis of the Akaike information criterion (adjusted for finite-size samples) [179] rather than log-likelihoods:

$$\text{AIC} = 2(h - \log \mathcal{L}) + \frac{2h(h+1)}{n-h-1}, \tag{3.13}$$

where h is the number of fitting parameters, $\mathcal{L}$ is the likelihood, and n is the number of data points. For each model we can calculate the AIC, which accounts for both the benefits of higher log-likelihood and the costs of additional parameters. A better fit is reflected in a lower AIC value.

Table 3.2 shows the AIC differences between the unconstrained Fermi-Dirac fits (UFD, $h = 4$) and the constrained Fermi-Dirac fits with $f_0 = 0.99$ (CFD, $h = 3$) for each TF. Positive AIC differences indicate that UFD is more favorable. We also calculate the Akaike weights $w \propto e^{-\text{AIC}/2}$, which give the relative likelihood that a given model is the best [179].

For the six TFs in the $1 - f_0 \ll 1$ regime, the constrained Fermi-Dirac fits perform somewhat but not drastically better than the unconstrained Fermi-Dirac fits (Table 3.2). Indeed, the Akaike weights for the constrained Fermi-Dirac fits exceed the full fits for these TFs consistently by about a factor of $e \approx 2.7$, since their raw likelihoods are essentially equivalent and they only differ in the number of fitted parameters h. Out of the five TFs

TF	$\mathrm{AIC}_{\mathrm{CFD}} - \mathrm{AIC}_{\mathrm{UFD}}$	$\mathrm{AIC}_{\mathrm{EXP}} - \mathrm{AIC}_{\mathrm{UFD}}$	w_{UFD}	w_{CFD}	w_{EXP}
REB1	3.054	35.734	0.822	0.178	1.43×10^{-8}
ROX1	-2.042	34.753	0.265	0.735	7.53×10^{-9}
MET32	-2.233	10.540	0.246	0.752	0.001
RPN4	5.674	19.959	0.945	0.055	4.38×10^{-5}
MET31	-1.469	-3.869	0.100	0.208	0.692
PDR3	-2.124	6.123	0.254	0.734	0.012
YAP7	-2.049	10.722	0.264	0.735	0.001
BAS1	-2.107	1.061	0.224	0.644	0.132
STB5	-2.732	-7.145	0.025	0.097	0.879
AFT1	-2.069	6.096	0.259	0.729	0.012
CUP9	-2.251	1.560	0.220	0.679	0.101
MCM1	-3.343	-0.175	0.135	0.718	0.147

Table 3.2: **Comparison of fitness function models.** For each TF, shown are the AIC differences between the unconstrained Fermi-Dirac fit ("UFD"), the constrained Fermi-Dirac fit with $f_0 = 0.99$ ("CFD"), and the exponential fit ("EXP"). Also shown are Akaike weights w, which indicate the relative likelihood of each model.

for which $f_0 \approx 0$, YAP7, BAS1, and AFT1 fit slightly better to the constrained Fermi-Dirac, suggesting that their fitted values of f_0 are not significant. For RPN4 and MET31, the AIC analysis shows preference for the fits with low f_0. This preference is especially strong for RPN4. Both RPN4 and MET31 are listed as nonessential in the Yeast Deletion Database [167, 175], suggesting either an inconsistency in our analysis or that growth media tested in Refs. 167, 175 do not reveal essentiality of these TFs.

We may also consider a purely exponential fitness landscape of the form $\mathcal{F}(E) = e^{\alpha E}$. The reasons for including this case are threefold. First, exponential fitness emerges in the limit $E - \mu \gg 0$ of the Fermi-Dirac landscape, the regime into which many of the TF binding sites fall. Second, the fitness landscapes in Fig. 3.6 appear close to linear on the logarithmic scale, implying that to a good approximation fitness depends exponentially on energy. Third, the model has just one fitting parameter, making it a useful null case for AIC evaluation.

The steady-state distribution $\pi(\sigma)$ with exponential fitness is given by

$$\pi(\sigma) = \frac{1}{Z}\pi_0(\sigma)e^{\nu\alpha E(\sigma)} = \prod_{i=1}^{L} \frac{\pi_0^i(\sigma_i)}{Z_i} e^{\nu\alpha\epsilon_i^{\sigma_i}}, \tag{3.14}$$

where $E(\sigma)$ is given by Eq. 3.2, $\pi_0(\sigma)$ is the neutral probability of sequence σ, $\pi_0^i(\sigma_i)$ is the background probability of nucleotide σ_i at position i, and Z_i is a single-site partition function: $\pi_0(\sigma)/Z = \prod_{i=1}^{L} \pi_0^i(\sigma_i)/Z_i$. Here we assumed that the background probability of a sequence is a product of probabilities of its constituent nucleotides. In this case, positions in the binding site decouple and the distribution of sites $\pi(\sigma)$ completely factorizes. The assumption of factorization underlies the common practice of inferring energy matrices from log-odds scores of observed genomic binding sites [163]. The log-odds score of a nucleotide σ_i is defined as

$$S(\sigma_i) = \log \frac{p_i^{\sigma_i}}{\pi_0^i(\sigma_i)} = -\beta\epsilon_i^{\sigma_i} - \log Z_i, \tag{3.15}$$

where $p_i^{\sigma_i}$ is the probability of observing base $\sigma_i \in \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ at position i within the set of known sites, β is an effective inverse temperature, and Z_i is the normalization constant. Equation 3.15 shows that the log-odds score, which is computed using observed nucleotide probabilities, is equivalent to $\epsilon_i^{\sigma_i}$ (up to an overall scale and shift) under the assumption of site independence.

We can quantitatively compare the exponential fitness landscape with the unconstrained and constrained Fermi-Dirac landscapes using the Akaike information criterion, Eq. 3.13. The AIC analysis shows that the exponential landscape is significantly poorer than the Fermi-Dirac landscape in all cases except MET31 (Table 3.2), where the exponential fit is marginally better than the Fermi-Dirac fits, and STB5, where the exponential landscape does perform much better than the Fermi-Dirac models. This observation provides statistical support for the fitness landscapes of Fermi-Dirac type, and for the non-lethality of deleting most TFs (the exponential fitness decays to zero rather than a nonzero f_0 found in most of our Fermi-Dirac fits).

3.4 Discussion

In this work, we have considered how fitness of a single-cell eukaryote *S. cerevisiae* is affected by interactions between TFs and their cognate genomic sites. Changing the energy of a site, or creating new sites in gene promoters may change how genes are activated and repressed, which in turn alters the cell's chances of survival. Under the assumptions of a haploid monomorphic population in which the evolution of binding sites has reached steady state, the fitness landscape as a function of TF binding energy can be inferred from the distribution of TF binding sites observed in the genome, using a biophysical model which assigns binding energies to sites. We use a simple energy matrix model of TF-DNA energetics in which the energy contribution of each position in the site is independent of all the other positions. The energy matrix parameters are inferred from a high-throughput data set in which TF-DNA interactions were studied *in vitro* using a microfluidics device [166]. We consider two types of fitness functions: Fermi-Dirac, which appears naturally from considering TF binding as a two-state process (Eq. 3.1), and exponential, which is motivated by the observation that for many TFs, the logarithm of fitness appears to decrease linearly as energy increases.

A single fitness landscape for all genomic binding sites of a given TF can only exist in the absence of site-specific selection. Indeed, it is possible that TF sites experience different selection pressures depending on the genes they regulate: for example, sites in promoters of essential genes may be penalized more for deviating from the consensus sequence. In this case, the fitness function is an average over all sites which evolve under different selection constraints: as an extreme example, consider the case where each site i has a Fermi-Dirac fitness function (Eq. 3.5) with different parameters μ_i, β_i, and f_0^i. The resulting observed distribution of energies would then be the average of the distributions predicted by Eq. 3.3:

$$\pi(E) = \frac{1}{Z}\pi_0(E)\langle \mathcal{F}(E;\mu_i,\beta_i,f_0^i)^{\nu}\rangle_i \equiv \frac{1}{Z}\pi_0(E)\mathcal{F}(E;\bar{\mu},\bar{\beta},\bar{f}_0)^{\bar{\nu}}, \tag{3.16}$$

which defines the "average" fitness function with effective parameters $\bar{\mu}$, $\bar{\beta}$, $\bar{f}_0$, $\bar{\nu}$. Thus the fit can be carried out even in the presence of site-dependent selection, but the fitted parameters correspond to fitness functions of individual sites only in an average sense.

In order to gauge the importance of site-specific selection in TF binding site evolution, we have performed several statistical tests aimed at discovering correlations between binding site energies and biological properties of the sites and the genes they regulate. These tests considered gene essentiality, growth rates of strains with nonessential genes knocked out, gene expression levels, K_A/K_S ratios based on alignments with *S. paradoxus*, and the distance of the site to the TSS. We find no consistent correlations among these properties, indicating that for a given TF, the evolution of regulatory sites is largely independent of the properties of regulated genes and the specific biological functions of the sites.

Previously, low correlations have been observed between essentiality and conservation of protein and coding sequences [180–186], which has fueled considerable speculation as it contradicts the prediction of the neutral theory of evolution that higher selection pressures lead to lower evolutionary rates. It has also been found that the growth rate of strains with nonessential genes knocked out are significantly (though weakly) correlated with conservation of those genes [187]. It has therefore been suggested that selection pressures are so strong that only the most nonessential genes experience significant genetic drift [180]. Previous studies have also found that gene expression levels are a more reliable (though still very weak) predictor of selection pressures than essentiality [184], but we do not find this to be the case for TF binding sites, nor do we observe a significant correlation between gene expression levels and TF binding energies.

Available data does not rule out the possibility of time-dependent selection in combination with forms of site-dependent selection for which we have not accounted. In this scenario, the variation in site binding affinity is not due to genetic drift, but to variable selection pressures across sites and over time, such that the sites are strongly tuned to particular binding energies which change from locus to locus. Indeed, there is evidence that there is frequent gain and loss of TF binding sites and that the gene regulatory network is highly dynamic [188–194]. However, it is possible that rapid turnover of binding sites in eukaryotes may be due to evolution acting on whole promoters rather than individual binding sites. Many promoters contain multiple binding sites for a single TF, and it may be that while individual binding sites are lost and gained frequently, the overall binding affinity of a promoter to a TF may be held constant [195–197]. Our evolutionary model can account for this scenario using a promoter-level fitness function, which we intend to consider in future work.

Out of 12 TFs with sufficient binding site data, five have $f_0 \approx 0$, indicating a large fitness penalty for deleting such sites. However, this conclusion is strongly supported by the AIC differences between unconstrained and non-lethal Fermi-Dirac fits for only one TF, RPN4 (Table 3.2). RPN4 is classified as nonessential in the Yeast Deletion Database. It may be that this misclassification is due to a mismatch between genomic sites, in which the core GCCACC motif is preceded by TTT, and the energy matrix in which the binding energies upstream of the core motif are non-specific. We also classify REB1 and MCM1 binding sites as nonessential, although knocking out these TFs is lethal in yeast. This discrepancy may be due to a minority of essential sites being averaged with the majority of nonessential sites to produce a single fitness function, as described above. In addition, although a penalty for deleting any single site may be small, the cumulative penalty for deleting all sites (or, equivalently, deleting the TF) may be lethal. Overall, on the basis of AIC we classify 8 out of 12 TFs correctly (Table 3.2).

We find that in 11 out of 12 cases, fitting an exponential fitness function is less supported by the data than fitting a Fermi-Dirac function (Table 3.2). This is interesting since constructing a position-specific weight matrix by aligning genomic sites is a common practice which implicitly assumes factorization of exponential fitness and independence of each position in the binding site. Our results show the limitations of this approximation. It is important to note that a key difference between the Fermi-Dirac fitness landscape and the exponential landscape is that the former contains magnitude epistasis [11, 32] (i.e., the magnitude of a mutation's fitness effect depends on the background sequence), while the latter is non-epistatic. Thus, our results indicate that epistasis is widespread in the evolution of TF binding sites [72].

Finally, we find that depending on the TF, the distribution of TF binding energies may fall on the exponential tail, across the threshold region, or on the saturated plateau where the sites are always occupied (Table 3.1). In the first two categories, variation of TF concentration in the cell will lead to graded responses, which may be necessary to achieve precise and coordinated gene regulation. In the third regime, TF binding is robust and not dynamic. We also find that the fitted inverse temperature β is typically not close to the value based on room temperature (Table 3.1). This observation suggests selection pressures in addition to those dictated by the energetics of TF binding to its cognate sites.

Chapter 4

Statistical Physics of Stochastic Paths

In the previous two chapters we focused on populations in steady state. While some systems may be reasonably described by steady state, many important questions in evolution are inherently dynamical. As discussed in Chapter 1, we are interested not just in the eventual fates of populations, but also the pathways they take to get there. This is of particular importance for understanding how predictable evolution is [85]. In this chapter we take a detour to develop some fundamental tools from stochastic processes and statistical physics, which we will use to understand the evolutionary paths of populations. This chapter is drawn from Refs. 10, 11, as well as some unpublished material.

A **stochastic process** describes a system whose dynamics over time is probabilistic rather than deterministic [131]. In physics the most important example of a stochastic process is a **random walk** [198], historically used to describe the motion of a large particle suspended in a fluid at finite temperature [5]. Although the traditional notion of a random walk describes the physical motion of a particle in space, by analogy we can think of any stochastic system as randomly walking through its state space. In this sense random walks are ubiquitous across physics, chemistry, and biology, including molecular evolution [31–33], protein folding [199], chemical reactions [200], transport and search in complex media [201, 202], stochastic phenotypes [203], and cell-type differentiation [42, 204, 205].

In this chapter we develop a formalism for stochastic processes that explicitly decomposes them into sums over all possible paths [10, 11, 206–210]. The formalism is general to any continuous-time random walk on a discrete state space with arbitrary complexity (e.g., a simple lattice or a complex network). The approach is particularly well-suited for obtaining statistics that describe the diversity of paths, such as the distribution of path lengths and path entropy. We will first describe the general formalism along with an efficient numerical implementation, then illustrate the method on a few simple examples.

4.1 The ensemble of stochastic paths

Let $\mathcal{S}$ be a discrete set of possible states for a system. A **semi-Markov process** (also known as a **continuous-time random walk** [198]) on $\mathcal{S}$ consists of jumps between states and continuous-time waiting within states. Let $\mathbf{Q}$ be the jump matrix, with $\langle\sigma'|\mathbf{Q}|\sigma\rangle$ being the probability of jumping to σ' given the current state σ ($\sigma, \sigma' \in \mathcal{S}$). The space $\mathcal{S}$ equipped with the jump matrix $\mathbf{Q}$ defines a network with directed and weighted edges, so we can think of this process as a random walk on the network of states. The waiting time distributions will be denoted by $\psi(t|\sigma)$, which is the probability density of waiting exactly time t in state σ before jumping out. Define $\theta(\sigma)$ to be the mean of $\psi(t|\sigma)$; we will generally assume this is finite for all σ, although for cases when it is not (e.g., $\psi(t|\sigma)$ is a power law distribution) it only invalidates results having to do with mean times. We also define the "partial waiting" probability $\Psi(t|\sigma) = \int_t^\infty dt' \psi(t'|\sigma)$ as the probability of waiting *at least* time t in state σ.

We say this process is "semi-Markovian" in the sense that the jump process is **memoryless** — the probability of jumping to a new state depends only on the current state — but the waiting process may not be. Memoryless waiting means that the probability of waiting an additional time t in a state, given the system has already waited time t_0, equals the probability of just waiting time t in the first place (there is no "memory" of having waited the original time t_0):

$$\frac{\Psi(t + t_0|\sigma)}{\Psi(t_0|\sigma)} = \Psi(t|\sigma). \tag{4.1}$$

The only function satisfying this condition is the exponential function; hence a fully-Markov process has partial waiting probabilities $\Psi(t|\sigma) = e^{-t/\theta(\sigma)}$ and waiting time distributions

$\psi(t|\sigma) = e^{-t/\theta(\sigma)}/\theta(\sigma)$. Non-Markov waiting time distributions will be non-exponential and can arise due to coarse-graining a Markov process [211–213].

A general semi-Markov process is completely specified by the jump matrix $\mathbf{Q}$ and the waiting time distributions $\psi(t|\sigma)$. The special case of a fully-Markov process, however, is usually specified by a rate matrix $\mathbf{W}$, and the probability distribution $|\pi(t)\rangle$ of system states at time t is given by the master equation as defined in Eq. 1.6. This was the approach of our evolutionary model in Chapters 1 and 2. The mean waiting times are related to the rate matrix via

$$\theta(\sigma) = -\frac{1}{\langle\sigma|\mathbf{W}|\sigma\rangle} = \left(\sum_{\sigma'}\langle\sigma'|\mathbf{W}|\sigma\rangle\right)^{-1}, \tag{4.2}$$

i.e., the inverse total escape rate from σ, while the jump probabilities are

$$\langle\sigma'|\mathbf{Q}|\sigma\rangle = \langle\sigma'|\mathbf{W}|\sigma\rangle\theta(\sigma). \tag{4.3}$$

It is possible to write a generalized master equation, analogous to Eq. 1.6, for the semi-Markov case [206, 208, 214]; however we omit it here because we will instead study the dynamics of the process in the path formalism.

4.1.1 The path probability functional

Define a **path** through state space as the time-ordered sequence of states $\varphi = \{\sigma_0, \sigma_1, \ldots, \sigma_\ell\}$. Suppose the system spends times $t_0, t_1, \ldots, t_\ell$ waiting in each state along the path. The probability functional of starting in the initial state σ_0 and completing the path φ (reaching the final state σ_ℓ) no later than time t is

$$\Pi_{\mathrm{FP}}[\varphi, t] = \pi(\sigma_0)\left(\prod_{i=0}^{\ell-1}\langle\sigma_{i+1}|\mathbf{Q}|\sigma_i\rangle\right)\left(\prod_{i=0}^{\ell-1}\int_0^\infty dt_i\ \psi(t_i|\sigma_i)\ \Theta\left(t - \sum_{i=0}^{\ell-1} t_i\right)\right), \tag{4.4}$$

where the first factor is the initial state probability $\pi(\sigma_0)$, the second is the product of jump probabilities, and the third integrates over waiting times while constraining the total waiting time to be less than t (Θ is the Heaviside step function). We will refer to this as the "first-passage" path probability (hence the subscript FP), since it considers the path finished once it has reached the final state. The Laplace transform of Eq. 4.4 results in a simpler expression through deconvolution [206]:

$$\tilde{\Pi}_{\mathrm{FP}}[\varphi, s] = \int_0^\infty dt\; e^{-st}\; \Pi_{\mathrm{FP}}[\varphi, t] = \frac{\pi(\sigma_0)}{s} \prod_{i=0}^{\ell-1} \langle \sigma_{i+1} | \mathbf{Q} | \sigma_i \rangle \; \tilde{\psi}(s|\sigma_i), \tag{4.5}$$

where $\tilde{\psi}(s|\sigma_i)$ is the Laplace transform of $\psi(t|\sigma_i)$. For a fully-Markov process,

$$\tilde{\psi}(s|\sigma) = \frac{1}{1 + s\theta(\sigma)}, \qquad \tilde{\Psi}(s|\sigma) = \frac{\theta(\sigma)}{1 + s\theta(\sigma)}, \tag{4.6}$$

and therefore [207]

$$\tilde{\Pi}_{\mathrm{FP}}[\varphi, s] = \frac{\pi(\sigma_0)}{s} \prod_{i=0}^{\ell-1} \frac{\langle \sigma_{i+1} | \mathbf{Q} | \sigma_i \rangle}{1 + s\theta(\sigma_i)}. \tag{4.7}$$

Other definitions of the path probability functional may be more appropriate in some cases and are straightforwardly constructed in s-space. For instance, we may want to consider the probability of observing the system still waiting in final state σ_ℓ at time t after having taken the path φ. We will consider this to be the path propagator $\Pi_{\mathrm{prop}}[\varphi, t]$, since it is related to the overall propagator for the system (defined below). In s-space this is

$$\tilde{\Pi}_{\mathrm{prop}}[\varphi, s] = \pi(\sigma_0) \tilde{\Psi}(s|\sigma_\ell) \prod_{i=0}^{\ell-1} \langle \sigma_{i+1} | \mathbf{Q} | \sigma_i \rangle \; \tilde{\psi}(s|\sigma_i). \tag{4.8}$$

The factor of $\tilde{\Psi}(s|\sigma_\ell)$ accounts for the extra waiting in final state σ_ℓ. For the fully-Markov process,

$$\tilde{\Pi}_{\mathrm{prop}}[\varphi, s] = \pi(\sigma_0) \frac{\theta(\sigma_\ell)}{1 + s\theta(\sigma_\ell)} \prod_{i=0}^{\ell-1} \frac{\langle \sigma_{i+1} | \mathbf{Q} | \sigma_i \rangle}{1 + s\theta(\sigma_i)}. \tag{4.9}$$

Finally, in the $t \to \infty$ limit we obtain the probability of the path φ for any duration (which does not depend on the waiting time distributions):

$$\Pi_\infty[\varphi] = \pi(\sigma_0) \prod_{i=0}^{\ell-1} \langle \sigma_{i+1} | \mathbf{Q} | \sigma_i \rangle, \tag{4.10}$$

which is just the product of jump probabilities.

4.1.2 Path ensemble averages

Usually we are interested in not a single path but an ensemble Φ of paths that define some dynamical process; for example, this may be all paths from a set of initial states $\mathcal{S}_i$ to a set of final states $\mathcal{S}_f$. We are especially interested in ensembles of **first-passage** paths, defined as paths that reach the final states only once. The partition function for the ensemble Φ is the sum over all path probabilities:

$$\mathcal{Z}_\Phi(t) = \sum_{\varphi \in \Phi} \Pi[\varphi, t]. \tag{4.11}$$

For the FP path functional in Eq. 4.4, the partition function represents the total probability of reaching $\mathcal{S}_f$ from $\mathcal{S}_i$ by time t via paths in Φ. For the path propagator (Eq. 4.8), the partition function represents the total probability of observing the system in any final state at time t. For the fully-Markov case, the standard propagator is $\mathbf{P}(t) = e^{t\mathbf{W}}$, where $\langle \sigma' | \mathbf{P}(t) | \sigma \rangle$ gives the total probability of observing the system in state σ' at time t given an initial state of σ. (This is essentially the solution to the Markov master equation, i.e., Eq. 1.9.) Hence the path partition function must equal this propagator:

$$\langle \sigma' | e^{t\mathbf{W}} | \sigma \rangle = \sum_{\varphi \in \Phi} \Pi_{\mathrm{prop}}[\varphi, t], \tag{4.12}$$

where Φ is the ensemble of all paths beginning in σ and ending in σ'. We give a simple proof of this path decomposition for the fully-Markov case in Appendix C.

Besides total probabilities, we are interested in average properties of the path ensemble. We define the following path functionals:

$$\mathcal{L}[\varphi] = \text{length (number of jumps) of } \varphi, \qquad \mathcal{I}_\sigma[\varphi] = \begin{cases} 1 \text{ if } \sigma \in \varphi, \\ 0 \text{ otherwise,} \end{cases}$$
$$\mathcal{T}[\varphi] = \sum_{i=0}^{\ell-1} \theta(\sigma_i), \qquad \mathcal{T}_\sigma[\varphi] = \sum_{i=0}^{\ell-1} \delta_{\sigma,\sigma_i}\theta(\sigma_i), \tag{4.13}$$

where δ is the Kronecker delta. We can now express various path statistics as averages of these functionals over the ensemble. We use a generic path functional $\Pi[\varphi, t]$ in these definitions; Eq. 4.4, Eq. 4.8, or other constructions are used depending on the problem of interest. The average path time is given by [210]

$$\bar{t}_\Phi(t) = \langle \mathcal{T}(t) \rangle_\Phi = \frac{1}{\mathcal{Z}_\Phi(t)} \sum_{\varphi\in\Phi} \mathcal{T}[\varphi]\Pi[\varphi, t]. \tag{4.14}$$

The distribution of path lengths is given by

$$\rho_\Phi(\ell, t) = \frac{1}{\mathcal{Z}_\Phi(t)} \sum_{\varphi\in\Phi} \delta_{\ell,\mathcal{L}[\varphi]}\Pi[\varphi, t], \tag{4.15}$$

from which the average length $\bar{\ell}_\Phi(t) = \langle \mathcal{L}(t) \rangle_\Phi$ and standard deviation of length $\ell^{\text{sd}}_\Phi(t)$ are readily obtained.

Averages over state-dependent functionals can be used to characterize the spatial structure of paths. For example, the fraction of time paths spend in a state σ can be expressed as $\langle \mathcal{T}_\sigma(t) \rangle_\Phi / \bar{t}_\Phi(t)$; this is a normalized distribution over all states $\sigma \in \mathcal{S}$ and therefore it represents the density of states on the paths in the ensemble Φ. The quantity $\langle \mathcal{T}_\sigma(t) \rangle_\Phi / \theta(\sigma)$ gives the average number of visits to state σ. The probability that a path will visit a state

σ at all is given by $\langle \mathcal{I}_\sigma(t) \rangle_\Phi$, which we will refer to as the density of paths in the ensemble Φ. We can also construct the two-point correlation function $\langle \mathcal{I}_{\sigma'}(t) \mathcal{I}_\sigma(t) \rangle_\Phi$, which gives the probability of paths passing through both σ and σ'.

In many cases we are interested in the time-independent versions of these quantities, i.e., statistical properties of paths taking any amount of time to finish. These can be obtained as the $t \to \infty$ limit of the above expressions, which amounts to replacing $\Pi[\varphi, t]$ with $\Pi_\infty[\varphi]$ (Eq. 4.10). We will denote these time-independent properties by simply omitting the time dependence, e.g., $\lim_{t\to\infty} \bar{t}_\Phi(t) = \bar{t}_\Phi$. The convergence of these limits depends on the path ensemble Φ. For example, if Φ includes all possible paths connecting the initial and final states, including those that visit the final state multiple times, then these limits generally diverge: typically, there is no finite average time or length for these paths. However, restriction to first-passage paths in Φ, as is often our focus, guarantees convergence.

This formalism also allows for development of path thermodynamics. The entropy of the path ensemble is given by

$$\begin{aligned} S_\Phi(t) &= -\frac{1}{\mathcal{Z}_\Phi(t)} \sum_{\varphi \in \Phi} \Pi[\varphi, t] \log \left(\frac{\Pi[\varphi, t]}{\mathcal{Z}_\Phi(t)} \right) \\ &= -\langle \log \Pi(t) \rangle_\Phi + \log \mathcal{Z}_\Phi(t). \end{aligned} \tag{4.16}$$

Indeed, if we define the path Hamiltonian to be

$$\mathcal{H}[\varphi, t] = -\log(\Pi[\varphi, t]), \tag{4.17}$$

(so that $\Pi[\varphi, t] = e^{-\mathcal{H}[\varphi, t]}$), we can express the path ensemble free energy as

$$F_\Phi(t) = \langle \mathcal{H}(t) \rangle_\Phi - S_\Phi(t) = -\log \mathcal{Z}_\Phi(t). \tag{4.18}$$

The partition function $\mathcal{Z}_\Phi(t)$ monotonically increases with time. Therefore the free energy $F_\Phi(t)$ monotonically decreases as $t \to \infty$, corresponding to equilibration of the path ensemble.

For recurrent processes (i.e., where the system will almost surely reach the final states eventually [212]), $\lim_{t\to\infty} \mathcal{Z}_\Phi(t) = \mathcal{Z}_\Phi = 1$, and hence equilibrium free energy is zero. In these cases, equilibrium path entropy is equal to the average Hamiltonian. If the ensemble Φ consists of only a single path with nonzero probability, its entropy is $S_\Phi = 0$. This situation may arise if the process is so constrained that only a single viable pathway exists between the initial and final states. In contrast, consider a purely random walk on a homogeneous network with γ nearest neighbors per node. The jump probability between any pair of neighboring nodes is thus γ^{-1}, so any path φ has probability $\Pi_\infty[\varphi] = \gamma^{-\mathcal{L}[\varphi]}$, and the entropy of the ensemble is given by

$$S_\Phi = -\langle \log \Pi_\infty \rangle_\Phi = \bar{\ell}_\Phi \log \gamma. \tag{4.19}$$

Thus the path entropy depends on two distinct factors: the average path length $\bar{\ell}$ and the network connectivity γ. Note that path entropy and the average path Hamiltonian scale with the average path length, which defines a notion of extensivity in the path ensemble. This is sensible if we think of a path as a gas of particles, where each jump in the path corresponds to a particle. The path ensemble, which includes paths of many lengths, therefore is equivalent to the grand canonical ensemble of the gas. In the case of the gas, extensive quantities like entropy and energy scale with the number of particles, and hence these quantities here scale with the path length.

4.2 Simple analytical examples

We now illustrate the path formalism on two simple examples.

4.2.1 Two-state system

Suppose the state space $\mathcal{S}$ consists only of two states, σ_1 and σ_2, with rate matrix

$$\mathbf{W} = \begin{bmatrix} -\lambda_1 & \lambda_2 \\ \lambda_1 & -\lambda_2 \end{bmatrix}, \tag{4.20}$$

where $\lambda_1, \lambda_2 > 0$. So the total propagator is given by the matrix exponential:

$$\mathbf{P}(t) = e^{t\mathbf{W}} = \frac{1}{\lambda_1 + \lambda_2} \begin{bmatrix} \lambda_2 + \lambda_1 e^{-(\lambda_1+\lambda_2)t} & \lambda_2(1 - e^{-(\lambda_1+\lambda_2)t}) \\ \lambda_1(1 - e^{-(\lambda_1+\lambda_2)t}) & \lambda_1 + \lambda_2 e^{-(\lambda_1+\lambda_2)t} \end{bmatrix}. \tag{4.21}$$

Let us now compute the matrix element $\langle\sigma_2|\mathbf{P}(t)|\sigma_1\rangle$ using the path expansion. The transition $\sigma_1 \to \sigma_2$ can occur through the following paths:

$$\begin{gathered} \sigma_1 \to \sigma_2 \\ \sigma_1 \to \sigma_2 \to \sigma_1 \to \sigma_2 \\ \sigma_1 \to \sigma_2 \to \sigma_1 \to \sigma_2 \to \sigma_1 \to \sigma_2 \\ \vdots \end{gathered} \tag{4.22}$$

In s-space the path expansion for the propagator has the form

$$\begin{aligned} \langle\sigma_2|\tilde{\mathbf{P}}(s)|\sigma_1\rangle &= \frac{\lambda_1}{(s+\lambda_1)(s+\lambda_2)} + \frac{\lambda_1^2\lambda_2}{(s+\lambda_1)^2(s+\lambda_2)^2} + \frac{\lambda_1^3\lambda_2^2}{(s+\lambda_1)^3(s+\lambda_2)^3} + \cdots \\ &= \frac{\lambda_1}{(s+\lambda_1)(s+\lambda_2)} \sum_{\ell=0}^{\infty} \left(\frac{\lambda_1\lambda_2}{(s+\lambda_1)(s+\lambda_2)}\right)^{\ell} \\ &= \left(\frac{\lambda_1}{(s+\lambda_1)(s+\lambda_2)}\right)\left(\frac{(s+\lambda_1)(s+\lambda_2)}{(s+\lambda_1)(s+\lambda_2) - \lambda_1\lambda_2}\right) \\ &= \frac{\lambda_1}{(s+\lambda_1)(s+\lambda_2) - \lambda_1\lambda_2}. \end{aligned} \tag{4.23}$$

Carrying out the inverse Laplace transform,

$$\langle\sigma_2|\mathbf{P}(t)|\sigma_1\rangle = \frac{1}{2\pi i}\oint ds\ e^{ts}\langle\sigma_2|\tilde{\mathbf{P}}(s)|\sigma_1\rangle = \frac{\lambda_1}{\lambda_1+\lambda_2}(1-e^{-(\lambda_1+\lambda_2)t}). \tag{4.24}$$

This matches the matrix exponential solution (Eq. 4.21).

We can also calculate the average number of jumps $\bar{\ell}(t)$ for the transition $\sigma_1 \to \sigma_2$ as a function of time:

$$\begin{aligned}\bar{\ell}(t) &= \frac{1}{\mathcal{Z}(t)}\sum_{\varphi}\mathcal{L}[\varphi]\Pi[\varphi,t] \\ &= \left(\frac{\lambda_1+\lambda_2}{\lambda_1(1-e^{-(\lambda_1+\lambda_2)t})}\right)\frac{1}{2\pi i}\oint ds\ e^{ts}\frac{\lambda_1}{(s+\lambda_1)(s+\lambda_2)}\sum_{\ell=0}^{\infty}(2\ell+1)\left(\frac{\lambda_1\lambda_2}{(s+\lambda_1)(s+\lambda_2)}\right)^{\ell} \\ &= \frac{1}{(\lambda_1+\lambda_2)^2}\left((\lambda_1-\lambda_2)^2+2\lambda_1\lambda_2(\lambda_1+\lambda_2)t\coth\left(\frac{1}{2}(\lambda_1+\lambda_2)t\right)\right)\end{aligned} \tag{4.25}$$

In the $t \gg (\lambda_1+\lambda_2)^{-1}$ limit, $\coth((\lambda_1+\lambda_2)t/2) \approx 1$ and we get the expected linear time dependence:

$$\bar{\ell}(t) \approx \frac{2\lambda_1\lambda_2}{(\lambda_1+\lambda_2)}t. \tag{4.26}$$

4.2.2 Random walk in one dimension

We now consider a slightly more complex example: an asymmetric random walk on a one-dimensional lattice. Thus the states can be labeled by integers. Suppose the jump rates are

$$\langle n'|\mathbf{W}|n\rangle = \begin{cases}\alpha & \text{if } n' = n+1, \\ \beta & \text{if } n' = n-1, \\ 0 & \text{otherwise.}\end{cases} \tag{4.27}$$

Without loss of generality we consider paths starting at 0 and ending at $n > 0$. Then any path $0 \to n$ must involve $n+\ell$ rightward jumps and ℓ leftward jumps, for $\ell \in \{0, 1, \ldots\}$. The number of paths of length $n + 2\ell$ is

$$\frac{(n+2\ell)!}{(n+\ell)!\ell!} = \binom{n+2\ell}{\ell}, \tag{4.28}$$

since each path can be uniquely specified by a sequence of right and left jumps. Then the propagator path expansion is

$$\langle n|\tilde{\mathbf{P}}(s)|0\rangle = \sum_{\ell=0}^{\infty} \binom{n+2\ell}{\ell} \frac{\alpha^{n+\ell}\beta^{\ell}}{(s+\alpha+\beta)^{n+2\ell+1}}. \tag{4.29}$$

We carry out the inverse Laplace transform term-by-term to obtain

$$\langle n|\mathbf{P}(t)|0\rangle = e^{-(\alpha+\beta)t} \sum_{\ell=0}^{\infty} \binom{n+2\ell}{\ell} \frac{\alpha^{n+\ell}\beta^{\ell}}{(n+2\ell)!} t^{n+2\ell}. \tag{4.30}$$

We can rewrite this in terms of a modified Bessel function [215]:

$$\langle n|\mathbf{P}(t)|0\rangle = e^{-(\alpha+\beta)t} \left(\frac{\alpha}{\beta}\right)^{n/2} \sum_{\ell=0}^{\infty} \frac{(\sqrt{\alpha\beta}t)^{n+2\ell}}{(n+\ell)!\ell!} = e^{-(\alpha+\beta)t} \left(\frac{\alpha}{\beta}\right)^{n/2} I_n(2\sqrt{\alpha\beta}t). \tag{4.31}$$

This agrees with the standard solution of the asymmetric random walk using the characteristic function [131]. We can also calculate the average number of steps:

$$\begin{aligned} \bar{\ell}(t) &= e^{-(\alpha+\beta)t} \left(\frac{\alpha}{\beta}\right)^{n/2} \sum_{\ell=0}^{\infty} (n+2\ell) \frac{(\sqrt{\alpha\beta}t)^{n+2\ell}}{(n+\ell)!\ell!} \\ &= e^{-(\alpha+\beta)t} \left(\frac{\alpha}{\beta}\right)^{n/2} (nI_n(2\sqrt{\alpha\beta}t) + 2\sqrt{\alpha\beta}tI_{n+1}(2\sqrt{\alpha\beta}t)). \end{aligned} \tag{4.32}$$

4.3 Transfer-matrix numerical algorithm

Unfortunately, analytical implementations of the path expansion appear to be limited to very simple cases such as those in the previous section due to the difficulty of enumerating over all paths. (For a more sophisticated analytical example, see Ref. 206.) However, the factorized form of the path probability distribution functional (Eqs. 4.5, 4.7, and 4.10) permits efficient numerical calculation of path ensemble averages via a recursive algorithm based on transfer matrices. Here for simplicity we consider the time-independent case, and thus assume that Φ consists of first-passage paths to guarantee convergence of path averages. Let $|\pi\rangle = \sum_\sigma \pi(\sigma)|\sigma\rangle$ be the vector of initial state probabilities. For each jump ℓ and intermediate state σ, we calculate the transfer matrix elements $P_\ell(\sigma) = \langle\sigma|\mathbf{Q}^\ell|\pi\rangle$, the total probability of all paths that end at σ in ℓ steps; $T_\ell(\sigma)$, the total average time of all such paths; and $\Gamma_\ell(\sigma)$, the total entropy of all such paths. These quantities obey the following recursion relations:

$$\begin{aligned}
P_\ell(\sigma') &= \sum_{\text{nn } \sigma \text{ of } \sigma'} \langle\sigma'|\mathbf{Q}|\sigma\rangle P_{\ell-1}(\sigma), \qquad (4.33)\\
T_\ell(\sigma') &= \sum_{\text{nn } \sigma \text{ of } \sigma'} \langle\sigma'|\mathbf{Q}|\sigma\rangle \left[T_{\ell-1}(\sigma) + \theta(\sigma)P_{\ell-1}(\sigma)\right],\\
\Gamma_\ell(\sigma') &= \sum_{\text{nn } \sigma \text{ of } \sigma'} \langle\sigma'|\mathbf{Q}|\sigma\rangle \left[\Gamma_{\ell-1}(\sigma) - \log\langle\sigma'|\mathbf{Q}|\sigma\rangle P_{\ell-1}(\sigma)\right],
\end{aligned}$$

where $P_0(\sigma) = \pi(\sigma)$ and $T_0(\sigma) = \Gamma_0(\sigma) = 0$ for all $\sigma \in \mathcal{S}$, and the sums run over all nearest neighbors (nn) σ of σ'. The final states $\sigma \in \mathcal{S}_f$ are treated as absorbing to ensure that only first-passage paths are counted. This procedure can be considered a generalization of the exact-enumeration algorithm of Ref. 216. Path ensemble averages are then given by

$$\mathcal{Z}_\Phi = \sum_{\ell=1}^{\infty} \sum_{\sigma \in \mathcal{S}_f} P_\ell(\sigma), \qquad \rho_\Phi(\ell) = \frac{1}{\mathcal{Z}_\Phi} \sum_{\sigma \in \mathcal{S}_f} P_\ell(\sigma), \tag{4.34}$$

$$\bar{t}_\Phi = \frac{1}{\mathcal{Z}_\Phi} \sum_{\ell=1}^{\infty} \sum_{\sigma \in \mathcal{S}_f} T_\ell(\sigma), \qquad S_\Phi = \frac{1}{\mathcal{Z}_\Phi} \sum_{\ell=1}^{\infty} \sum_{\sigma \in \mathcal{S}_f} \Gamma_\ell(\sigma).$$

We can similarly calculate state-dependent quantities such as $\langle \mathcal{I}_\sigma \rangle_\Phi$ and $\langle \mathcal{T}_\sigma \rangle_\Phi$. The two quantities to be recursively updated are $P_\ell(\sigma'; \sigma)$, the total probability of all paths currently at σ' at step ℓ that have visited σ at least once previously, and $T_\ell(\sigma'; \sigma)$, the total average time that all such paths have spent in σ. These obey the following recursion relations:

$$P_\ell(\sigma'; \sigma) = \begin{cases} \displaystyle\sum_{\text{nn } \sigma'' \text{ of } \sigma'} \langle \sigma' | \mathbf{Q} | \sigma'' \rangle P_{\ell-1}(\sigma''; \sigma), & \sigma' \neq \sigma, \\ \\ P_\ell(\sigma), & \sigma' = \sigma, \end{cases} \tag{4.35}$$

$$T_\ell(\sigma'; \sigma) = \sum_{\text{nn } \sigma'' \text{ of } \sigma'} \langle \sigma' | \mathbf{Q} | \sigma'' \rangle [T_{\ell-1}(\sigma''; \sigma) + \delta_{\sigma,\sigma''} \theta(\sigma'') P_{\ell-1}(\sigma''; \sigma)],$$

with the initial conditions $P_0(\sigma'; \sigma) = T_0(\sigma'; \sigma) = 0$ for all $\sigma, \sigma' \in \mathcal{S}$, $\sigma \neq \sigma'$ ($P_0(\sigma; \sigma) = \pi(\sigma)$, $T_0(\sigma; \sigma) = 0$). Averages are then expressed as

$$\langle \mathcal{I}_\sigma \rangle_\Phi = \frac{1}{\mathcal{Z}_\Phi} \sum_{\ell=1}^{\infty} \sum_{\sigma' \in \mathcal{S}_f} P_\ell(\sigma'; \sigma), \qquad \langle \mathcal{T}_\sigma \rangle_\Phi = \frac{1}{\mathcal{Z}_\Phi} \sum_{\ell=1}^{\infty} \sum_{\sigma' \in \mathcal{S}_f} T_\ell(\sigma'; \sigma). \tag{4.36}$$

Furthermore, we can calculate mean path divergence that characterizes the spatial diversity of the paths in Φ:

$$\mathcal{D}_\Phi = \sum_{\ell=1}^{\infty} \sum_{\sigma, \sigma' \in \mathcal{S}} d(\sigma, \sigma') P_\ell(\sigma) P_\ell(\sigma'), \tag{4.37}$$

where $d(\sigma, \sigma')$ is a distance metric on $\mathcal{S}$. This definition is distinct from that proposed in Refs. 35, 85 (Eq. 1.13) in that it dynamically calculates distances between points on paths as they propagate, rather than comparing the minimal distance between complete paths. Thus for a path that revisits some states multiple times, the divergence with a path that travels through the same set of states without revisiting any of them will be zero according to Eq. 1.13, but nonzero with the definition in Eq. 4.37.

This algorithm allows for very general definitions of the path ensemble Φ without having to explicitly enumerate all the paths. For instance, Φ can include paths that begin and end at arbitrary sets of states, or are prohibited from passing through arbitrary sets of intermediate states. The time complexity of the algorithm is $\mathcal{O}(\gamma N \Lambda)$ for $\mathcal{Z}_\Phi$, $\rho_\Phi(\ell)$, $\bar{t}_\Phi$, S_Φ, and $\mathcal{O}(\gamma N^2 \Lambda)$ for $\langle \mathcal{I}_\sigma \rangle_\Phi$, $\langle \mathcal{T}_\sigma \rangle_\Phi$, $\mathcal{D}_\Phi$, where γ is the average number of nearest neighbors, N is the number of states visited by paths in Φ, and Λ is the cutoff path length. The cutoff Λ scales with network size N in the same way as the average path length $\bar{\ell}_\Phi$; for simple random walks, it is known that

$$\Lambda \sim \bar{\ell}_\Phi \sim \begin{cases} N^{d_w/d_f}, & d_w \geq d_f \quad \text{(compact exploration)}, \\ N, & d_w < d_f \quad \text{(non-compact exploration)}, \end{cases} \tag{4.38}$$

where d_w is the dimension of the walk and d_f is the fractal dimension of the space [202, 217]. Therefore, the algorithm scales as

$$\mathcal{O}(\gamma N \Lambda) = \begin{cases} \mathcal{O}(\gamma N^{1+d_w/d_f}), & d_w \geq d_f, \\ \mathcal{O}(\gamma N^2), & d_w < d_f, \end{cases} \tag{4.39}$$

automatically accounting for the sparseness of network connections. This scaling compares favorably with standard linear algebra algorithms, which in general require $\mathcal{O}(N^3)$ operations [218] to solve the backward equation [199, 219].

4.4 Example: evolution on a neutral network

As a simple application of this approach, we consider a population evolving on a neutral network [34], as described in Sec. 1.3.4. In the space of all sequences of length L and with an alphabet of size k, we assign each sequence fitness 1 with probability p or fitness zero with probability $1-p$. The subset of fit states connected to each other forms a neutral network; there can be several disconnected neutral networks in each landscape realization. All jumps between neighboring fit states occur at the same rate u, and waiting times are Markovian. We choose $L=8$ and a binary alphabet $\{\mathsf{A},\mathsf{B}\}$ ($k=2$), which gives $2^8=256$ total nodes in the network, and we consider the ensemble Φ of first-passage paths from the sequence AAAAAAAA to the sequence BBBBBBBB.

Figure 4.1A shows $\rho_\Phi(\ell)$ for a single realization of this model with $p=0.9$. The exponential tail of $\rho_\Phi(\ell)$ is a universal feature of first-passage processes on finite spaces [217]; other path statistics, such as the average time $\bar{t}_\Phi(\ell)$ of paths with length ℓ, also show asymptotic behavior that is exponential for long paths. We can use this feature to determine the cutoff path length Λ for the algorithm: Λ is set at a length such that $\rho_\Phi(\ell)$ and $\bar{t}_\Phi(\ell)$ are close to exponential in a region around Λ. Then one need only consider paths with $\ell<\Lambda$ and infer the contributions of all longer paths from an exponential fit to the tail, which considerably improves the efficiency of the algorithm. This procedure takes advantage of the fact that information about longer paths is already contained in the structure of shorter paths; the longer paths are built on the shorter paths by adding loops. The maximum length Λ of the shorter paths that must be explicitly calculated depends on the chemical distance between the initial and final states and the lengths over which the landscape is correlated. This essentially implements a numerical renormalization scheme on the ensemble of paths [207].

In Fig. 4.1B,C,D we show distributions of the mean path time $\bar{t}_\Phi$, mean path length $\bar{\ell}_\Phi$, path length standard deviation ℓ_Φ^{sd}, and path entropy S_Φ for multiple realizations of the neutral network with high and low values of p. We see that long paths are likely in

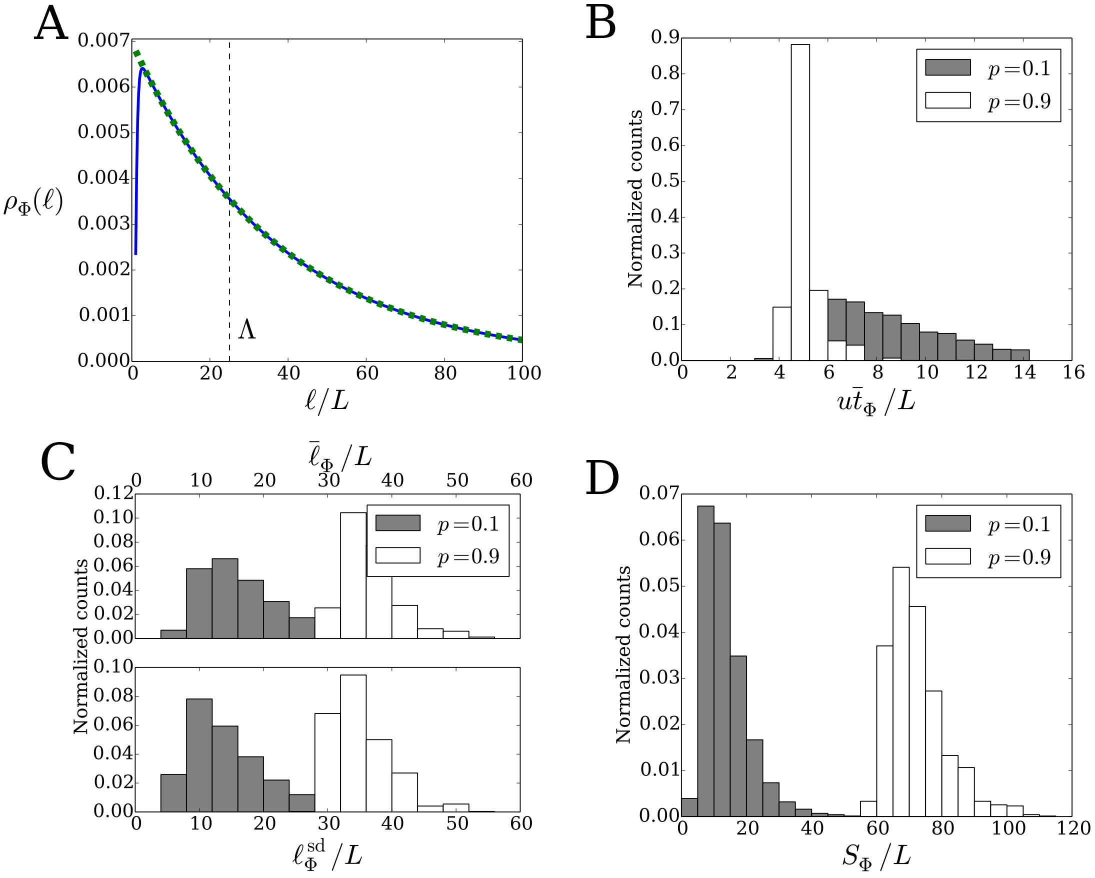


Figure 4.1: **First-passage path ensemble statistics in a neutral network.** (A) The path length distribution $\rho_\Phi(\ell)$ (solid, blue) and exponential fit (dashed, green) in the interval $[\Lambda - 5, \Lambda]$ for $\Lambda = 25$ in a single realization of the neutral network with $p = 0.9$. (B) Distribution of mean path times $\bar{t}_\Phi$, (C) distribution of mean path lengths $\bar{\ell}_\Phi$ and standard deviations of path lengths $\ell_\Phi^{\rm sd}$, and (D) distribution of path entropies S_Φ for $p = 0.1$ and $p = 0.9$. Histograms in (B)–(D) are generated from 10^4 successful random realizations of the neutral network for each value of p; a realization is considered successful if both initial and final states are included in a single connected network.

these models: dozens of substitutions can occur at each site before the final state is reached. The larger size of the neutral network for $p = 0.9$ allows longer paths on average than for $p = 0.1$. However, the mean time of paths for the larger neutral network is usually smaller (Fig. 4.1B), since the increased connectivity of the network leads to shorter waiting times at individual nodes. Larger p leads to substantially more diversity of paths and path lengths, as expected due to the increased size and connectivity of the network (Fig. 4.1C,D). Note that the distributions of $\bar{\ell}_\Phi$ and ℓ_Φ^{sd} in Fig. 4.1C are nearly the same, owing to the nearly exponential distribution of $\rho_\Phi(\ell)$ in this model (cf. Fig. 4.1A).

In an unconstrained sequence space, the number of nearest neighbors is $\gamma = L(k-1)$, and the average path length $\bar{\ell}_\Phi$ scales as $N = k^L$ (Eq. 4.38). According to Eq. 4.19, the entropy of paths in sequence space is

$$S_\Phi = \bar{\ell}_\Phi \log L(k-1) \sim k^L \log L(k-1). \tag{4.40}$$

When $p = 0.9$ the neutral network is nearly the size of the entire sequence space, and these results hold approximately. Indeed, we see that $\bar{\ell}_\Phi$ and S_Φ differ by roughly a factor of $\log L(k-1) \approx 2.1$ (Fig. 4.1C,D).

4.5 Applications to reaction rate theory

Besides molecular evolution, a particularly important application of the path-based approach is **reaction rate theory** [220], which studies rare transitions between metastable states that model phenomena ranging from protein folding [221] to chemical reactions [200]. In these systems, quantities of interest include not only mean first-passage times and reaction rates but also the spatial distribution of transition paths and identification of kinetic bottlenecks.

So-called transition state theory is a well-known approach to these problems; however, it relies on the existence and *a priori* identification of key transition states [220]. A more

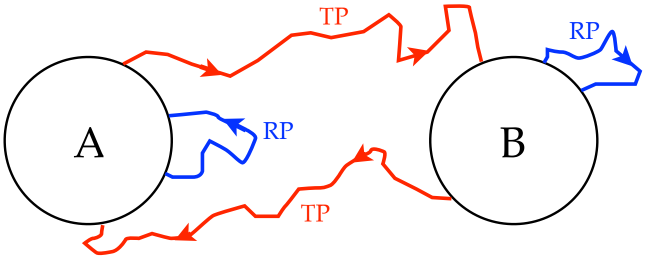


Figure 4.2: **Schematic of reactions between metastable states.** The regions A and B are metastable states. When the system leaves one of these states, it can either follow a transition path (TP, red) and reach the other metastable state, or it can follow a return path (RP, blue) and return to the state in which it started. The general aim of reaction rate theory is to study the statistical properties of the transition path ensemble, especially their overall rate.

recent development has been transition path sampling [200, 210, 222–225], in which paths are directly sampled via Monte Carlo to estimate their statistical properties. Similar methods have been used in phylogenetic analysis of protein sequences [183, 226–229]. These techniques are based on a finite sample of paths and do not provide natural cutoffs for the size of the sample, which may lead to noisy estimates of various path statistics. Another technique, called transition path theory [199, 219, 230–232], relies on explicit solutions to the backward equation. This approach, though more systematic, does not directly address the diversity of paths.

We now illustrate our approach on two simple reaction rate problems. In a typical reaction rate problem, a system has two or more metastable states, usually defined such that the system in steady state spends almost all of its time in those states. Transitions between metastable states are therefore rare. Figure 4.2 shows a schematic of a system with two metastable states. When the system exits one of the metastable states, there are two possible outcomes: the system can either return to the state in which it started, without having reached any other metastable state, or it can transition over to another metastable state. The ensemble of first-passage paths leaving the boundary of one metastable state can thus be partitioned into two sub-ensembles, one called “return paths” (RP) and another called “transition paths” (TP).

Assume there are just two metastable states A and B; it is straightforward to generalize these results for additional metastable states. Let TP denote the ensemble of transition paths from A to B and vice versa, while RP denotes the ensemble of return paths for both A and B [223, 224]. We can use the foregoing formalism and numerical algorithm to calculate statistical properties of these path ensembles. In particular, define the density of states on TPs as the probability of observing the system at a state σ, given it is on a TP:

$$p(\sigma|\mathrm{TP}) = \frac{1}{\bar{t}_{\mathrm{TP}}} \langle \mathcal{T}_\sigma \rangle_{\mathrm{TP}}, \tag{4.41}$$

where we calculate the probability as the fraction of total time spent at σ. A related quantity is the TP density, defined as the probability of being on a TP given the system is observed at σ:

$$p(\mathrm{TP}|\sigma) = \frac{\mathcal{Z}_{\mathrm{TP}} \langle \mathcal{I}_\sigma \rangle_{\mathrm{TP}}}{\mathcal{Z}_{\mathrm{TP}} \langle \mathcal{I}_\sigma \rangle_{\mathrm{TP}} + \mathcal{Z}_{\mathrm{RP}} \langle \mathcal{I}_\sigma \rangle_{\mathrm{RP}}}. \tag{4.42}$$

These are related via Bayes rule [224]:

$$p(\mathrm{TP}|\sigma) = \frac{p(\sigma|\mathrm{TP}) p(\mathrm{TP})}{\pi(\sigma)}, \tag{4.43}$$

where $\pi(\sigma)$ is the steady-state probability of the system being at σ and $p(\mathrm{TP})$ is the probability of being on a TP at any state. We will generally assume the dynamics of the system are described by a potential energy function $V(\sigma)$, and that the system is in thermal equilibrium with $\pi(\sigma) = e^{-\beta V(\sigma)}/Z$, although more general cases are possible. The probability $p(\mathrm{TP})$ can be calculated as

$$\begin{aligned} p(\mathrm{TP}) &= p(\mathrm{TP}|\text{system is in transition region})\; p(\text{system is in transition region}) \\ &= \frac{\mathcal{Z}_{\mathrm{TP}} \bar{t}_{\mathrm{TP}}}{\mathcal{Z}_{\mathrm{TP}} \bar{t}_{\mathrm{TP}} + \mathcal{Z}_{\mathrm{RP}} \bar{t}_{\mathrm{RP}}} (1 - \pi_A - \pi_B), \end{aligned} \tag{4.44}$$

where π_A and π_B are the equilibrium probabilities of being in metastable states A and B. The probability of being in the transition region $(1-\pi_A-\pi_B)$ is very small by construction.

We approximate the overall flux of TPs as the probability of being on a TP divided by the average time of a TP [224]:

$$\lambda \approx \frac{p(\mathrm{TP})}{\bar{t}_{\mathrm{TP}}} = \frac{(1-\pi_A-\pi_B)\mathcal{Z}_{\mathrm{TP}}}{\mathcal{Z}_{\mathrm{TP}}\bar{t}_{\mathrm{TP}}+\mathcal{Z}_{\mathrm{RP}}\bar{t}_{\mathrm{RP}}}. \tag{4.45}$$

The reaction rates are then given by $k_{A\to B} = \lambda/(2\pi_A)$ and $k_{B\to A} = \lambda/(2\pi_B)$.

4.5.1 Double-well potential in two dimensions

As a simple example we consider a random walk on a two-dimensional square lattice $\mathcal{S} = [-1.6, 1.6] \times [-1.3, 1.3]$ with spacing $\Delta x = 0.05$. Let

$$V(x,y) = \frac{1}{6}(4(1-x^2-y^2)^2 + 2(x^2-2)^2 + ((x+y)^2-1)^2 + ((x-y)^2-1)^2 - 2) \tag{4.46}$$

be the potential energy landscape over this space (Fig. 4.3A). The two metastable states (outlined in Fig. 4.3A) are $A = [-1.5, -0.5] \times [-0.5, 0.5]$ and $B = [0.5, 1.5] \times [-0.5, 0.5]$. We consider Monte Carlo jump rates between nearest neighbors on the lattice:

$$\langle x', y'|\mathbf{W}|x, y\rangle = (\Delta x)^{-2}\min[1, e^{-\beta(V(x',y')-V(x,y))}]. \tag{4.47}$$

The rates are rescaled by $(\Delta x)^{-2}$ so that Brownian dynamics with a fixed diffusion constant is recovered in the $\Delta x \to 0$ limit.

Figure 4.3B shows the equilibrium distribution $\pi(x,y)$, the density of states on TPs $p(x,y|\mathrm{TP})$, and the TP density $p(\mathrm{TP}|x,y)$ for several inverse temperatures β. The density of states on transition paths $p(x,y|\mathrm{TP})$ shows two symmetric channels by which most reactions between A and B occur. We also show the distribution of TP lengths $\rho_{\mathrm{TP}}(\ell)$ in Fig. 4.4A,

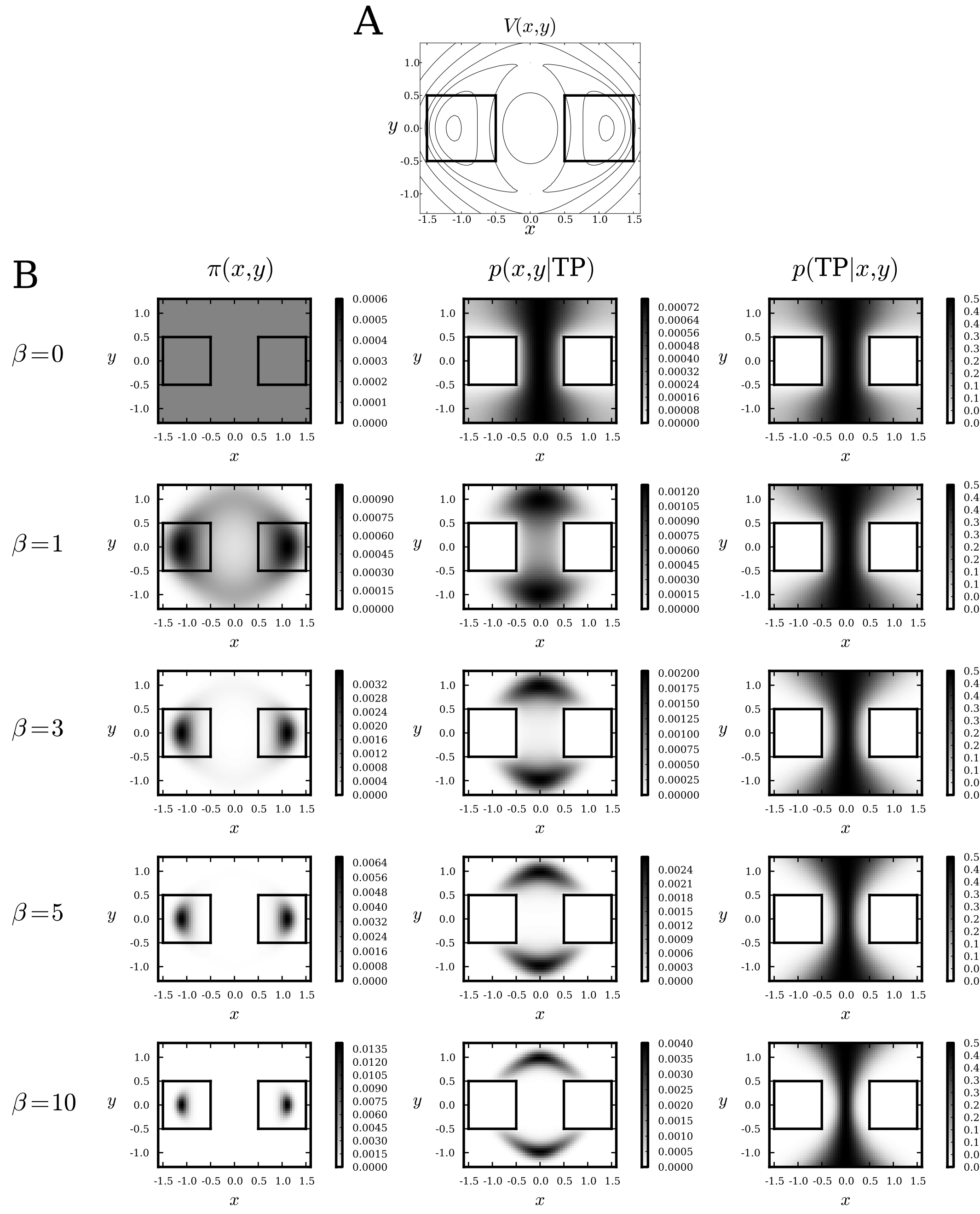


Figure 4.3: **Reactions on a two-dimensional double-well potential.** (A) Contours of the potential energy function (Eq. 4.46). The metastable states A and B are shown as boxes around the energy minima. (B) The equilibrium distribution of states $\pi(x,y)$, density of states on TPs $p(x,y|\text{TP})$, and TP densities $p(\text{TP}|x,y)$ at different values of inverse temperature β. The lattice spacing is fixed at $\Delta x = 0.05$.

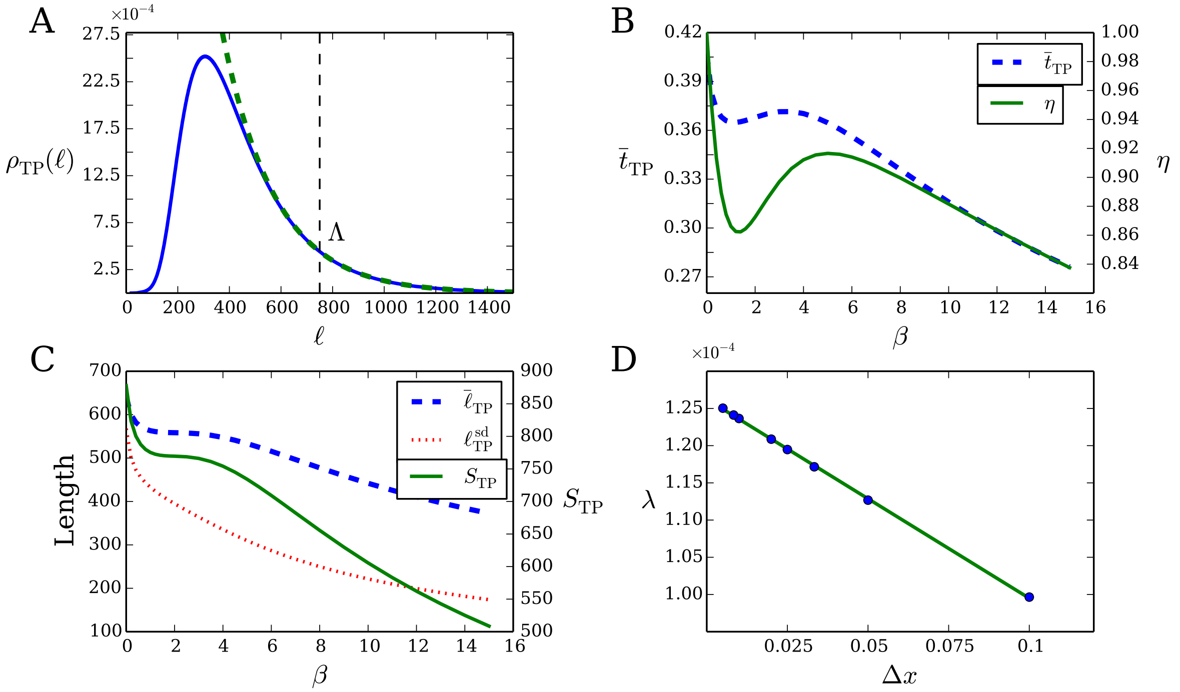


Figure 4.4: **Statistics of transition paths for the double-well potential.** (A) For the ensemble of TPs in the 2D double-well potential, path length distribution $\rho_{\mathrm{TP}}(\ell)$ (solid, blue) versus path length ℓ, and exponential fit in the interval $[\Lambda - 50, \Lambda]$ (dashed, green) are shown. (B) Relative mean path divergence $\eta = (\mathcal{D}_{\mathrm{TP+RP}}(\beta)/\mathcal{D}_{\mathrm{TP+RP}}(\beta = 0))^{1/2}$ (solid, green) and average time of TPs $\bar{t}_{\mathrm{TP}}$ (dashed, blue) versus β. The divergence η is calculated using Eq. 4.37 with $d(x, y; x', y') = (x - x')^2 + (y - y')^2$. (C) Mean length $\bar{\ell}_{\mathrm{TP}}$ (dashed, blue), standard deviation $\ell^{\mathrm{sd}}_{\mathrm{TP}}$ (dotted, red), and entropy S_{TP} of TPs (solid, green) versus β. (D) Total flux λ of TPs as a function of lattice spacing Δx. The green line is the best fit straight line, which we use to infer the continuous limit flux of $\lambda_0 \approx 1.3 \times 10^{-4}$. In (A)–(C), the lattice spacing is $\Delta x = 0.05$.

which peaks at an intermediate path length but decays exponentially for long paths as previously argued.

In general, we expect paths to be longer and more diverse at higher temperatures. However, between $\beta = 5$ and $\beta = 1$ the paths become shorter and less diverse as $T = 1/\beta$ increases (Fig. 4.4B, C). This is a signature of entropic switching [231]: at a critical value of T, the two most energetically-favored pathways that dominate the low-T behavior become less favorable than the shorter path through the middle. Entropic switching is reflected in plots of the relative path divergence, $\bar{t}_{\mathrm{TP}}$, $\bar{\ell}_{\mathrm{TP}}$, and S_{TP} (Fig. 4.4B, C).

We can also calculate the continuous-space limit of the TP flux λ and the reaction rates. We analytically continue λ as a function of the lattice spacing Δx:

$$\lambda(\Delta x) = \lambda_0 + \lambda_1 \Delta x + \mathcal{O}(\Delta x^2), \tag{4.48}$$

where λ_0 is the continuous-limit flux and Δx should be smaller then the smallest length scale of the potential. Indeed, $\lambda(\Delta x)$ is linear (Fig. 4.4D), yielding continuous-limit rates of $k_{A\to B} = k_{B\to A} \approx 1.3 \times 10^{-4}$. Therefore, one need only calculate λ at a few finite lattice spacings in order to infer continuous-limit rates.

4.5.2 Triple-well potential on a fractal

As a more complex example, we consider reactions on a fractal in a triple-well potential, which may serve as a model of transport in disordered media [201]. We embed the Sierpinski triangle of side length 1 in a two-dimensional triple-well potential:

$$V(x,y) = 10\sum_{i=1}^{3}((x-x_i)^2 + (y-y_i)^2)e^{-5(x-x_i)^2-5(y-y_i)^2}, \tag{4.49}$$

where $(x_1, y_1) = (0, 1/\sqrt{3})$, $(x_2, y_2) = (1/2, -1/(2\sqrt{3}))$, and $(x_3, y_3) = (-1/2, -1/(2\sqrt{3}))$ are the corners of the triangle. The metastable states are defined around these corners and are shown in Fig. 4.5B. We consider Monte Carlo jump rates as before (Eq. 4.47) but with rates rescaled by $(\Delta x)^{-d_w}$, where $\Delta x = 2^{-n}$ (n is the fractal order) is the spacing between neighboring points on the triangle and $d_w = \log 5/\log 2$ is the dimension of a random walk on the Sierpinski triangle [201].

Figure 4.5B shows that TPs bottleneck in the middle of the three sides of the triangle. As before, we use analytical continuation to infer the continuous-limit reaction rate k between any pair of metastable states from finite-order realizations of the fractal, yielding $k \approx 2.0 \times 10^{-2}$ (Fig. 4.5A).

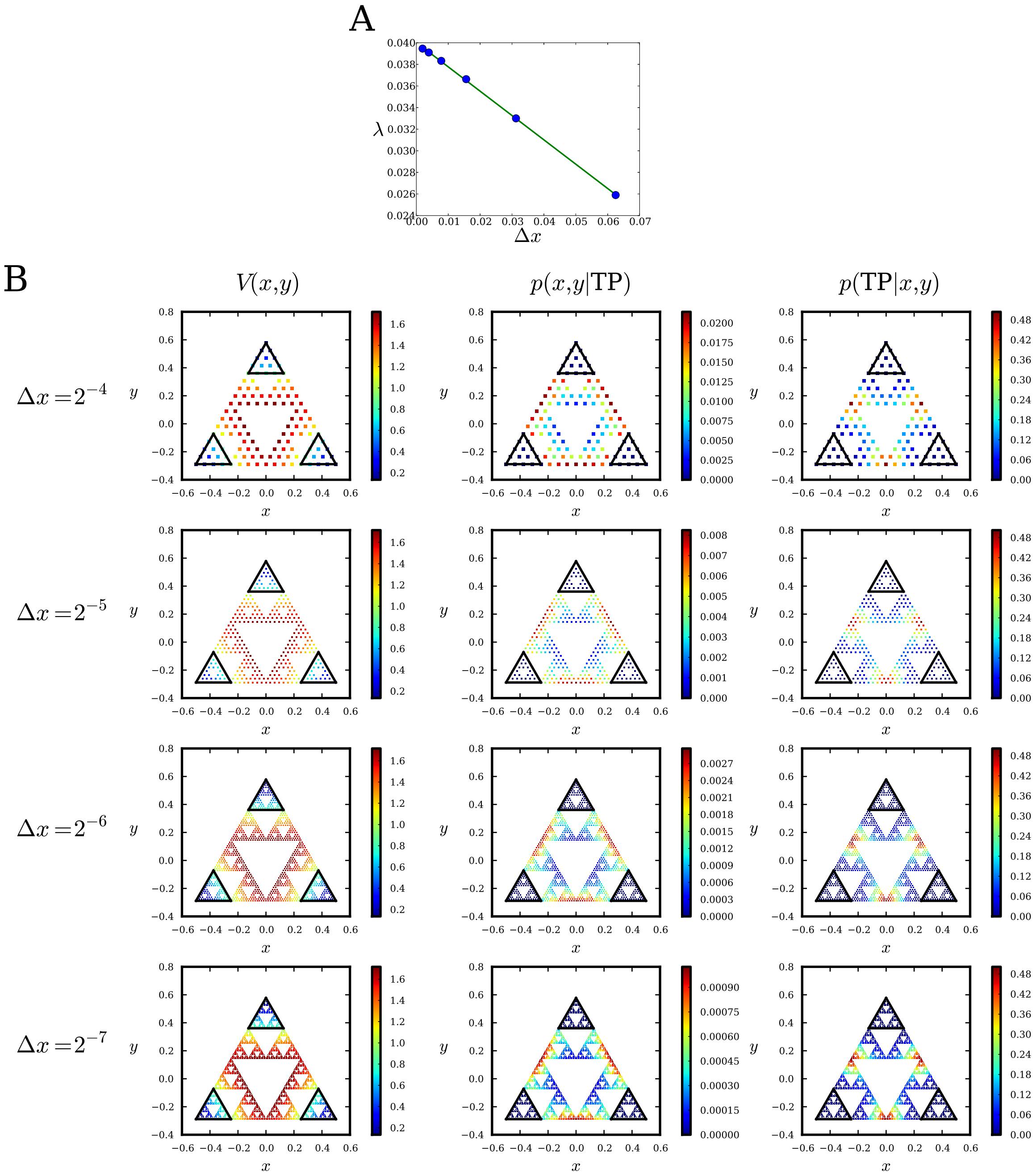


Figure 4.5: **Reactions on a Sierpinski triangle embedded in a triple-well potential.** (A) Transition path flux λ as a function of lattice spacing Δx. As with the double-well potential, analytic continuation of $\lambda(\Delta x)$ allows us to infer the reaction rate $k \approx \lambda/2 \approx 2.0 \times 10^{-2}$ between any pair of metastable states in an infinite-order fractal using a few finite-order realizations. (B) The potential $V(x, y)$, the density of states on TPs $p(x, y|\text{TP})$, and TP densities $p(\text{TP}|x, y)$ for $\beta = 6$.

Chapter 5

Evolution of Protein Binding and Folding Stability

In the previous chapter we developed a general methodology for studying the ensemble of paths in a stochastic process. We now apply this approach to a model of protein evolution. This chapter reproduces Ref. 12. Appendix D contains additional material from Refs. 10, 11 on a related model.

Proteins carry out a diverse array of chemical and mechanical functions in the cell, ranging from metabolism to signaling [233]. Therefore proteins serve as central targets for natural selection in wild populations, as well as a key toolbox for bioengineers to design novel molecules with medical and industrial applications [40, 234–237]. For many proteins, structure is essential for their function [233]: the protein must fold into its native state, a unique three-dimensional conformation, in order to perform its function, which typically involves binding a target molecule such as DNA, RNA, another protein, or a small ligand. Misfolded proteins may also form toxic aggregates and divert valuable protein synthesis and quality control resources [238–241]. It is therefore imperative that the folded state be stable against the thermal fluctuations present at physiological temperatures. However, biophysical experiments and computational studies reveal that most random mutations in proteins destabilize the folded state [242, 243], including mutations that improve function [243–245]. As a result many natural proteins tend to be only marginally stable, mutationally teetering at the brink of substantial unfolding [59, 246]. With proteins in such a precarious evolutionary position, how can they evolve new functions while maintaining sufficient folding stability?

Directed evolution experiments have offered a window into the dynamics of this process [40, 234–237], indicating the importance of compensatory mutations, limited epistasis, and mutational robustness. Theoretical efforts to describe protein evolution in biophysical terms have focused on evolvability [54, 58], global properties of protein interaction networks [66, 247], and reproducing observed distributions of protein stabilities and evolutionary rates [55, 57, 59, 61, 62]. However, a subtle but key property of proteins has not been explored in this context: the structural coupling of folding and binding (the fact that folding is required for function) implies an evolutionary coupling of folding stability and binding strength. This raises the possibility that selection acting directly on only one of these traits may produce apparent, indirect selection for the other. The importance of such indirect selection for coupled traits was popularized by Gould and Lewontin in their influential paper on evolutionary "spandrels" [248], defined as traits that evolve in the absence of direct selection. In particular this includes traits that emerge as byproducts (via indirect selection) when there is direct selection on another property coupled to the spandrel. Since then the importance of coupling between traits has been explored in many areas of evolutionary biology [249], including various molecular examples [246, 250–253].

How do coupled traits affect protein evolution? We consider a simple biophysical and evolutionary model that describes evolution of a new binding interaction in the context of a directed evolution experiment [40], as a result of gene duplication and divergence [254], or in response to a change in the protein's chemical or physical environment, including availability and concentrations of various ligands [31, 37] and temperature [103, 255]. We postulate a fitness landscape as a function of two quantitative protein traits: the free energy of folding (stability) and the free energy of binding a target molecule. This fitness function allows us to parameterize the distinct selection pressures acting on folding and binding. We then use an exact numerical algorithm [10, 11] to quantitatively characterize the fitness landscape and the resulting adaptive paths, addressing key evolutionary questions of epistasis [24, 32, 256], repeatability [14, 31, 85], and the tempo and rhythm of adaptation [15, 55].

We find that both binding and folding can readily emerge as evolutionary spandrels: they evolve as protein traits even in the absence of direct selective advantage. In particular, proteins can evolve strong binding interactions that confer no intrinsic fitness advantage but merely serve to stabilize the protein if misfolding is deleterious. The evolution of these nonfunctional interactions may be highly stochastic: random mutation events can determine whether or not a protein evolves to bind a ligand. This offers a compelling interpretation of widespread nonfunctional interactions observed among proteins genome-wide [66, 247]. Moreover, when there are distinct selection pressures on both folding and binding, we predict strongly-constrained adaptive paths that gain extra stability first and then partially lose it before new function is acquired. This suggests the evolution of many natural proteins is highly predictable at the level of folding and binding energy traits.

5.1 Biophysical model of evolution

5.1.1 Protein energetics

We consider a protein with two-state folding kinetics [233]. The protein has an interface that binds a ligand, such as a small molecule, DNA, RNA, or another protein. If the protein is folded, it may be bound or unbound, but it cannot form a binding interface in the unfolded state. Because the protein can bind *only* when it is folded, the binding and folding processes are structurally coupled (see Appendix D for an investigation of the model when folding and binding are *not* coupled). Under the thermodynamic equilibrium assumption (valid when protein folding and binding are faster than typical cellular processes), the probabilities of the three structural states — folded and bound ($p_{\mathrm{f,b}}$), folded and unbound ($p_{\mathrm{f,ub}}$), and unfolded and unbound ($p_{\mathrm{uf,ub}}$) — are given by their corresponding Boltzmann weights:

State	Free energy	Probability
folded, bound	$E_f + E_b$	$p_{\text{f,b}} = \frac{e^{-\beta(E_f+E_b)}}{1+e^{-\beta E_f}+e^{-\beta(E_f+E_b)}}$
folded, unbound	E_f	$p_{\text{f,ub}} = \frac{e^{-\beta E_f}}{1+e^{-\beta E_f}+e^{-\beta(E_f+E_b)}}$
unfolded, unbound	0	$p_{\text{uf,ub}} = \frac{1}{1+e^{-\beta E_f}+e^{-\beta(E_f+E_b)}}$

(5.1)

Here β is inverse temperature, E_f is the free energy of folding (also known as ΔG), and $E_b = E_b' - \mu$, where E_b' is the binding free energy and μ is the chemical potential of the target molecule. For simplicity, we will refer to E_b as the binding energy, unless indicated otherwise. Note that $E_f < 0$ for intrinsically-stable proteins and $E_b < 0$ for favorable binding interactions.

The folding and binding energies depend on the protein's genotype (amino acid sequence) σ. We assume adaptation only affects "hotspot" residues at the binding interface [257, 258]; the rest of the protein does not change on relevant time scales because it is assumed to be already optimized for folding. If positions away from the binding interface can accept stabilizing mutations (and are not functionally constrained), they may be explicitly included into the model as "folding hotspots." In the present study we focus on L binding hotspot residues which, to a first approximation, make additive contributions to the total folding and binding free energies [259, 260]:

$$E_f(\sigma) = E_f^{\text{ref}} + \sum_{i=1}^{L} \epsilon_f(i, \sigma^i), \quad E_b(\sigma) = E_b^{\text{min}} + \sum_{i=1}^{L} \epsilon_b(i, \sigma^i), \tag{5.2}$$

where $\epsilon_f(i, \sigma^i)$ and $\epsilon_b(i, \sigma^i)$ capture the energetic contributions of amino acid σ^i at position i. The reference energy E_f^{ref} is the fixed contribution to the folding energy from all other residues in the protein. Furthermore, by construction it is also the total folding energy of a reference sequence σ_{ref}, so that $\epsilon_f(i, \sigma^i)$ can be interpreted as the $\Delta\Delta G$ value (change in folding free energy E_f) resulting from a single point mutation away from σ_{ref}. The parameter E_b^{min} is equal to the minimum binding energy among all genotypes.

Folding energetics are probed experimentally and computationally by measuring $\Delta\Delta G$ values, which are the changes in E_f (also known as ΔG) resulting from single point mutations. Values of $\Delta\Delta G$ are observed to be universally distributed over many proteins [242]; consistent with this observation, we sample entries of ϵ_f from a Gaussian distribution with mean 1.25 kcal/mol and standard deviation 1.6 kcal/mol. For the reference sequence σ_{ref}, $\epsilon_f(i, \sigma^i_{\text{ref}}) = 0$ for all $i \in \{1, \ldots, L\}$, such that $E_f(\sigma_{\text{ref}}) = E_f^{\text{ref}}$. The parameter E_b^{min} is defined as the binding energy of the genotype σ_{bb} with the lowest E_b: $\epsilon_b(i, \sigma^i_{\text{bb}}) = 0$ for all $i \in \{1, \ldots, L\}$. Since binding hotspot residues typically have a 1–3 kcal/mol penalty for mutations away from the wild-type amino acid (this requirement is used to define which residues make up the hotspot) [257, 258], we sample the other entries of ϵ_b from an exponential distribution defined in the range of $(1, \infty)$ kcal/mol, with mean 2 kcal/mol. This distribution is consistent with alanine-scanning experiments which probe energetics of amino acids at the binding interface [261]. The exact shape of these distributions, however, is unimportant for large enough L due to the central limit theorem. We consider $L = 6$ hotspot residues and a reduced alphabet of $k = 5$ amino acids (grouped into negative, positive, polar, hydrophobic, and other), resulting in $5^6 = 15625$ possible genotypes.

5.1.2 Fitness landscape

We construct a simple fitness landscape based on the molecular traits E_f and E_b. Without loss of generality, we assume that the protein contributes fitness 1 to the organism if it is always folded and bound. Let $f_{\text{ub}}, f_{\text{uf}} \in [0, 1]$ be the multiplicative fitness penalties for being unbound and unfolded, respectively: the fitness is f_{ub} if the protein is unbound but folded, and $f_{\text{ub}} f_{\text{uf}}$ if the protein is both unbound and unfolded. Then the fitness of the protein averaged over all three possible structural states in Eq. 5.1 is given by

$$\mathcal{F}(E_f, E_b) = p_{\mathrm{f,b}} + f_{\mathrm{ub}} p_{\mathrm{f,ub}} + f_{\mathrm{ub}} f_{\mathrm{uf}} p_{\mathrm{uf,ub}}. \tag{5.3}$$

This fitness landscape is divided into three nearly-flat plateaus corresponding to the three protein states of Eq. 5.1, separated by steep thresholds corresponding to the folding and binding transitions (Fig. 5.1A). The heights of the plateaus are determined by the values of f_{ub} and f_{uf}, leading to three qualitative regimes of the global landscape structure (Fig. 5.1B–D).

In the first case (Fig. 5.1B), a protein that is perfectly folded but unbound has no fitness advantage over an unbound and unfolded protein: $f_{\mathrm{ub}} = f_{\mathrm{ub}} f_{\mathrm{uf}}$. Thus we say that selection only acts directly on the binding trait. This regime is realized when either $f_{\mathrm{ub}} = 0$ (binding is essential, e.g., in the context of conferring antibiotic resistance to the cell [31, 262]) or $f_{\mathrm{uf}} = 1$ (misfolded proteins are not toxic). In contrast, when $f_{\mathrm{ub}} = 1$ and $0 \leq f_{\mathrm{uf}} < 1$ (Fig. 5.1C), a perfectly folded and bound protein has no fitness advantage over a folded but unbound protein, and thus this case entails direct selection only for folding. These proteins are harmful to the cell in the misfolded state (e.g., due to aggregation or significant costs of degrading unfolded proteins [238–241]), while binding provides no intrinsic fitness advantage (the protein may have other, functional binding interfaces). Finally, it is also possible that there are distinct selection pressures on both binding and folding. This occurs when $0 < f_{\mathrm{ub}} < 1$ and $0 \leq f_{\mathrm{uf}} < 1$ (Fig. 5.1D).

It is straightforward to generalize our three-state model to proteins with additional structural states (other local minima on the folding energy landscape, other binding modes) and allow for simultaneous adaptation at multiple binding interfaces. Furthermore, the fitness landscape in Eq. 5.3 can be made an arbitrary nonlinear function of state probabilities. However, these more complex scenarios would still share the essential features of our basic model: coupling between folding and binding traits and sharp fitness thresholds between

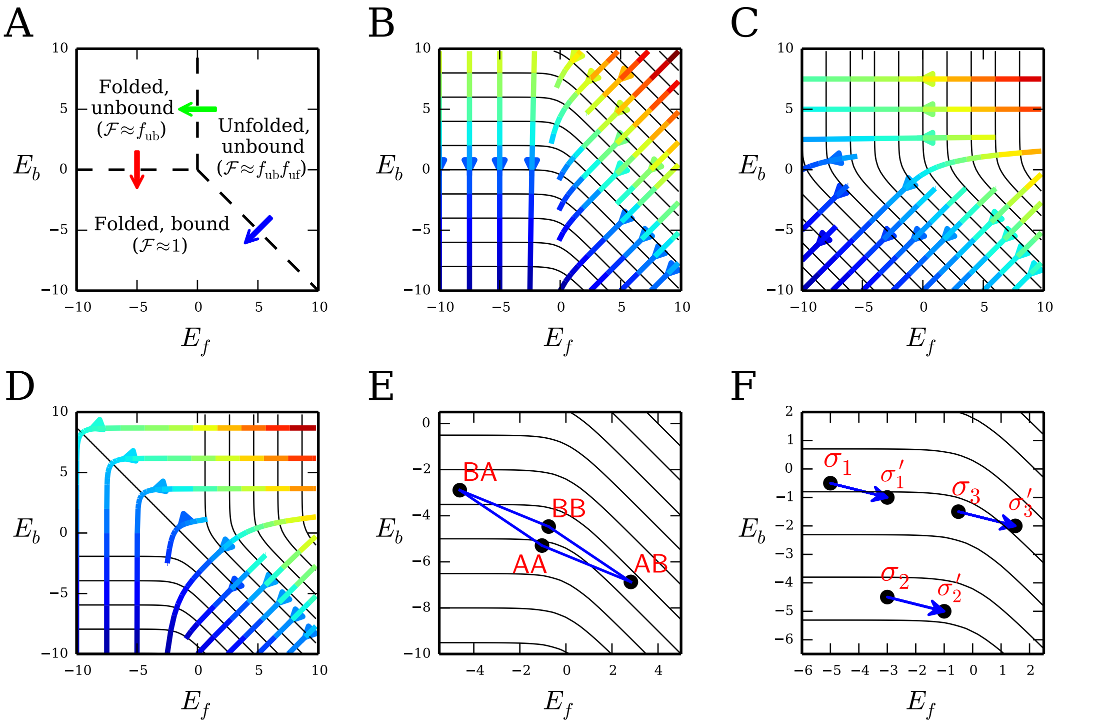


Figure 5.1: **Fitness, selection, and epistasis in energy trait space.** (A) Phase diagram of protein structure (Eq. 5.1) and fitness (Eq. 5.3). Dashed lines separate structural phases of the protein corresponding to plateaus on the fitness landscape; arrows represent the folding transition (green), binding transition (red), and the coupled folding-binding transition (blue). Fitness landscapes $\mathcal{F}(E_f, E_b)$ with direct selection (B) for binding only ($f_{\text{ub}} = f_{\text{uf}} = 0$), (C) for folding only ($f_{\text{ub}} = 1$, $f_{\text{uf}} = 0$), and (D) for both binding and folding ($f_{\text{ub}} = 0.9$, $f_{\text{uf}} = 0$). Black contours indicate constant fitness values. The contours are uniformly placed in energy space; fitness differences between adjacent contours are not all equal. Streamlines indicate the direction of the selection "force" $\vec{\nabla} \log \mathcal{F}$, with color showing its magnitude (decreasing from red to blue). (E) Example projection of a genotype distribution and mutational network into energy space for $L = 2$ and a two-letter ($k = 2$) alphabet. (F) Blue arrows indicate the same mutation on different genetic backgrounds. When the fitness contours are straight, the mutation is beneficial regardless of the background (σ_1 or σ_2). However, with curved contours, the same mutation can become deleterious ($\sigma_3 \to \sigma_3'$), indicative of sign epistasis. Sign epistasis from curved contours can give rise to multiple local fitness maxima (e.g., AA and BB in (E)).

bound/unbound and folded/unfolded states. Thus our qualitative conclusions do not depend on the specific model in Eq. 5.3.

5.1.3 Epistasis and local maxima

For protein sequences of length L with an alphabet of size k, each of the k^L possible genotypes is projected into the two-dimensional trait space of E_f and E_b (Eq. 5.2) and connected to $L(k-1)$ immediate mutational neighbors, forming a network of states that the population must traverse (a simple example is shown in Fig. 5.1E). Adaptive dynamics are determined by the interplay between the structure of the fitness landscape in the energy trait space (Fig. 5.1A–D) and the distribution of genotypes in trait space.

This interplay gives rise to the possibility of epistasis and multiple local fitness maxima. Our model is non-epistatic at the level of the energy traits, since residues make additive contributions to the total energies E_f and E_b (Eq. 5.2). Thus, mutations can be represented as vectors in energy space, resulting in the same displacement in energies regardless of the genetic background on which they occur (Fig. 5.1F). When the fitness contours are straight parallel lines, there can be no sign epistasis on the fitness landscape: a mutation that is beneficial on one background will be beneficial on all backgrounds. Magnitude epistasis, on the other hand, is widespread due to the nonlinear dependence of fitness on folding and binding energies. Curved fitness contours, which occur near the folding or binding thresholds in our model (Fig. 5.1B–D), can produce sign epistasis in fitness, giving rise to multiple local fitness maxima in the genotype space (Fig. 5.1E).

5.1.4 Evolutionary dynamics

We assume a population encoding the protein of interest evolves in the monomorphic limit (defined in Sec. 1.4.1) as in previous chapters. We moreover consider the strong-selection limit of $N|s| \gg 1$. In this case the Wright-Fisher fixation probability (Eq. 1.11) can be approximated as

$$\phi(s) \approx \begin{cases} 1 - e^{-2s} & \text{for } s > 0, \\ 0 & \text{for } s < 0. \end{cases} \tag{5.4}$$

The substitution rate is Eq. 1.5 with the above fixation probability. Thus the effective population size N sets the overall time scale $(Nu)^{-1}$ of substitutions but does not affect fixation probabilities. In this regime, deleterious mutations never fix and adaptive paths have a finite number of steps, terminating at a global or local fitness maximum. For compact genomic units such as proteins, the monomorphic condition is generally met in multicellular species, although it may be violated in some unicellular eukaryotes and prokaryotes [263]. Sequential fixation of single mutants is also a typical mode of adaptation in directed evolution experiments [40]. For simplicity, we neglect more complex mutational moves such as indels and recombination.

5.1.5 Validity of the strong-selection approximation

Far from the binding and folding thresholds the fitness landscape becomes flat (Fig. 5.1A) and the strong-selection assumption of Eq. 5.4 may be violated. To establish the limits of validity for our model, we calculate average selection coefficients of accessible substitutions (defined as $s = \mathcal{F}_{\text{final}}/\mathcal{F}_{\text{initial}} - 1$, where $\mathcal{F}_{\text{initial}}$ and $\mathcal{F}_{\text{final}}$ are the initial and final fitness values of a substitution), both throughout the landscape and at the local maxima (Fig. 5.2). The calculations were done for all three regimes of the model described above (Fig. 5.1B–D) and for a wide range of folding and binding energies. We observe that for typical values of the effective population size $N \in (10^4, 10^7)$ [83, 263], the selection strengths in the model justify our strong-selection approximation for realistic choices of energy parameters.

5.1.6 Quantitative description of adaptation

Although our model accommodates a general evolutionary process with any initial condition for the protein, for concreteness we focus on a specific but widely-applicable scenario. A

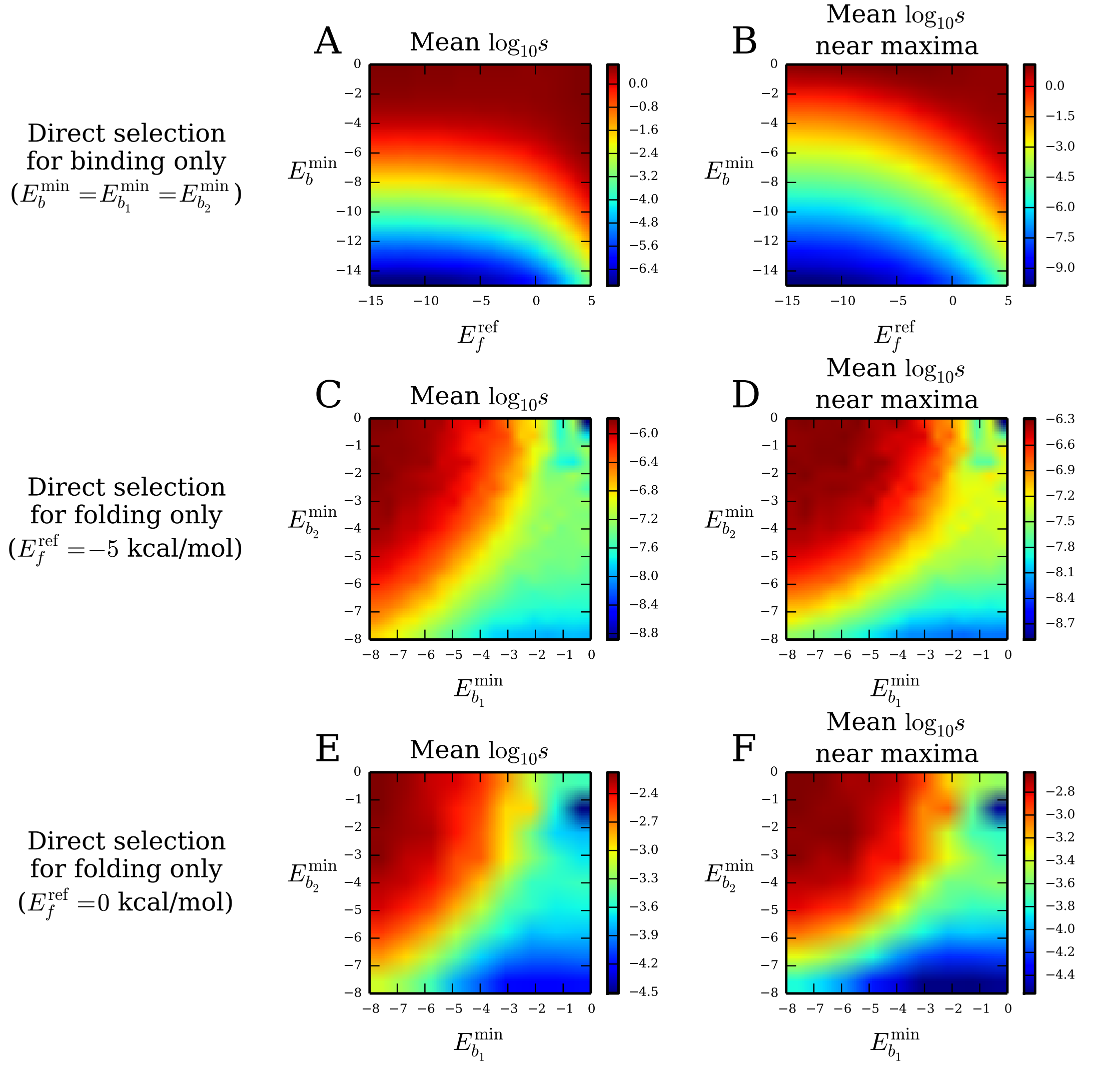


Figure 5.2: **Average selection strength.** (A) Average $\log_{10} s$ (s is the selection coefficient) of all accessible beneficial substitutions as a function of E_f^{ref} and $E_b^{\min} = E_{b_1}^{\min} = E_{b_2}^{\min}$ in the case of direct selection for binding only ($f_{\text{uf}} = f_{\text{ub}} = 0$). Due to the E_b symmetry of this case (Fig. 5.1B), we can neglect differences in $E_{b_1}^{\min}$ and $E_{b_2}^{\min}$ without loss of generality. (B) Same as (A) but limited to accessible substitutions that end at local fitness maxima. (C) Average $\log_{10} s$ of all accessible beneficial substitutions as a function of $E_{b_1}^{\min}$ and $E_{b_2}^{\min}$ in the case of selection for folding only ($f_{\text{uf}} = 0$, $f_{\text{ub}} = 1$, $E_f^{\text{ref}} = -5$ kcal/mol). (D) Same as (C) but limited to accessible substitutions that end at local fitness maxima. (E, F) Same as (C, D) but for $E_f^{\text{ref}} = 0$ kcal/mol. Simultaneous selection for both binding and folding yields qualitatively similar results. All data points are averages over 10^4 landscape realizations.

population begins as perfectly adapted to binding an original target molecule characterized by an energy matrix ϵ_{b_1} with minimum binding energy $E_{b_1}^{\text{min}}$ (defining a fitness landscape $\mathcal{F}_1$). The population is then subjected to a selection pressure which favors binding a new target, with energy matrix ϵ_{b_2} and minimum binding energy $E_{b_2}^{\text{min}}$ (fitness landscape $\mathcal{F}_2$). The population proceeds to adapt on this new landscape via the substitution dynamics of Eqs. 1.5 and 5.4. The adaptive paths are first-passage paths leading from the initial state to a local or global maximum on $\mathcal{F}_2$, with fitness increasing monotonically along each path.

Each adaptive path φ is a sequence of genotypes connecting the initial state (global maximum on $\mathcal{F}_1$) with the final state (local or global maximum on $\mathcal{F}_2$). We calculate various statistics of the adaptive path ensemble as discussed in Chapter 4. Specifically, we determine the path-length distribution $\rho(\ell)$, which gives the probability of taking an adaptive path with ℓ amino acid substitutions, and the mean adaptation time $\bar{t}$. We also determine S_{path} (S_Φ in the previous chapter), the entropy of the adaptive paths:

$$S_{\text{path}} = -\sum_{\varphi} \Pi[\varphi] \log \Pi[\varphi], \tag{5.5}$$

where $\Pi[\varphi]$ is the path probability functional defined in Eq. 4.10. The path entropy is maximized when evolution is neutral, resulting in all paths of a given length being accessible and equally likely: $S_{\text{path}} = \bar{\ell} \log L(k-1)$ [11], where k is the size of the amino acid alphabet, L is the number of residues, and $\bar{\ell}$ is the average path length.

Finally, we consider $\psi(\sigma)$, the probability of an adaptive path passing through or ending at a genotype σ. For final states σ this corresponds to their commitment probability, defined as the total probability of reaching the final state σ. We calculate the entropy S_{com} of the commitment probabilities as

$$S_{\text{com}} = -\sum_{\text{final states } \sigma} \psi(\sigma) \log \psi(\sigma), \tag{5.6}$$

where the sum is over all final states σ (local fitness maxima). The commitment entropy

$S_{\rm com}$ ranges from zero (for a single accessible final state) to $\log m_{\rm acc}$, where $m_{\rm acc}$ is the total number of accessible local maxima on $\mathcal{F}_2$.

We calculate these quantities using the exact numerical algorithm defined in Chapter 4. The substitution rate $W(\sigma'|\sigma)$ defines $\theta(\sigma) = (\sum_{\text{nn } \sigma' \text{ of } \sigma} W(\sigma'|\sigma))^{-1}$, the mean waiting time in genotype σ before a substitution occurs, where the sum is over all genotypes σ' one mutation away from σ (nearest mutational neighbors, "nn"). The substitution rates also determine the probability $Q(\sigma'|\sigma) = W(\sigma'|\sigma)\theta(\sigma)$ of making the substitution $\sigma \to \sigma'$, given that a substitution occurs out of σ.

For each substitution ℓ and intermediate genotype σ, we calculate $P_\ell(\sigma)$, the total probability of all paths that end at σ in ℓ substitutions; $T_\ell(\sigma)$, the total average time of all such paths; and $\Gamma_\ell(\sigma)$, their total entropy. As in Chapter 4, these quantities obey the recursion relations of Eq. 4.34. We use these transfer matrix objects to calculate the following path averages:

$$\begin{aligned} \rho(\ell) &= \sum_{\sigma \in \mathcal{S}_f} P_\ell(\sigma), & \psi(\sigma) &= \sum_{\ell=1}^{\Lambda} P_\ell(\sigma), \\ \bar{t} = \sum_{\ell=1}^{\Lambda} \sum_{\sigma \in \mathcal{S}_f} T_\ell(\sigma) &= \sum_{\ell=1}^{\Lambda} \tau(\ell) = \sum_{\sigma \in \mathcal{S}} \tau(\sigma), & \tau(\sigma) &= \sum_{\ell=1}^{\Lambda} \theta(\sigma) P_\ell(\sigma), \\ S_{\rm path} &= \sum_{\ell=1}^{\Lambda} \sum_{\sigma \in \mathcal{S}_f} \Gamma_\ell(\sigma), & \tau(\ell) &= \sum_{\sigma \in \mathcal{S}} \theta(\sigma) P_\ell(\sigma). \end{aligned} \tag{5.7}$$

The sums are calculated up to a path length cutoff Λ, which we choose such that $1 - \sum_{\ell=1}^{\Lambda} \rho(\ell) < 10^{-6}$. Note that the calculations for the state-dependent quantities $\psi(\sigma)$ and $\tau(\sigma)$ are simplified in this model (compared to the more general cases in Chapter 4) since the strong-selection dynamics prevents the population from traversing loops in genotype space.

5.2 Modes of adaptation

5.2.1 Direct selection for binding only

We first focus on the case where selection only acts directly on the binding trait: $f_{\mathrm{ub}} = f_{\mathrm{uf}} f_{\mathrm{ub}}$ in Eq. 5.3. This describes proteins with essential binding function (e.g., conferring antibiotic resistance), or nonessential proteins with a functional binding interface for which misfolding carries no additional fitness penalty beyond loss of function. This also includes directed evolution experiments where only function is artificially selected *in vitro*. The geometry of the fitness contours in this case is invariant under overall shifts in the binding energy E_b (Fig. 5.1B); equivalently, the direction (but not the magnitude) of the selection force ($\vec{\nabla} \log \mathcal{F} / |\vec{\nabla} \log \mathcal{F}|$) does not depend on E_b. Thus without loss of generality, we set $E_{b_1}^{\mathrm{min}} = E_{b_2}^{\mathrm{min}}$ in this section. The contours of constant fitness on this landscape are parallel to the E_f axis when E_f is low, indicating that, as expected, selection acts only on binding when proteins are sufficiently stable.

However, for marginally-stable proteins (E_f negative but close to zero), the fitness contours begin to curve downward. Most natural proteins appear to fall into this range of stabilities [59, 246, 264]. Even though selection only acts directly on the binding trait, this regime of the fitness landscape includes apparent, indirect selection for folding induced by the structural coupling between folding and binding: the protein can only bind when folded. Thus, adaptation will produce a trait (more stability) that is neutral at the level of the fitness function simply because it is coupled with another trait (binding) that is under selection. Folding stability can therefore be considered an evolutionary spandrel [248]. Proteins may even be intrinsically unstable ($E_f > 0$) and only fold when bound ($E_f + E_b < 0$), which we refer to as binding-mediated stability. In this regime, the fitness contours approach diagonal lines: selection effectively acts to improve both binding and folding equally (Fig. 5.1B), even though better folding is not advantageous *per se*.

An example realization of evolutionary dynamics in the marginally-stable regime is shown in Fig. 5.3A,B (see Fig. 5.4 for stable and intrinsically-unstable examples, and Fig. 5.5 for average distributions of initial, intermediate, and final states). There is typically just one or two fitness maxima (Fig. 5.3C), with a higher probability of multiple maxima for marginally-stable proteins that have genotype distributions situated near the region of curved fitness contours, leading to sign epistasis. Whenever there are multiple maxima, they are usually all accessible (Fig. 5.3C). For stable proteins, the global maximum almost always coincides with the best-binding genotype and is usually as far as a randomly-chosen genotype from the best-folding genotype (Fig. 5.3D; two random sequences are separated by $1 - 1/k = 0.8$ for $k = 5$). However, as E_f becomes greater, the average distance between the maxima and the best-binding genotype increases while the average distance between the maxima and the best-folding genotype decreases, until they meet halfway for intrinsically-unstable proteins, where effective selection for binding and folding is equally strong (Fig. 5.3D). In general the maxima lie on or near the Pareto front [77, 265], defined here as the set of genotypes such that either E_f or E_b cannot be decreased further without increasing the other (the global maximum is always on the front, while local maxima may not be). The location of the maxima along the Pareto front varies with the overall regime of the folding energy (Fig. 5.3A, Fig. 5.4).

As E_f increases, the average distance between initial and final states for adaptation decreases. As a result the average path length (number of substitutions) decreases as well, although the variance in path length is relatively constant over all energies (Fig. 5.3E). The path entropy per-substitution $S_{\text{path}}/\bar{\ell}$ also decreases with E_f: starting closer to the final state means there are fewer possible beneficial substitutions, resulting in more constraints on adaptive paths. Note that $S_{\text{path}}/\bar{\ell} = \log L(k-1) \approx 3.2$ for neutral evolution, more than twice as large as the path entropy observed in all regimes; this indicates that adaptive paths are significantly constrained by selection. Finally, the entropy S_{com} of the commitment probabilities ranges from zero for a single maximum to ≈ 0.31 in the marginally-stable regime (Fig. 5.3F). Since the average number of maxima is ≈ 1.9 in this regime (Fig. 5.3C),

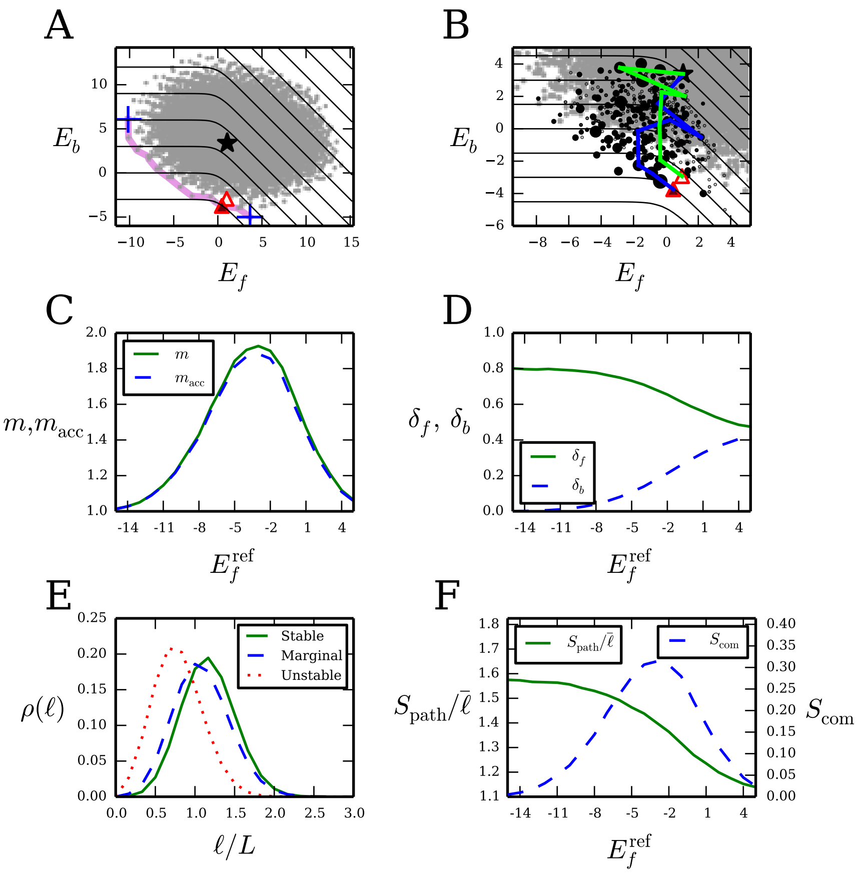


Figure 5.3: **Properties of adaptation with direct selection for binding only.** (A) Global distribution of folding and binding energies for all $k^L = 5^6$ genotypes in a single realization of the model with a marginally-stable protein ($E_f^{\mathrm{ref}} = -3$ kcal/mol). The black star indicates the initial state for adaptation (global maximum on $\mathcal{F}_1$), red triangles indicate local fitness maxima on $\mathcal{F}_2$ (shaded according to their commitment probabilities $\psi(\sigma)$), and the blue crosses indicate best-folding and best-binding genotypes. The magenta line connects genotypes on the Pareto front, and the black contours indicate constant fitness $\mathcal{F}_2$. (B) The region of energy space accessible to adaptive paths (zoomed in from (A)). Example paths are shown in blue and green; black circles indicate intermediate states along paths, sized proportional to their path density $\psi(\sigma)$; small gray circles are genotypes inaccessible to adaptation. (C) Average number m of local fitness maxima (solid, green) and average number m_{acc} of local maxima accessible to adaptation (dashed, blue) versus E_f^{ref}. The average number of maxima is greatest at $E_f^{\mathrm{ref}} \approx -3$ kcal/mol, where multiple local maxima are separated by ≈ 2.23 substitutions on average. (D) Average per-residue Hamming distance between the maxima and the best-folding genotype (δ_f; solid, green) and the best-binding genotype (δ_b; dashed, blue) versus E_f^{ref}. (E) Average distributions $\rho(\ell)$ of path lengths (number of substitutions) ℓ for stable proteins ($E_f^{\mathrm{ref}} = -15$ kcal/mol), marginally-stable proteins ($E_f^{\mathrm{ref}} = -3$ kcal/mol), and intrinsically-unstable proteins ($E_f^{\mathrm{ref}} = 5$ kcal/mol). (F) Per-substitution path entropy $S_{\mathrm{path}}/\bar{\ell}$ (solid, green) and entropy S_{com} (dashed, blue) of the final state commitment probability distribution $\psi(\sigma)$ versus E_f^{ref}. Panel (E) is averaged over 10^5 realizations of the model; all other averages are taken over 10^4 realizations. In all panels $f_{\mathrm{ub}} = f_{\mathrm{uf}} = 0$ and $E_{b_1}^{\mathrm{min}} = E_{b_2}^{\mathrm{min}} = -5$ kcal/mol.

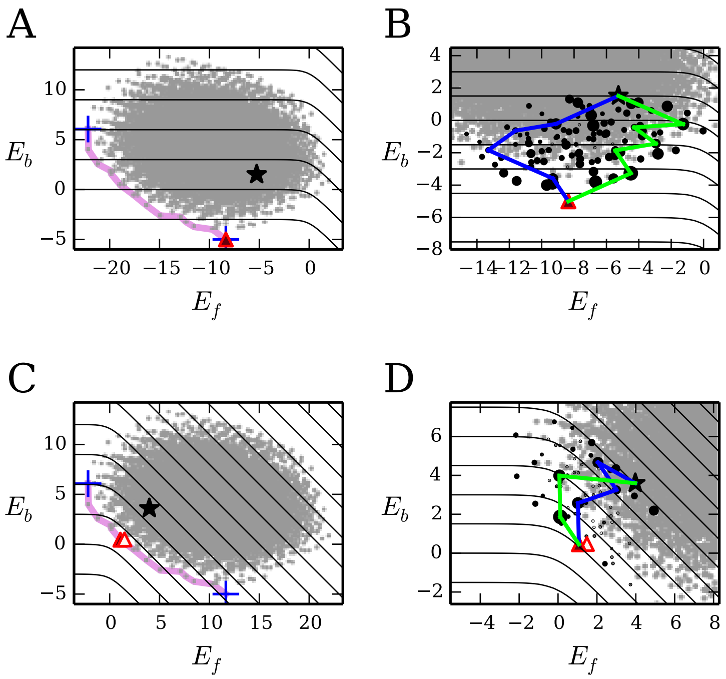


Figure 5.4: **Example landscapes for stable and intrinsically-unstable proteins with direct selection for binding only.** Symbols and randomly-generated energy matrices (ϵ_f, ϵ_{b_1}, and ϵ_{b_2}) are the same as in Fig. 5.3A,B. (A, B) Stable protein ($E_f^{\text{ref}} = -15$ kcal/mol). (C, D) Intrinsically-unstable protein ($E_f^{\text{ref}} = 5$ kcal/mol). As in Fig. 5.3A,B, $f_{\text{ub}} = f_{\text{uf}} = 0$ and $E_{b_1}^{\min} = E_{b_2}^{\min} = -5$ kcal/mol.

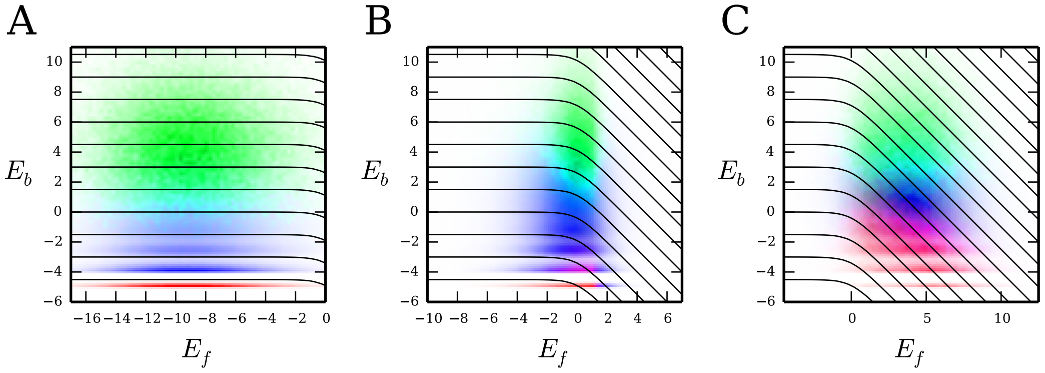


Figure 5.5: **Average landscapes for direct selection for binding only.** As in Fig. 5.8C, the distribution of initial states is shown in green, intermediate states in blue (weighted by their path density $\psi(\sigma)$), and final states in red (weighted by their commitment probabilities $\psi(\sigma)$). (A) Stable proteins ($E_f^{\text{ref}} = -15$ kcal/mol). (B) Marginally-stable proteins ($E_f^{\text{ref}} = -3$ kcal/mol). (C) Intrinsically-unstable proteins ($E_f^{\text{ref}} = 5$ kcal/mol). All landscapes are averaged over 10^5 realizations. As in Fig. 5.3A,B, $f_{\text{ub}} = f_{\text{uf}} = 0$ and $E_{b_1}^{\min} = E_{b_2}^{\min} = -5$ kcal/mol.

the maximum value of the commitment entropy is $\log 1.9 \approx 0.64$, over twice as large. This indicates that even when multiple maxima are present, there is significant bias towards one of them.

5.2.2 Direct selection for folding only

We next consider proteins for which only folding is directly selected: $f_{\mathrm{ub}} = 1$ and $0 \leq f_{\mathrm{uf}} < 1$ in Eq. 5.3. This situation may arise if misfolded proteins are toxic, if they impose an increased burden on protein synthesis and quality control machinery, or if misfolding interferes with another function of the protein besides the binding interaction considered here [238–241]. Here the binding interface under consideration is nonfunctional in the sense that binding confers no fitness advantage, although the protein may also have other, functional binding interfaces. Similar to the previous case, the geometry of the fitness contours and thus most landscape properties are now independent of E_f (Fig. 5.1C); equivalently, normalized selection force $\vec{\nabla} \log \mathcal{F} / |\vec{\nabla} \log \mathcal{F}|$ does not depend on E_f.

When the nonfunctional binding is weak, the fitness contours are parallel to the E_b axis, indicating that selection acts only on folding (Fig. 5.1C). However, with increasing binding strength the fitness contours begin to curve such that the effective selection force attempts to improve both binding and folding equally. Thus binding emerges as an evolutionary spandrel: although it is nonfunctional by itself, adaptation may produce strong binding anyway due to its coupling with folding. Similar to the previous case with no misfolding penalty, the weak-binding regime yields a single fitness maximum due to the lack of sign epistasis (Fig. 5.6A). This maximum predominantly coincides with the best-folding genotype. However, once the binding interaction becomes stronger (lower E_b), there is an increased likelihood of multiple local maxima, located between the best-folding and best-binding genotypes.

Depending on the abundance of the old and new ligands in the cell and their binding properties, several adaptive scenarios may take place. First, the best-binding strengths

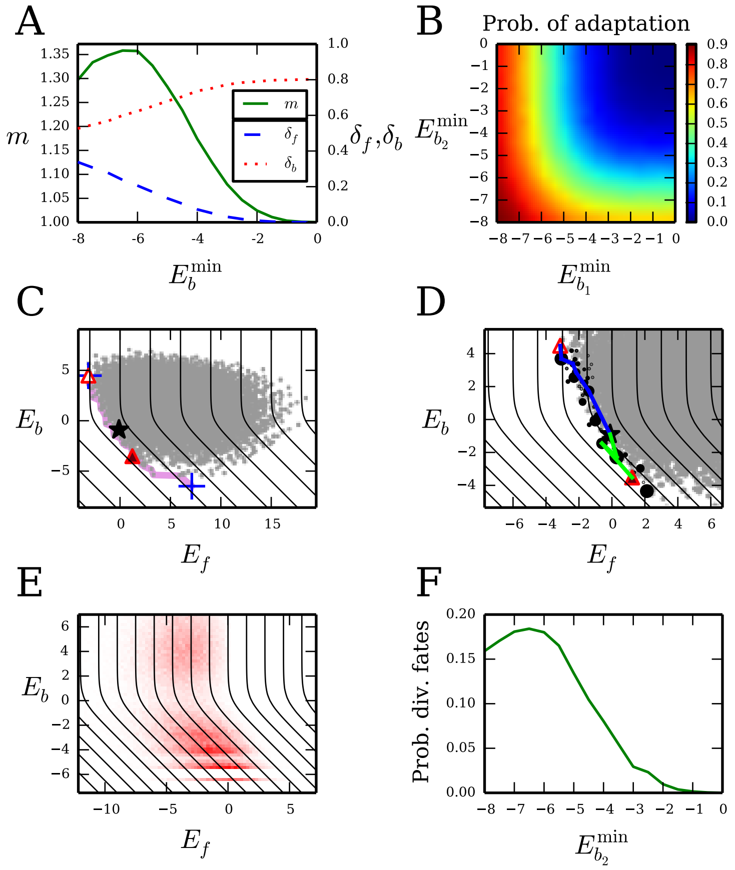


Figure 5.6: **Properties of adaptation with direct selection for folding only.** (A) The average number of local maxima m (solid, green) and their average per-residue Hamming distance from the best-folding (δ_f; dashed, blue) and the best-binding (δ_b; dotted, red) genotypes versus E_b^{min}. (B) Probability that adaptation occurs when the binding target is changed (i.e., the initial state is not coincident with any of the final states), as a function of $E_{b_1}^{\text{min}}$ and $E_{b_2}^{\text{min}}$. (C,D) Example landscape with divergent binding fates: there are two accessible local maxima, one with $E_b < 0$ (favorable binding) and the other with $E_b > 0$ (negligible binding). All symbols are the same as in Fig. 5.3A,B. The commitment entropy is $S_{\text{com}} \approx 0.67$ (probabilities of 0.6 for the $E_b < 0$ maximum and 0.4 for the $E_b > 0$ maximum). (E) Average distribution of local maxima (weighted by their commitment probabilities). The average commitment entropy for realizations with divergent fates is $S_{\text{com}} \approx 0.43$. In (C)–(E) we use $E_{b_1}^{\text{min}} = E_{b_2}^{\text{min}} = -6.5$ kcal/mol. (F) The probability of having divergent fates versus $E_{b_2}^{\text{min}} = E_{b_1}^{\text{min}}$. Panel (E) is averaged over 10^5 realizations of the model; all other averages are taken over 10^4 realizations. In all panels $f_{\text{ub}} = 1$, $f_{\text{uf}} = 0$, and $E_f^{\text{ref}} = 0$ kcal/mol.

$E_{b_1}^{\min}$ and $E_{b_2}^{\min}$ of the old and new targets may be similar in magnitude. If both are weak, initial and final states are likely to be the best-folding genotype or close to it (Fig. 5.6A); in this case, there is a high probability that no adaptation will occur (Fig. 5.6B). When $E_{b_1}^{\min}$ and $E_{b_2}^{\min}$ are both low, adaptation usually occurs to accommodate the binding specificity of the new ligand (Fig. 5.6B, Fig. 5.7A). Effective selection favors improving both folding and binding equally in this regime, even though binding is not under direct selection. Surprisingly, we see that proteins frequently evolve stronger binding at the expense of folding (bottom panel of Fig. 5.7A). This happens due to the constraints of the genotype-phenotype map: not enough genotypes are available to optimize both traits simultaneously.

It is also possible to gain or lose binding affinity at the nonfunctional interface through adaptation. In the first case, the new target has stronger binding than the old one ($E_{b_2}^{\min} < E_{b_1}^{\min}$). Thus the initial state is the best-folding genotype or close to it, and the protein adapts toward a genotype with intermediate folding and binding (Fig. 5.7B). As before, adaptation is tightly constrained by the genotype-phenotype map, sacrificing the trait (folding stability) under direct selection in order to affect the spandrel (nonfunctional binding interaction). Effectively, the protein switches from being "self-reliant" to needing a binding partner to remain viable. In the second case ($E_{b_1}^{\min} < E_{b_2}^{\min}$), the dynamics is opposite: the initial state is an intermediate genotype between the best-folding and best-binding genotype, and the final state is likely to be the best-folding genotype as the binding interface is effectively lost (Fig. 5.7C). Here, the protein becomes self-reliant. Thus proteins may acquire or lose binding interfaces depending on the availability of ligands (or classes of ligands) that can participate in binding-mediated stability. More generally, if the protein stability is initially suboptimal due to an environmental change, the stability may be restored not only through stabilizing mutations, but also by developing a novel binding interface which may be specific or non-specific.

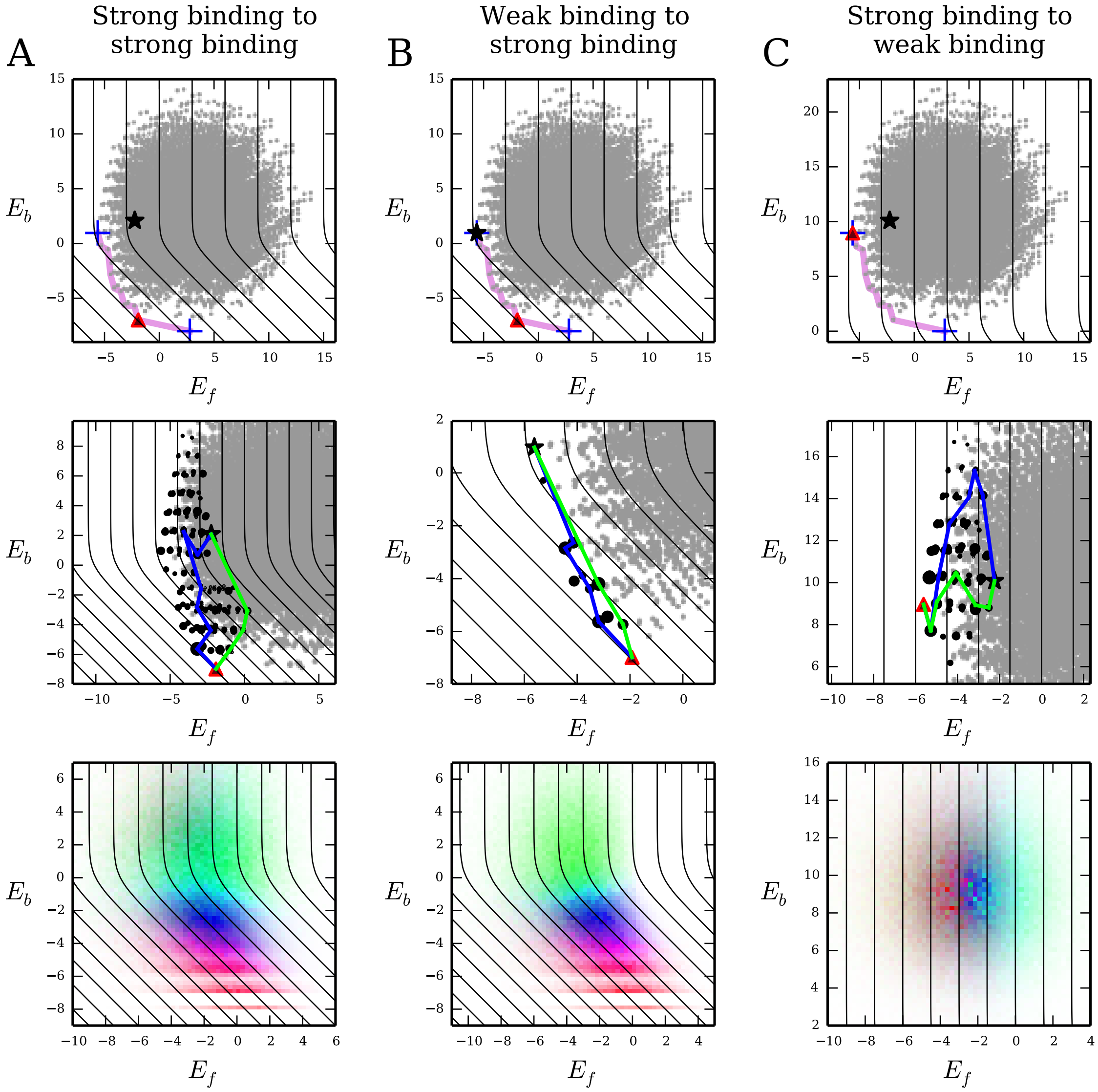


Figure 5.7: **Example and average landscapes for direct selection for folding only.** Symbols in top and middle panels are the same as in Fig. 5.3A,B, and the color scheme in the bottom panels is the same as in Fig. 5.8C and Fig. 5.5. (A) Strong binding to both old and new targets ($E_{b_1}^{\min} = E_{b_2}^{\min} = -8$ kcal/mol). (B) Weak binding to old target and strong binding to new target ($E_{b_1}^{\min} = 0$ kcal/mol, $E_{b_2}^{\min} = -8$ kcal/mol). (C) Strong binding to old target and weak binding to new target ($E_{b_1}^{\min} = -8$ kcal/mol, $E_{b_2}^{\min} = 0$ kcal/mol). We use $f_{\text{ub}} = 1$, $f_{\text{uf}} = 0$, and $E_f^{\text{ref}} = 0$ kcal/mol in all cases. In the bottom panels, the landscapes are averaged over 10^5 realizations.

5.2.3 Divergent evolutionary fates

In the region where the fitness contours in Fig. 5.1C are curved, it is possible to have two or more local maxima accessible to adaptation, with at least one having negative E_b (strong binding) and at least one having positive E_b (negligible binding) (see Fig. 5.6C,D for an example landscape). The selection streamlines are divergent in this regime (Fig. 5.1C), leading to the possibility of maxima that are widely separated in energy space. Thus a protein may have two fates available to it: one in which it evolves to bind the target and another in which it does not. The eventual fate of the protein is determined by random mutation events. Indeed, the distribution of final states is strongly bimodal (Fig. 5.6E), yielding a sizable probability of divergent fates across a range of binding strengths (Fig. 5.6F).

5.2.4 Simultaneous selection for binding and folding

Finally we consider a general case in which there are distinct selection pressures on both binding and folding ($0 < f_{\mathrm{ub}} < 1$ and $0 \leq f_{\mathrm{uf}} < 1$ in Eq. 5.3; Fig. 5.1D). This scenario is realized when the binding interaction is functional but nonessential, while protein misfolding entails a fitness penalty beyond mere loss of function. The fitness landscape is divided into two regions by a straight diagonal contour with fitness f_{ub} and slope -1. Below this contour, the landscape is qualitatively similar to the case of selection for binding only (Fig. 5.1B), while above the contour the landscape resembles that of the folding-only selection scenario (Fig. 5.1C). Thus evolutionary dynamics for proteins with favorable binding and folding energies will largely resemble the case of selection for binding only. However, a qualitatively different behavior will be observed if the distribution of genotypes straddles the diagonal contour (Fig. 5.8). This will occur when initial folding stability is marginal and initial binding is unfavorable. In this case, selection streamlines around the diagonal contour (Fig. 5.1D) and the genotype-phenotype map tightly constrain the adaptive paths to gain extra folding stability first, and then lose it as the binding function is improved.

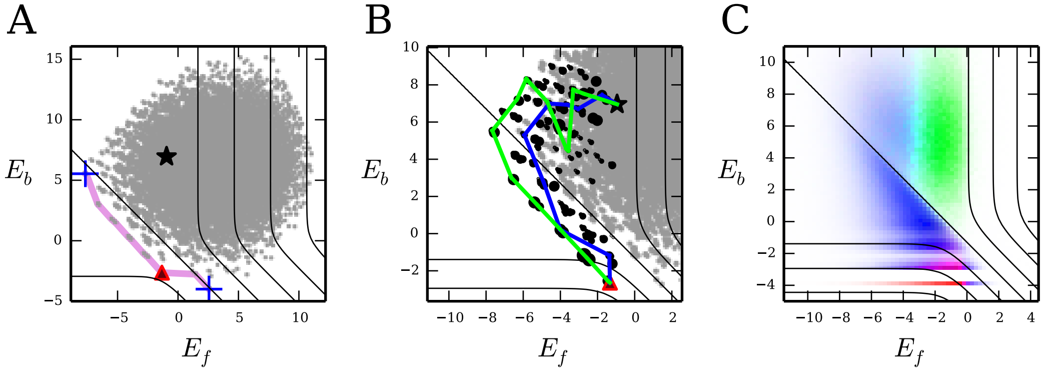


Figure 5.8: **Properties of adaptation with direct selection for both folding and binding.** (A, B) Distribution of folding and binding energies in an example landscape for a marginally-stable and marginally-bound protein; all symbols are the same as in Fig. 5.3A,B. (C) Landscape averaged over 10^5 realizations. Distribution of initial states is shown in green, intermediate states in blue (weighted by their path densities $\psi(\sigma)$), and final states in red (weighted by their commitment probabilities $\psi(\sigma)$). In all panels $f_{\mathrm{ub}} = 0.9$, $f_{\mathrm{uf}} = 0$, and $E_f^{\mathrm{ref}} = E_{b_1}^{\mathrm{min}} = E_{b_2}^{\mathrm{min}} = -4$ kcal/mol.

5.2.5 Tempo and rhythm of adaptation

Lastly we consider the temporal properties of protein adaptation. The strength of selection is the primary determinant of the average adaptation time $\bar{t}$. If the selection coefficient s is small (but $Ns > 1$), the substitution rate $W(\sigma'|\sigma)$ in Eq. 1.5 is proportional to s. Thus the dependence of total adaptation time on the energies E_f and E_b is very similar to the dependence of selection strength on energies (Fig. 5.2): as selection becomes exponentially weaker for lower energies, adaptation becomes exponentially slower. The distribution of total adaptation time over an individual adaptive path is highly nonuniform on average. For example, in the case of selection for binding only and a marginally-stable protein, the adaptation time is concentrated at the end of the path, one mutation away from the final state (Fig. 5.9A,B). Substitutions at the beginning of the path occur quickly because there are many possible beneficial substitutions and because selection is strong; in contrast, at the end of the path adaptation slows down dramatically as beneficial mutations are depleted and selection strength weakens. This behavior is observed in most of the other model regimes as well.

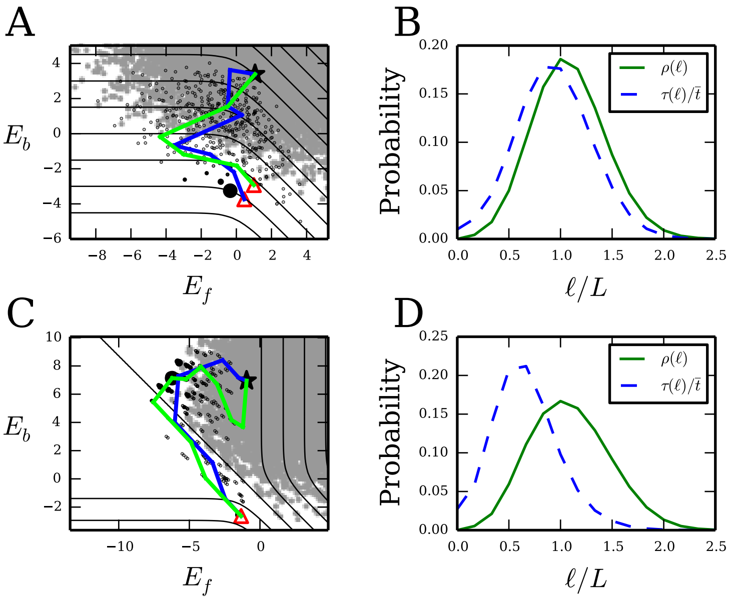


Figure 5.9: **Distribution of adaptation times over intermediate states.** (A) The same landscape realization as in Fig. 5.3A,B (selection for binding only on a marginally-stable protein), but with each intermediate state σ sized proportional to $\tau(\sigma)$, the average time spent in that state. (B) The probability $\rho(\ell)$ (solid, green) of taking an adaptive path of exactly ℓ substitutions and the average time $\tau(\ell)$ (dashed, blue) spent by paths at the ℓth substitution, averaged over 10^5 realizations with $f_{\text{ub}} = f_{\text{uf}} = 0$, $E_f^{\text{ref}} = -3$ kcal/mol, and $E_{b_1}^{\text{min}} = E_{b_2}^{\text{min}} = -5$ kcal/mol. (C, D) Same as (A, B), but with the landscape realization used in Fig. 5.8A,B (selection for both binding and folding, $f_{\text{ub}} = 0.9$, $f_{\text{uf}} = 0$, $E_f^{\text{ref}} = E_{b_1}^{\text{min}} = E_{b_2}^{\text{min}} = -4$ kcal/mol).

The exception to this pattern occurs in the case of selection for both binding and folding in marginally-stable and marginally-bound proteins, due to the unique contour geometry (Fig. 5.1D). As the adaptive paths wrap around the diagonal contour in the region of high E_b and low E_f, the landscape flattens, making selection weaker and substitutions slower (Fig. 5.9C). Thus the intermediate states near the diagonal contour dominate the total adaptation time, and most of the waiting occurs in the middle of the path rather than the end (Fig. 5.9D). Adaptation accelerates toward the end of the path as the strength of selection increases again. If the intermediate slow-down is significant enough, a protein may not have time to complete the second half of its path before environmental conditions change, so that it will never actually evolve the new binding function.

5.3 Discussion

Protein folding and binding as evolutionary spandrels. In the decades since Gould and Lewontin's influential paper [248], the existence of evolutionary spandrels has emerged as a critical concept to understand as we attempt to infer evolutionary history from present-day traits of organisms. There are many possible scenarios in which spandrels can evolve [248, 249], although two key mechanisms are neutral processes, such as genetic drift and biases in mutation and recombination [266], and indirect selection arising from coupled traits. Here we have focused on the latter, which we expect to be more important on the short time scales considered in our model.

Taverna and Goldstein [246] previously argued that the marginal stability of most proteins may be an evolutionary spandrel that evolved not due to direct selection (as argued by some authors, e.g. Ref. [55]), but due to mutation-selection balance [40, 59, 246]. Our model more broadly argues that having folding stability at all may be a spandrel for proteins with no misfolding toxicity. Even more striking is the possibility that some binding interactions may be spandrels that evolved solely to stabilize proteins with toxic misfolding. The structural coupling of folding and binding has been previously discussed in the context of intrinsically-disordered proteins, some of which only acquire an ordered structure upon binding [267, 268], as described by the high E_f regime of our model. More recently, Dixit and Maslov [269] studied the role of binding-mediated stability in the yeast proteome, estimating the effective extra protein stability conferred by binding. Our model reveals how readily such nonfunctional binding interactions can evolve in proteins. The coupling between folding and binding fundamentally arises from the fact that at finite temperature, proteins fluctuate between different structural states. In particular, we expect more widespread nonfunctional interactions among proteins with less intrinsic stability and therefore more sensitivity to thermal fluctuations. It is not possible to directly test this prediction in the absence of whole-proteome stability measurements, but it is strikingly consistent with available proteome-wide observations. Specifically, protein abundance is believed to correlate positively with stability ($-E_f$) to explain the observed negative correlation of abundance with evolutionary rate [57, 61, 62]. Furthermore, models of protein-protein interaction networks imply that protein abundance also correlates negatively with the number of interactions [66]. Together these argue that stability should indeed be negatively correlated with the number of interactions, as expected from our model. Experiments on specific proteins support this finding: for example, destabilizing mutations in *E. coli* dihydrofolate reductase (DHFR) were found to be compensated at high temperature by protein binding, which protected against toxic aggregation [255].

In Appendix D we contrast these findings with a version of the model in which folding and binding are *not* coupled, but occur independently. Some regimes produce qualitatively similar landscapes and adaptive dynamics, but independent folding and binding precludes the possibility of indirect selection and its consequences.

Pareto optimization of proteins. The Pareto front is a useful concept in problems of multi-objective optimization [77, 265]. Proteins are one such system, since in general they must optimize both folding stability and binding affinity for their targets. The Pareto front in our model consists of the protein sequences along the low E_f, low E_b edge of the genotype distribution (Figs. 5.3A, 5.6C, 5.8A, 5.4A,C, 5.7). Pareto optimization assumes that all states on the front are valid final states for adaptation; this in turn implies that fitness has linear dependence on the individual traits. However, nonlinear fitness functions with saturation-like effects will confound this assumption. Indeed, in the case of proteins, once stability or binding is "good enough," there should be little selection pressure to improve it further. Our model shows how this leads to a small subset of genotypes on or even off the front, usually just one of two, that are true final states for adaptation. Thus Pareto optimization does not capture a key feature of the underlying biophysics, providing only a rough approximation to the true dynamics.

Epistasis. Our model also sheds light on the role of epistasis — the correlated effects of mutations at different sites — in protein evolution. Epistasis underlies the ruggedness of fitness landscapes and their accessibility to evolving populations [24, 32]. Here we consider a model of folding and binding energies that is purely additive (Eq. 5.2). Such energies nevertheless result in epistasis through nonlinear dependence of fitness on energy (Eqs. 5.1 and 5.3). Magnitude epistasis is widespread in our model, while sign epistasis only arises in regions where the fitness contours are curved (Fig. 5.1E,F). This picture of prevalent magnitude epistasis but limited sign epistasis is qualitatively consistent with studies of empirical fitness landscapes [24], and with directed evolution experiments that generally report high accessibility of protein sequence space [40].

Double mutant experiments indicate that the additive energy model is a good approximation for residues that are not in direct physical contact [259, 260, 270, 271]. For spatially-close residues, the mutational effects are largely "sub-additive" (diminishing-returns magnitude epistasis): two (de)stabilizing mutations combined will still usually be (de)stabilizing, but less so than the sum of their individual effects [259, 270]. For example, Istomin et al. [270] find that while residues separated by more than 6 Å are nearly additive (correlation $R^2 = 0.97$ with a slope of 0.88 between the sum of $\Delta\Delta G$'s for two single mutants and $\Delta\Delta G$ for the double mutant), spatially-close residues are substantially sub-additive

($R^2 = 0.84$, slope of 0.54). Nonetheless, in regions with straight contours which represent most of the fitness landscape, sub-additive energies cannot produce sign epistasis; substantial deviations from energy sub-additivity are required to create additional local maxima or place significant constraints on adaptive paths. Thus it appears that deviations from energy additivity will not lead to qualitative changes in our model's predictions.

Repeatability of evolution. Epistasis determines the repeatability of evolution, an issue of paramount importance in biology [14, 31, 85]. How predictable are the intermediate pathways and final outcomes for a protein evolving a new binding interaction? We use the path entropy S_{path} and the commitment entropy S_{com} to address these issues quantitatively. In many cases of the model, we see a diverse ensemble of pathways (high S_{path}) due to the minimal sign epistasis, while the entropy S_{com} of final states is low because there is usually only a single final state. Thus low sign epistasis gives rise to less predictable intermediate pathways but highly predictable final outcomes.

However, there are two major exceptions to this pattern. First, for proteins with a binding interaction under no direct selection, there is a substantial probability of having multiple local maxima, with at least one having strong binding and another with weak binding (Fig. 5.6). Here both intermediate pathways and final states are unpredictable — pure chance, in the form of random mutations, drives the population to one binding fate or the other. The possibility of nonfunctional interactions randomly evolving must be accounted for in interpretations of observed protein interactions in naturally-evolved organisms [66, 247]. The second exception occurs in proteins with direct selection for both binding and folding. Here there is limited sign epistasis (usually a single maximum), but the adaptive paths are tightly constrained in energy space (Fig. 5.8), gaining extra stability first and then losing it as function is improved. Since most natural proteins appear to have functional binding as well as toxic misfolding, natural protein evolution may be repeatable at the level of folding and binding energy traits.

Chapter 6

Conclusion and Outlook

In the foregoing chapters we have combined biophysical models, methods from stochastic processes and statistical physics, and high-throughput data to address some questions on the physical principles of evolution at the molecular scale. We have developed theoretical tools in Chapters 2 and 4 and then applied them to specific systems in Chapters 3 and 5. We have focused especially on how DNA-protein interactions and protein binding and folding shape fitness landscapes, and how these landscapes in turn shape the evolutionary pathways available to populations.

For transcription factor binding sites (Chapter 3), we found a simple model inspired by thermodynamics adequately captures the observed diversity of binding sites in yeast; however, the mismatch between the effective parameters of the inferred fitness landscape and our biophysical expectations implies the importance of other factors beyond simple DNA-protein interactions in binding site evolution. Additional data on DNA-protein interactions, combined with a more sophisticated model of gene regulation and evolution, will be necessary to complete this picture.

We also studied a model of protein evolution based on the biophysics of folding and binding (Chapter 5). When there are distinct selection pressures on both folding and binding, as is believed to be the case for many natural proteins, our model predicts that evolutionary paths are tightly constrained at the level of biophysical traits: folding stability is first gained and then partially lost as the new binding function is developed. We look forward to directly verifying this qualitative prediction with *in vivo* or *in vitro* evolution experiments.

But the most important consequence of our model is conceptual: the structural coupling of folding and binding enables protein traits to evolve as “spandrels” in the absence of direct selection. In particular, proteins can evolve strong binding interactions that confer no intrinsic fitness advantage but merely serve to stabilize the protein if misfolding is deleterious. This observation is highly suggestive in light of recent experiments showing a broad-tailed distribution of interaction partners for naturally-evolved proteins, which point toward widespread nonfunctional interactions. We plan to use this existing high-throughput data as well as controlled evolution experiments to look for signatures of our model predictions.

However, these are only two areas in which biophysics is being fruitfully applied to understanding evolution at the molecular scale. The rapid development of this subfield of physical biology has been spurred by the enormous experimental and computational advances of the past 20 years. In particular, high-throughput techniques such as whole-genome sequencing [272], DNA microarrays [164], and microfluidics [166], combined with massive international databases of molecular properties [264, 273], interactomes [274], and genomics [275] have revolutionized the field and will likely continue to produce even more abundant and precise data in the coming years.

The wealth of data will further fuel our efforts to build quantitative models and comprehensive theories of living matter, especially in terms of evolution. A multitude of outstanding questions remains. What are the structures of fitness landscapes for complex molecular networks, such as regulatory and metabolic networks, and how do populations traverse them? What is the role of population dynamics — such as recombination, clonal interference, ecological interactions, demography, and spatial structure — in molecular evolution? How do collective behaviors seen in microbial colonies and multicellular organisms evolve? And what can we learn from natural evolution to help us engineer synthetic organisms that perform useful industrial and medical functions?

It is indeed an exciting time for science at the interface of physics and biology, as researchers from a host of traditionally-distinct disciplines unite in pursuit of common goals. An era of unprecedented growth is upon us as the ambitions of our forerunners such as Schrödinger come to fruition. It holds the promise to transform not only the technology in our daily lives, but also our understanding of the most basic question we can ask about ourselves: what is life?

Appendix A

Additional Results for the Scaling Law

A.1 The scaling law in the weak-selection limit

Here we present an argument that the leading-order behavior of $\psi(r)$ is always a power law in the diffusion limit. Since $\nu = 2N\phi'(1)$, this is equivalent to showing that $\phi'(1) \neq 0$, which means that the fixation probability must be locally linear around the neutral limit $r = 1$. The fixation probability in the diffusion approximation is given by [80]

$$\phi(r) = \frac{\int_0^{1/N} dx\ G(x,r)}{\int_0^1 dx\ G(x,r)}, \quad G(x,r) = \exp\left(-2\int_0^x dy\ \frac{M(y,r)}{V(y,r)}\right), \tag{A.1}$$

where $M(x,r)$ and $V(x,r)$ are the first two moments of the change in mutant fraction x per unit time. Define expansions of the moments:

$$\begin{aligned} M(x,r) &= M_0(x) + (r-1)M_1(x) + \mathcal{O}((r-1)^2) \\ V(x,r) &= V_0(x) + (r-1)V_1(x) + \mathcal{O}((r-1)^2). \end{aligned} \tag{A.2}$$

Since evolution under pure drift ($r = 1$) is unbiased, the mean change in mutant fraction without selection is zero: $M_0(x) = 0$. Substituting these expansions into Eq. A.1 and expanding to lowest order in $r - 1$, we obtain

$$\phi(r) = \frac{1}{N} + 2(r-1)\left(\frac{1}{N}\int_0^1 dx \int_0^x dy\ \frac{M_1(y)}{V_0(y)} - \int_0^{1/N} dx \int_0^x dy\ \frac{M_1(y)}{V_0(y)}\right) + \mathcal{O}((r-1)^2). \tag{A.3}$$

Therefore

$$\phi'(1) = 2\left(\frac{1}{N}\int_0^1 dx \int_0^x dy\ \frac{M_1(y)}{V_0(y)} - \int_0^{1/N} dx \int_0^x dy\ \frac{M_1(y)}{V_0(y)}\right), \tag{A.4}$$

where $\phi'(1) = d\phi(r)/dr|_{r=1}$. Note that $V_1(x)$ does not appear — the correction to the second moment by weak selection does not affect the fixation probability expanded to the lowest order. Thus, barring some coincidental cancellation of terms in Eq. A.4, $\phi'(1)$ should be nonzero as long as $M_1(x)$ is nonzero.

To argue that $M_1(x) \neq 0$, we invoke an operational definition of selection strength described in Chapter 1. Experimental measurements of selection strength are often made by inferring it as the exponential growth rate of a small mutant sub-population, at least for microorganisms [276], so we require that the population model show this behavior. If X is the random variable denoting the fraction of mutants in the population, its deterministic equation is

$$\frac{d}{dt}\mathbb{E}[X] = \mathbb{E}[M(X,r)], \tag{A.5}$$

where $\mathbb{E}[\cdot]$ is the expected value operator. In the limit of weak selection ($r \sim 1$) and small mutant fraction ($X \ll 1$),

$$\frac{d}{dt}\mathbb{E}[X] \approx (r-1)\mathbb{E}[M_1(X)] \propto (r-1)\mathbb{E}[X], \tag{A.6}$$

assuming that $M_1(x)$ is linear in x to the lowest order. This yields exponential growth at a rate proportional to the selection strength $s = r - 1$. Therefore $M_1(x)$ should be nonzero and hence $\phi'(1)$ is nonzero, establishing the power-law behavior of $\psi(r)$ in the limit of weak selection.

Equation A.4 suggests an interpretation of ν. Under the appropriate rescaling of time units, the pure drift $V_0(x)$ is proportional to $1/N$ and $M_1(x)$ is independent of N. For example, this is true in the Wright-Fisher model with generations as the time unit, and it also holds in the Moran model with the single birth/death time scaled by a factor of N. Then Eq. A.4 implies that $\phi'(1) \sim \mathcal{O}(N^0)$, and therefore $\nu \sim \mathcal{O}(N)$. This observation can be generalized to a broader class of models in which $V_0(x)$ is proportional to $1/N_e$, where N_e is the variance effective population size [20, 102].

A.2 Exact Wright-Fisher fixation probability from discrete Markov chain

Studying discrete Markov chain properties of the Wright-Fisher model is not new [20]. However, previous work has typically focused on explicit results using spectral theory, with particular emphasis placed on neutral evolution. In contrast, we will obtain an implicit result suitable for numerical application. These results will allow investigation of the dynamics of the model under large selection effects that are beyond the scope of diffusion theory.

The transition probabilities $\langle n'|\mathbf{P}|n\rangle$ from Eq. 2.28 are elements of an $(N+1)\times(N+1)$ matrix $\mathbf{P}$. We will adopt the convention in which the final state n' is the row index and the initial state n is the column index. Transition probabilities between different states at different time steps are given by the matrix elements of powers of $\mathbf{P}$. That is, the probability of transitioning from n to n' in m generations is given by $\langle n'|\mathbf{P}^m|n\rangle$. Therefore the probability of fixation in m generations from initial state n is given by $\langle N|\mathbf{P}^m|n\rangle$, and the probability of fixing a single mutant in the infinite time limit is given by

$$\lim_{m\to\infty}\langle N|\mathbf{P}^m|1\rangle = \phi(r). \tag{A.7}$$

This limit can be conveniently expressed by permuting the states to group the transient states ($n = 1,\ldots,N-1$) together and the absorbing states ($n = 0, N$) together. Define elements of the $(N-1)\times(N-1)$ submatrix $\langle i|\mathbf{A}|j\rangle = \langle i|\mathbf{P}|j\rangle$ for $i, j = 1,\ldots,N-1$; this matrix describes transitions between transient states only. Next, define elements of the

$2 \times (N-1)$ submatrix $\langle \alpha | \mathbf{B} | i \rangle = \langle \alpha | \mathbf{P} | i \rangle$ for $\alpha = 0, N$ and $i = 1, \ldots, N-1$; this matrix describes single-generation transitions from transient states to absorbing states. Now we permute the indices to put $\mathbf{P}$ in the canonical form [277]:

$$\mathbf{P} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{B} & \mathbf{1}_2 \end{bmatrix}, \tag{A.8}$$

where $\mathbf{0}$ is the $(N-1) \times 2$ zero matrix and $\mathbf{1}_k$ is a $k \times k$ identity matrix. We can now easily compute the infinite time limit:

$$\begin{aligned} \lim_{m\to\infty} \mathbf{P}^m = \lim_{m\to\infty} \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{B} & \mathbf{1}_2 \end{bmatrix}^m &= \lim_{m\to\infty} \begin{bmatrix} \mathbf{A}^m & \mathbf{0} \\ \mathbf{B}(\mathbf{1}_{N-1} + \mathbf{A} + \cdots + \mathbf{A}^{m-1}) & \mathbf{1}_2 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{B}(\mathbf{1}_{N-1} - \mathbf{A})^{-1} & \mathbf{1}_2 \end{bmatrix}, \end{aligned} \tag{A.9}$$

since $\mathbf{A}^m \to \mathbf{0}$ as $m \to \infty$ and

$$(\mathbf{1}_{N-1} - \mathbf{A})^{-1} = \sum_{j=0}^{\infty} \mathbf{A}^j. \tag{A.10}$$

The fixation probability of a single mutant is given by the element of the matrix $\mathbf{B}(\mathbf{1}_{N-1} - \mathbf{A})^{-1}$ in the second row (corresponding to the final state $n = N$) and the first column (corresponding to the initial state $n = 1$):

$$\phi(r) = \langle 2 | (\mathbf{B}(\mathbf{1}_{N-1} - \mathbf{A})^{-1}) | 1 \rangle. \tag{A.11}$$

Alternatively, this expression can be expanded in powers of $\mathbf{A}$:

$$\phi(r) = \langle 2|\mathbf{B}|1\rangle + \sum_{i=1}^{N-1} \langle 2|\mathbf{B}|i\rangle\langle i|\mathbf{A}|1\rangle + \sum_{i,j=1}^{N-1} \langle 2|\mathbf{B}|i\rangle\langle i|\mathbf{A}|j\rangle\langle j|\mathbf{A}|1\rangle + \cdots . \tag{A.12}$$

Each term in the expansion represents the probability of fixing in a certain finite number of generations: the first term is the probability of fixing in one generation, the second term is the probability of fixing in two generations, etc.

For small population sizes N, Eq. A.11 can be evaluated explicitly:

N	$\phi(r)$
2	$\frac{r^2}{1+r^2}$
3	$\frac{r^3(8r^3+48r^2+6r+1)}{8r^6+48r^5+6r^4+65r^3+6r^2+48r+8}$
$\vdots$	$\vdots$
N	$\frac{r^N a_N(r)}{b_N(r)}$

(A.13)

Empirically we observe that $a_N(r)$ is a degree $N(N-2)$ polynomial and $b_N(r)$ is a degree $N(N-1)$ polynomial. Note that $b_N(r)$ appears to be palindromic: $b_N(r) = r^{N(N-1)} b_N(1/r)$. Unfortunately, the polynomials in these exact expressions grow increasingly intractable with N, making a numerical computation of $\phi(r)$ the only option. Equation A.11 can be rewritten as

$$(\mathbf{1}_{N-1} - \mathbf{A})^T \mathbf{u}^T = \mathbf{B}^T, \tag{A.14}$$

where $\mathbf{u}$ is the $2 \times (N-1)$ matrix of fixation and extinction probabilities from all initial mutant frequencies. The resulting system of linear equations can be efficiently solved to find $\mathbf{u}$ for the arbitrary fitness ratio r. The solution agrees extremely well with explicit simulations (Fig. 2.2).

Appendix B

Supplementary Material on Transcription Factor Binding Site Evolution

B.1 Maximum-likelihood fits of fitness function parameters

For a given TF, let $S = \{\sigma\}$ be the set of binding site sequences, and $\theta = (\beta, \mu, f_0, \nu)$ the parameters of the fitness function (Eq. 3.5). The log-likelihood is given by

$$\log \mathcal{L}(S|\theta) = \sum_{\sigma \in S} \log \pi(\sigma|\theta) = \sum_{\sigma \in S} \log \left(\frac{1}{Z(\theta)} \pi_0(\sigma) (\mathcal{F}(\sigma|\theta))^{\nu} \right), \tag{B.1}$$

where $\mathcal{F}$ is the fitness function, and $Z(\theta) = \sum_{\sigma} \pi_0(\sigma)(\mathcal{F}(\sigma|\theta))^{\nu}$ is the normalization.

Because the log-likelihood function has degenerate or nearly-degenerate regions in the parameter space of θ, instead of maximizing by gradient ascent we obtain a global map of the likelihood by calculating the function over a mesh of points in the parameter domain $\beta \in (0.1, 10)$ generated from $\beta = 10^n$ for n running from -1 to 1 in steps of 0.1, $\mu \in (-20, 0)$ in steps of 0.2, $\nu \in (10^{-3}, 10^5)$ generated from $\nu = 10^n$ for n running from -3 to 5 in steps of 0.1, and $f_0 \in (4.5 \times 10^{-5}, 1 - 4.5 \times 10^{-5})$ generated from $f_0 = (1 + \tanh n)/2$ for n running from -5 to 5 in steps of 0.1. Our predicted maximum is the maximum likelihood point in the mesh, which is sufficiently fine to estimate all fitting parameters. We have made the code for this procedure and for the analysis of site-specific selection available at `www.physics.rutgers.edu/~morozov/publications.html`.

B.2 A model system to check the assumptions of monomorphism and steady state

We consider a haploid asexual Wright-Fisher process [20]. The population consists of $N = 1000$ organisms, each with a single locus of L nucleotides. The new generation is created by means of a selection step and a mutation step. In the selection step, sequences from the current population are sampled with replacement, weighted by their fitness, to construct a new population of size N. In the mutation step, each position in all sequences is mutated with probability u. For simplicity, the mutation rates between all pairs of nucleotides are the same.

We characterize the difference between the distribution expected by our model, π_{exp} (Eq. 2.13), and the distribution observed in simulations, π_{obs}, using the total variation distance (TVD):

$$\Delta(\pi_{\text{exp}}, \pi_{\text{obs}}) = \frac{1}{2}\sum_{x} |\pi_{\text{exp}}(x) - \pi_{\text{obs}}(x)|. \tag{B.2}$$

The TVD ranges from zero for identical distributions to unity for completely non-overlapping distributions. We calculate the TVD for the distributions in energy space, where the sum in Eq. B.2 is over discrete energy bins (we bin the observed sequences by energy by dividing the range from the minimum to the maximum sequence energy for a particular energy matrix into 100 bins of equal size).

We begin by randomly generating the energy matrix parameters $\epsilon_i^{\sigma_i}$. Each $\epsilon_i^{\sigma_i}$ in the energy matrix is sampled from a uniform distribution and then rescaled such that the distribution of all sequence energies has standard deviation of 1.0. This is achieved by dividing all entries in the energy matrix by a factor χ:

$$\chi^2 = \sum_{i=1}^{L} \sum_{\alpha \in \{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}} \pi_0(\alpha)(\epsilon_i^{\alpha} - \bar{\epsilon}_i)^2, \tag{B.3}$$

where ϵ_i^α is the energy matrix element for base α at position i, $L = 10$ is the binding site length, $\bar{\epsilon}_i = \sum_{\alpha \in \{\mathsf{A,C,G,T}\}} \epsilon_i^\alpha$ is the average energy contribution at position i, and $\pi_0(\alpha)$ is the background probability of nucleotide α ($\pi_0(\alpha) = 0.25$ for all α in our simulations). It can be shown that χ is the standard deviation of the random sequence energy distributution, which is approximately Gaussian [68]. We generate the energy matrix once and use it in all subsequent simulations and maximum likelihood fits.

We perform the Wright-Fisher simulations in a range of mutation rates from $u = 10^{-6}$ to $u = 10^{-1}$ with a "non-lethal" Fermi-Dirac fitness function (Eq. 3.5 with $f_0 = 0.99$, $\beta = 1.69$ (kcal/mol)$^{-1}$, and $\mu = -2$ kcal/mol). We run 10^5 simulations for each mutation rate for $100/u + 1000$ steps, enough to reach steady state. Each simulation starts from a monomorphic population with a randomly chosen sequence. We construct the steady state distribution for each mutation rate by randomly choosing a single sequence from the final population of each simulation. Collected across all simulations, these are used to construct a distribution of sequences at each mutation rate. Additionally, we record the average final number of unique sequences at each mutation rate.

We perform another set of Wright-Fisher simulations with the same fitness function and energy matrix as above, and $u = 10^{-6}$. We run 10^5 simulations, each starting from the same monomorphic population with a specific sequence of $E \approx 0$. At regular intervals in each simulation, we record a randomly chosen sequence from the population. Collected across all simulations, these are used to construct a distribution of sequences at each point in time.

B.3 Binding site and energy matrix data

We obtain curated binding site locations for 125 TFs from Ref. 162, which provides a posterior probability that each site is functional based on cross-species analysis. We only consider sites with a posterior probability above 0.9. For this analysis, we use the Saccharomyces Genome Database R53-1-1 (April 2006) build of the *S. cerevisiae* genome.

We obtain position-specific affinity matrices (PSAMs) for a set of 26 TFs from an *in* *vitro* microfluidics analysis of TF-DNA interactions [166]. This study provides PSAMs for each TF determined using the MatrixREDUCE package [165]. We convert the elements of the PSAM $w_{i\alpha}$ to energy matrix elements using $\epsilon_{i\alpha} = -\log(w_{i\alpha})/\beta$, where $\beta = 1.69$ (kcal/mol)$^{-1}$ at room temperature. For each of these 26 TFs, genomic sites are available in Ref. 162. We neglect PHO4 since it does not have any binding sites above the 0.9 threshold in Ref. 162, leaving us with 25 TFs for which both an energy matrix and a set of genomic binding sites are available. We align the binding site sequences from Ref. 162 to the corresponding energy matrices, choosing the alignment that produces the lowest average binding energy for the sites.

B.4 Essentiality data

The Yeast Deletion Database classifies genes as essential, tested (nonessential), and unavailable, which number 1156, 6343, and 529, respectively [167, 175]. For each essential or tested gene, we determine all TF binding sites less than 700 bp upstream of the gene's transcription start site (on either strand), which we designate as the sites regulating that gene. Growth rates for nonessential knockout strains are provided under YPD, YPDGE, YPG, YPE, and YPL conditions, relative to wild-type. We choose the lowest of these growth rates to represent the fitness effect of the knockout.

To measure the rate of nonsynonymous substitutions, we align the non-mitochondrial, non-retrotransposon ORFs taken from the Saccharomyces Genome Database R64-1-1 (February 2011) build [275] of *S. cerevisiae* to those of *S. paradoxus* using ClustalW [278]. We measure the rate of nonsynonymous mutations using PAML [279]. We ran PAML with a runMode of -2 (pairwise comparisons) and the CodonFreq parameter (background codon frequency) set to 2; we also tested CodonFreq set to zero and obtained very similar results. We find the rate of nonsynonymous substitutions to be 0.04, and a Spearman rank correlation of -0.16 ($p = 10^{-27}$) between growth rate of knockouts and the nonsynonymous substitution rate of the knocked-out gene. This is consistent with the results of Ref. 182, which found the rate of substitutions to be 0.04 and the rank correlation between growth rate and substitution rate to be -0.19 ($p = 10^{-35}$).

To compare binding energy to evolutionary conservation, we calculate the mean Hamming distance between *S. cerevisiae* sites and corresponding sites in *S. paradoxus* [162]. To test for significance in the difference of mean energies and Hamming distances of sites regulating essential and nonessential genes, we use a null model which assumes that the sites were randomly categorized into essential and nonessential. We randomly choose a subset of the sites in our dataset to be "nonessential," equal in size to the number of sites regulating nonessential genes as classified by the Yeast Deletion Database. By repeating this procedure 10^7 times, we build a probability distribution for the difference in the means of the nonessential and essential groups. The p-value is the probability of obtaining a difference in the means greater than the empirically measured value.

B.5 Neutral binding site energy distributions

We construct the neutral probability $\pi_0(\sigma)$ of a sequence σ with length L as

$$\pi_0(\sigma) = \pi_0(\sigma_1) \prod_{i=2}^{L} \pi_0(\sigma_{i-1}, \sigma_i), \tag{B.4}$$

where $\pi_0(\sigma_i)$ is the background probability of a nucleotide σ_i, and $\pi_0(\sigma_{i-1}, \sigma_i)$ is the background probability of a dinucleotide $\sigma_{i-1}\sigma_i$. These probabilities are determined from mono- and dinucleotide frequencies in the intergenic regions of the *S. cerevisiae* genome (build R61-1-1, June 2008). We project $\pi_0(\sigma)$ into energy space using Eq. 3.2 to obtain $\pi_0(E)$, the neutral distribution of binding energies for sequences of length L.

If intergenic sequences evolve under no selection with respect to their TF-binding energy, the neutral distribution of site energies should closely match the actual distribution of L-mer sequences obtained from intergenic regions. Table B.2, column B shows that these two distributions match very well except at the low-energy tail, which is enriched in functional binding sites. Note that accounting for dinucleotide frequencies is important; mononucleotide frequencies alone are insufficient to reproduce the observed distribution [178].

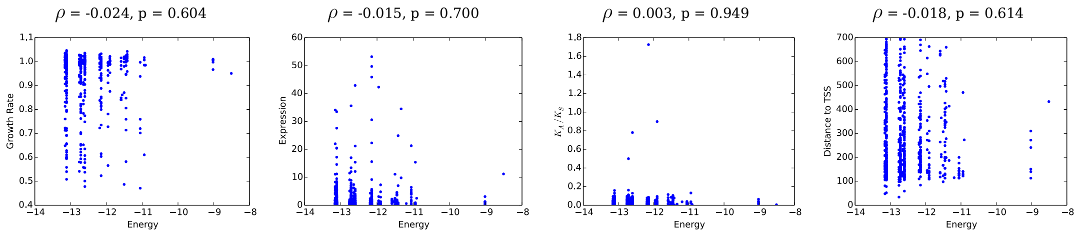
B.6 Supplementary tables

(A) Dataset **(B) Growth Rate** **(C) Expression Level** **(D) K_A/K_S Ratio** **(E) TSS Distance**

REB1 (essential TF)

Total sites: 749
Unique sites: 235

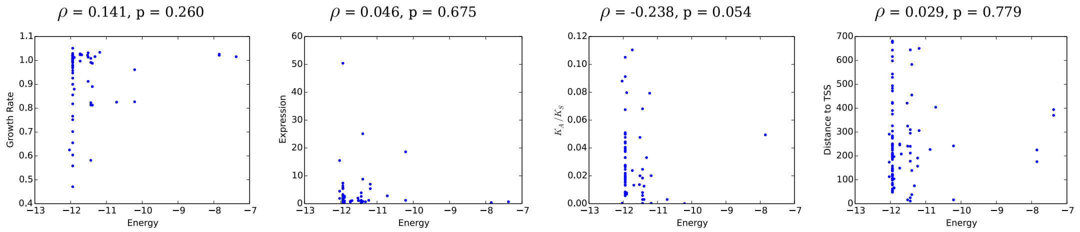
	Essential	**Noness.**	**p**
Total Data	213	468	
Expr Data	209	433	
S. Par. Data	194	358	
$\langle E \rangle$	-12.632	-12.54	0.090
V	0.336	0.471	
$\langle \Delta E^2 \rangle$	0.103	0.206	0.152
$\langle d \rangle$	0.847	0.904	0.484



ROX1 (nonessential TF)

Total sites: 93
Unique sites: 58

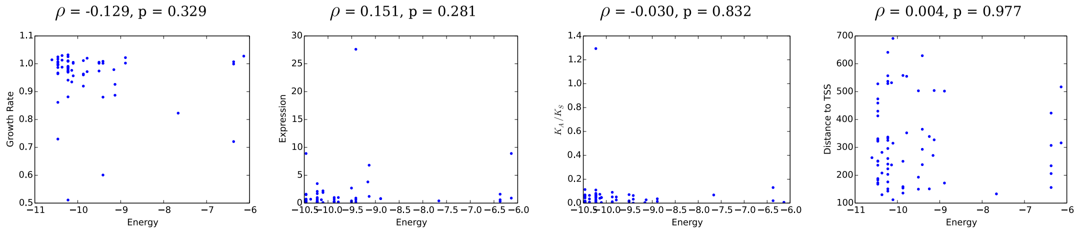
	Essential	**Noness.**	**p**
Total Data	18	65	
Expr Data	18	64	
S. Par. Data	18	47	
$\langle E \rangle$	-11.683	-11.575	0.681
V	0.205	0.833	
$\langle \Delta E^2 \rangle$	0.015	0.697	0.273
$\langle d \rangle$	0.824	0.737	0.679



MET32 (nonessential TF)

Total sites: 68
Unique sites: 39

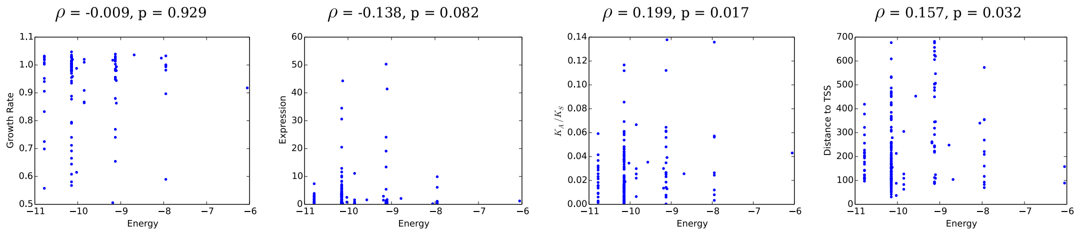
	Essential	**Noness.**	**p**
Total Data	5	59	
Expr Data	4	49	
S. Par. Data	4	48	
$\langle E \rangle$	-8.66	-9.738	0.053
V	3.917	1.136	
$\langle \Delta E^2 \rangle$	0.0	0.006	0.369
$\langle d \rangle$	0.25	0.377	0.627



RPN4 (nonessential TF)

Total sites: 188
Unique sites: 38

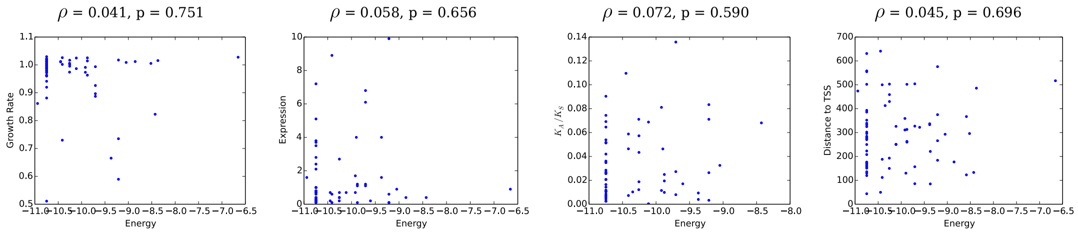
	Essential	**Noness.**	**p**
Total Data	71	101	
Expr Data	71	89	
S. Par. Data	66	77	
$\langle E \rangle$	-10.003	-9.776	0.048
V	0.304	0.705	
$\langle \Delta E^2 \rangle$	0.084	0.126	0.705
$\langle d \rangle$	0.167	0.231	0.342



MET31 (nonessential TF)

Total sites: 77
Unique sites: 35

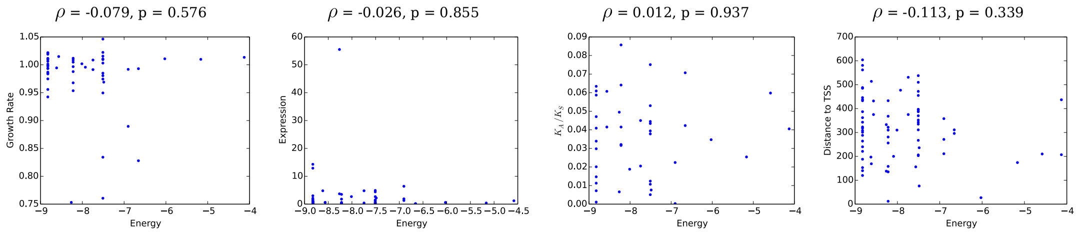
	Essential	**Noness.**	**p**
Total Data	10	60	
Expr Data	9	53	
S. Par. Data	10	48	
$\langle E \rangle$	-9.968	-10.201	0.390
V	0.328	0.649	
$\langle \Delta E^2 \rangle$	0.0	0.029	0.299
$\langle d \rangle$	0.0	0.151	0.084



PDR3 (nonessential TF)

Total sites: 73
Unique sites: 31

	Essential	**Noness.**	**p**
Total Data	6	51	
Expr Data	6	47	
S. Par. Data	5	36	
$\langle E \rangle$	-7.34	-7.942	0.174
V	1.826	0.968	
$\langle \Delta E^2 \rangle$	3.3	0.121	0.011
$\langle d \rangle$	1.0	0.304	0.009

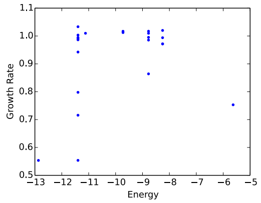


YAP7 (nonessential TF)

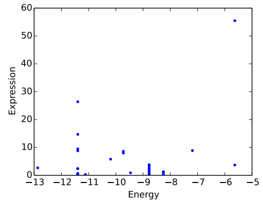
Total sites: 36
Unique sites: 22

	Essential	**Noness.**	**p**
Total Data	13	23	
Expr Data	13	23	
S. Par. Data	11	18	
$\langle E \rangle$	-9.076	-9.936	0.157
V	2.575	2.865	
$\langle \Delta E^2 \rangle$	1.718	1.073	0.714
$\langle d \rangle$	0.583	0.182	0.042

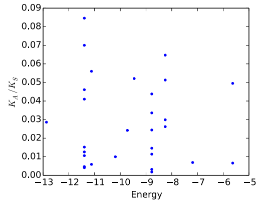
ρ = 0.200, p = 0.361



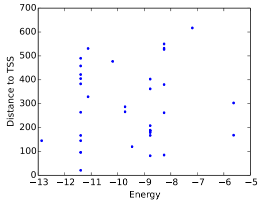
ρ = -0.015, p = 0.933



ρ = -0.012, p = 0.949



ρ = 0.187, p = 0.274

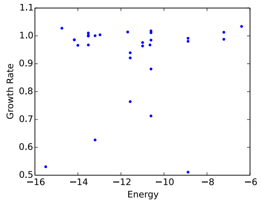


BAS1 (nonessential TF)

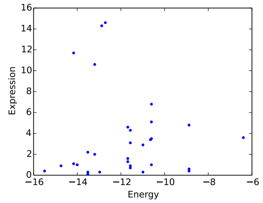
Total sites: 41
Unique sites: 21

	Essential	**Noness.**	**p**
Total Data	5	30	
Expr Data	5	26	
S. Par. Data	5	28	
$\langle E \rangle$	-12.107	-11.518	0.588
V	0.322	5.328	
$\langle \Delta E^2 \rangle$	0.128	0.6	0.935
$\langle d \rangle$	0.2	0.214	0.741

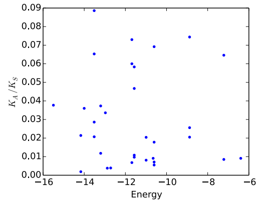
ρ = 0.058, p = 0.760



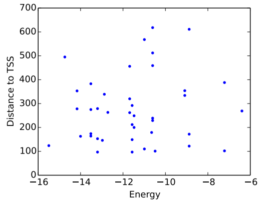
ρ = 0.166, p = 0.371



ρ = -0.109, p = 0.546



ρ = 0.078, p = 0.626

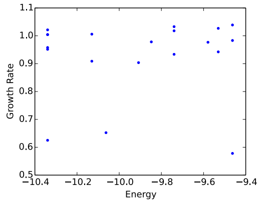


STB5 (nonessential TF)

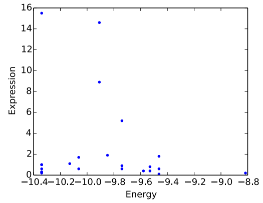
Total sites: 28
Unique sites: 19

	Essential	**Noness.**	**p**
Total Data	5	20	
Expr Data	5	18	
S. Par. Data	5	14	
$\langle E \rangle$	-9.893	-9.918	0.912
V	0.317	0.116	
$\langle \Delta E^2 \rangle$	0.002	0.0	0.000
$\langle d \rangle$	0.4	0.222	0.400

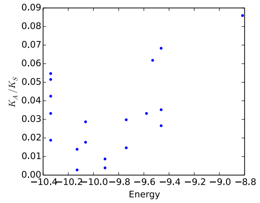
ρ = 0.116, p = 0.625



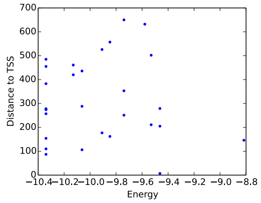
ρ = -0.154, p = 0.484



ρ = 0.232, p = 0.339



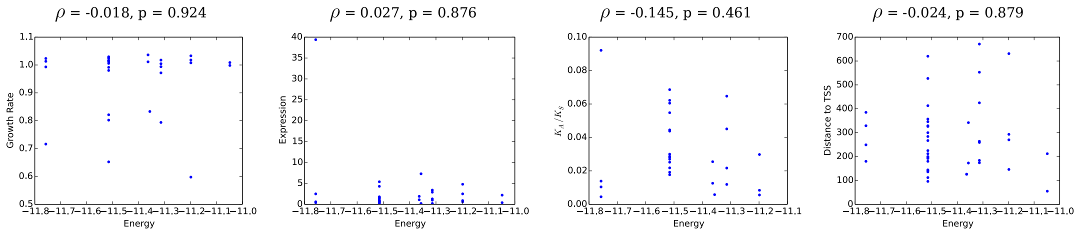
ρ = 0.004, p = 0.983



AFT1 (nonessential TF)

Total sites: 42
Unique sites: 18

	Essential	**Noness.**	**p**
Total Data	5	30	
Expr Data	5	29	
S. Par. Data	5	23	
$\langle E \rangle$	-11.475	-11.425	0.581
V	0.006	0.037	
$\langle \Delta E^2 \rangle$	0.0	0.018	0.445
$\langle d \rangle$	0.4	0.391	0.714


ρ = -0.018, p = 0.924

ρ = 0.027, p = 0.876

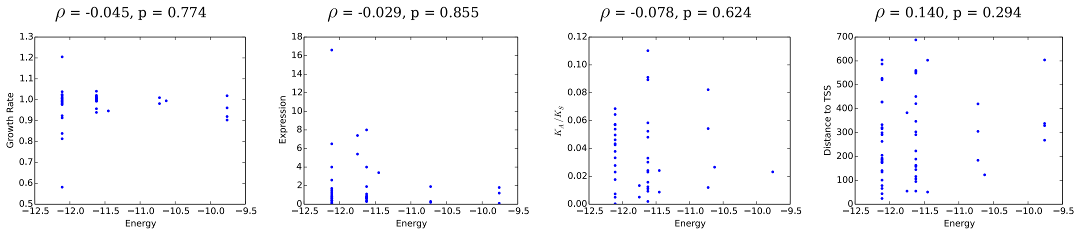
ρ = -0.145, p = 0.461

ρ = -0.024, p = 0.879

CUP9 (nonessential TF)

Total sites: 58
Unique sites: 13

	Essential	**Noness.**	**p**
Total Data	11	43	
Expr Data	11	32	
S. Par. Data	11	31	
$\langle E \rangle$	-11.681	-11.607	0.752
V	0.141	0.494	
$\langle \Delta E^2 \rangle$	0.024	0.109	0.797
$\langle d \rangle$	0.1	0.139	0.656


ρ = -0.045, p = 0.774

ρ = -0.029, p = 0.855

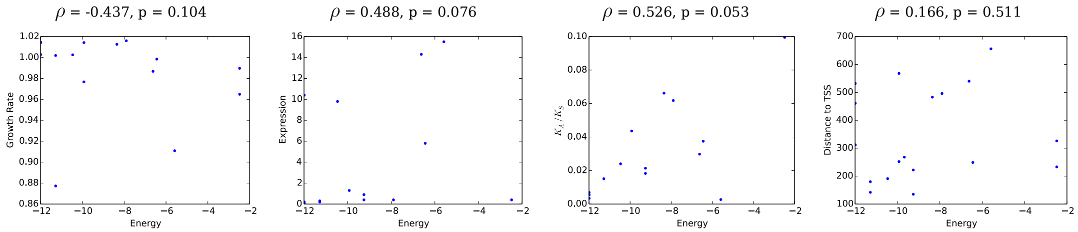
ρ = -0.078, p = 0.624

ρ = 0.140, p = 0.294

MCM1 (essential TF)

Total sites: 18
Unique sites: 13

	Essential	**Noness.**	**p**
Total Data	2	15	
Expr Data	2	12	
S. Par. Data	2	12	
$\langle E \rangle$	-9.252	-8.58	0.795
V	0.0	9.927	
$\langle \Delta E^2 \rangle$	0.013	1.36	0.845
$\langle d \rangle$	2.0	0.8	0.037


ρ = -0.437, p = 0.104

ρ = 0.488, p = 0.076

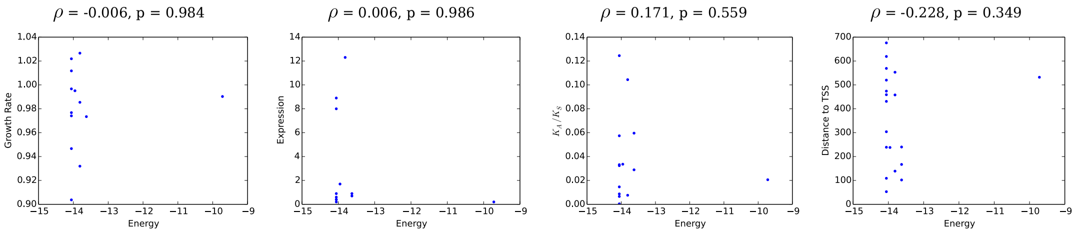
ρ = 0.526, p = 0.053

ρ = 0.166, p = 0.511

CIN5 (nonessential TF)

Total sites: 19
Unique sites: 12

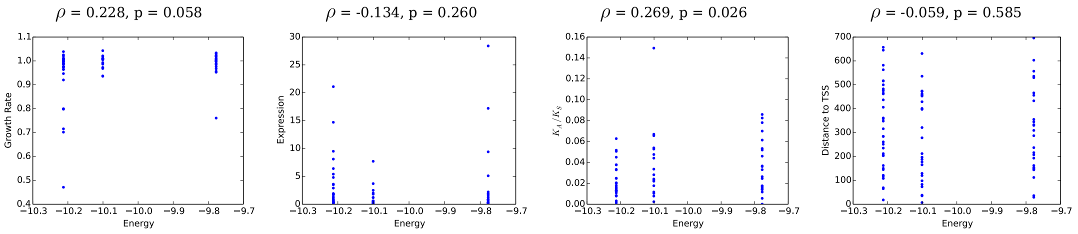
	Essential	**Noness.**	**p**
Total Data	2	13	
Expr Data	2	9	
S. Par. Data	2	11	
$\langle E \rangle$	-13.841	-13.625	0.835
V	0.046	1.29	
$\langle \Delta E^2 \rangle$	0.03	0.036	1.000
$\langle d \rangle$	1.0	0.417	0.228



GAT1 (nonessential TF)

Total sites: 88
Unique sites: 11

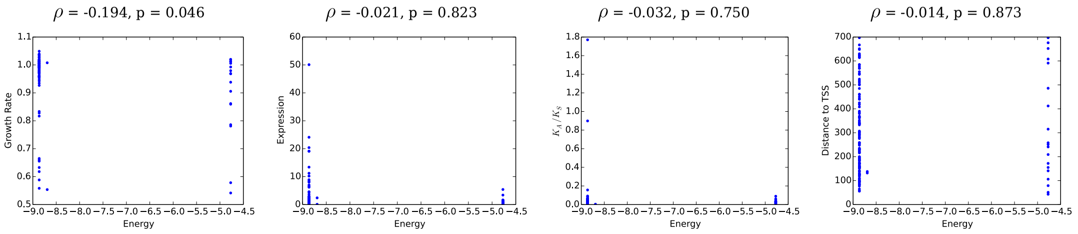
	Essential	**Noness.**	**p**
Total Data	8	70	
Expr Data	8	63	
S. Par. Data	7	60	
$\langle E \rangle$	-10.036	-10.046	0.866
V	0.041	0.035	
$\langle \Delta E^2 \rangle$	0.027	0.017	0.933
$\langle d \rangle$	0.429	0.353	0.731



MSN2 (nonessential TF)

Total sites: 141
Unique sites: 8

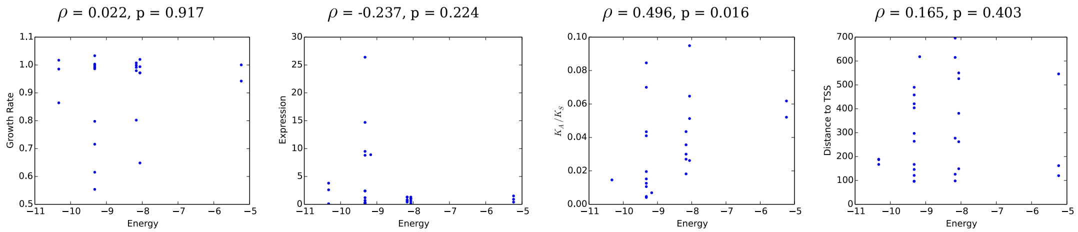
	Essential	**Noness.**	**p**
Total Data	19	106	
Expr Data	18	96	
S. Par. Data	15	86	
$\langle E \rangle$	-8.433	-8.204	0.653
V	1.577	2.251	
$\langle \Delta E^2 \rangle$	0.985	0.664	0.726
$\langle d \rangle$	0.059	0.079	0.614



CAD1 (nonessential TF)

Total sites: 28
Unique sites: 8

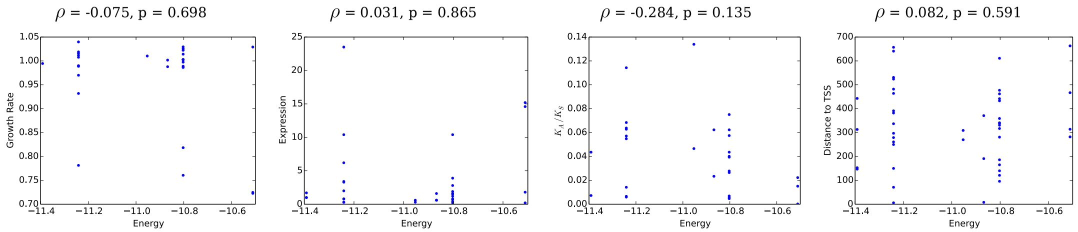
	Essential	**Noness.**	**p**
Total Data	3	25	
Expr Data	3	25	
S. Par. Data	3	20	
$\langle E \rangle$	-7.91	-8.635	0.443
V	3.585	1.571	
$\langle \Delta E^2 \rangle$	0.0	0.301	0.144
$\langle d \rangle$	0.0	0.2	0.215



ACE2 (nonessential TF)

Total sites: 45
Unique sites: 6

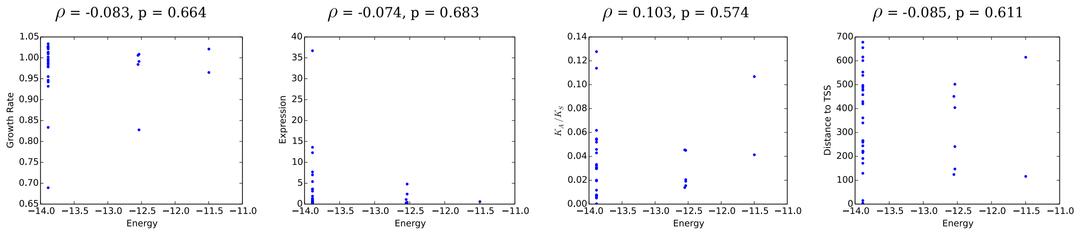
	Essential	**Noness.**	**p**
Total Data	7	29	
Expr Data	7	26	
S. Par. Data	6	23	
$\langle E \rangle$	-11.023	-10.954	0.554
V	0.094	0.065	
$\langle \Delta E^2 \rangle$	0.0	0.0	0.000
$\langle d \rangle$	0.0	0.0	0.000



YAP3 (nonessential TF)

Total sites: 38
Unique sites: 6

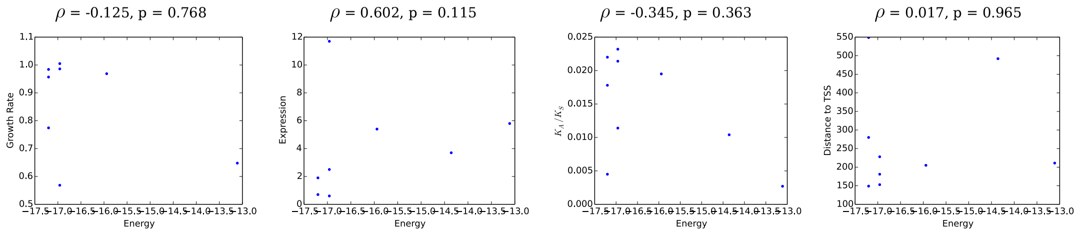
	Essential	**Noness.**	**p**
Total Data	8	30	
Expr Data	8	25	
S. Par. Data	8	24	
$\langle E \rangle$	-13.718	-13.503	0.367
V	0.199	0.535	
$\langle \Delta E^2 \rangle$	0.0	0.308	0.245
$\langle d \rangle$	0.0	0.276	0.030



GCN4 (nonessential TF)

Total sites: 9
Unique sites: 5

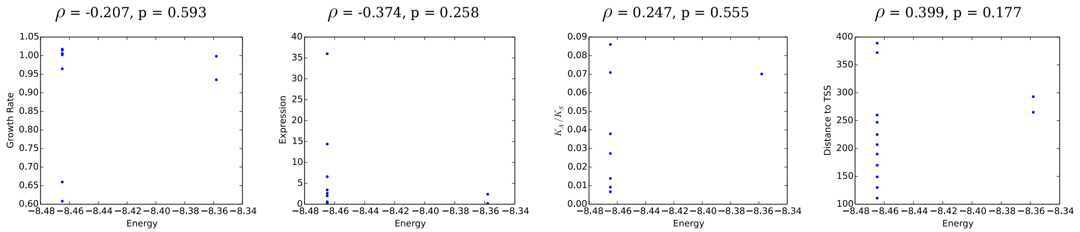
	Essential	Noness.	p
Total Data	1	8	
Expr Data	1	7	
S. Par. Data	1	8	
$\langle E \rangle$	-14.357	-16.442	0.183
V	0.0	1.736	
$\langle \Delta E^2 \rangle$	1.588	0.025	0.000
$\langle d \rangle$	2.0	0.429	0.000



MATA2 (nonessential TF)

Total sites: 13
Unique sites: 4

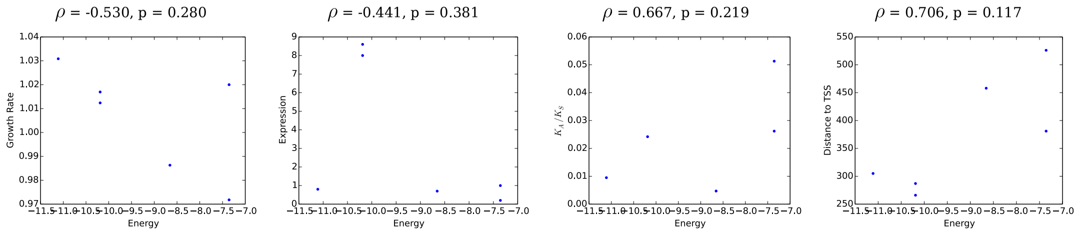
	Essential	Noness.	p
Total Data	1	9	
Expr Data	1	9	
S. Par. Data	1	6	
$\langle E \rangle$	-8.465	-8.441	0.866
V	0.0	0.002	
$\langle \Delta E^2 \rangle$	0.011	0.006	1.000
$\langle d \rangle$	1.0	0.5	0.443



YAP1 (nonessential TF)

Total sites: 6
Unique sites: 4

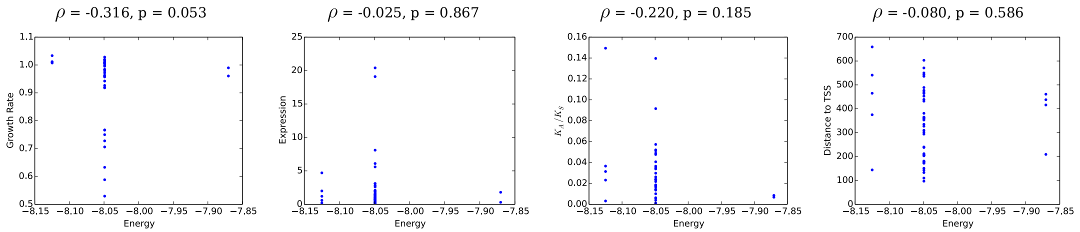
	Essential	Noness.	p
Total Data	0	6	
Expr Data	0	6	
S. Par. Data	0	5	
$\langle E \rangle$	nan	nan	0.000
V	nan	2.114	
$\langle \Delta E^2 \rangle$	nan	nan	0.000
$\langle d \rangle$	nan	nan	0.000



CBF1 (nonessential TF)

Total sites: 49
Unique sites: 3

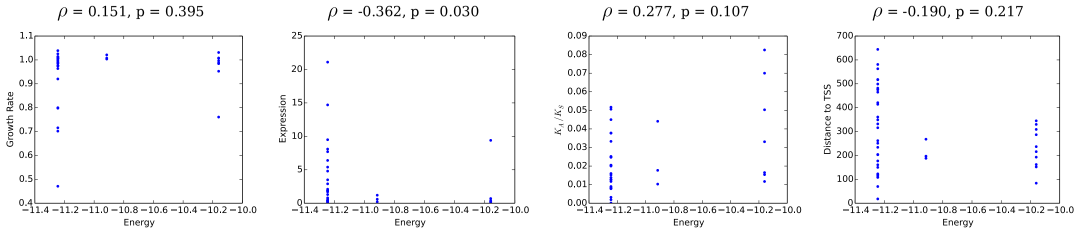
	Essential	**Noness.**	**p**
Total Data	7	38	
Expr Data	7	38	
S. Par. Data	6	31	
$\langle E \rangle$	-8.06	-8.048	0.380
V	0.001	0.002	
$\langle \Delta E^2 \rangle$	0.009	0.007	0.783
$\langle d \rangle$	0.143	0.139	0.569



DAL80 (nonessential TF)

Total sites: 44
Unique sites: 3

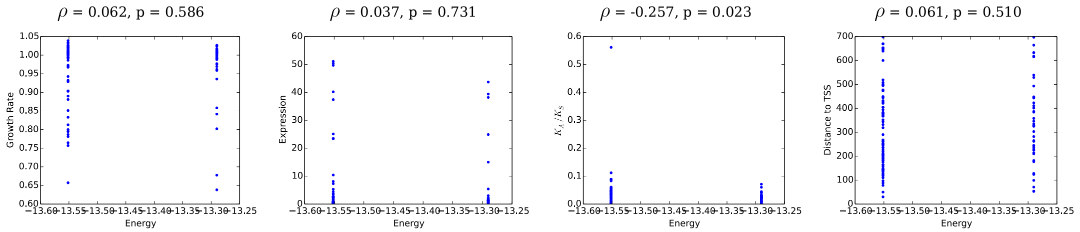
	Essential	**Noness.**	**p**
Total Data	4	34	
Expr Data	4	31	
S. Par. Data	4	30	
$\langle E \rangle$	-11.245	-10.961	0.425
V	0.0	0.205	
$\langle \Delta E^2 \rangle$	0.0	0.224	0.137
$\langle d \rangle$	0.0	0.303	0.182



AFT2 (nonessential TF)

Total sites: 118
Unique sites: 2

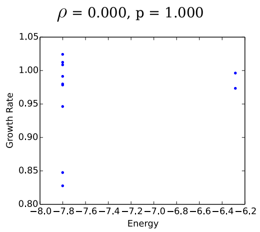
	Essential	**Noness.**	**p**
Total Data	17	80	
Expr Data	17	71	
S. Par. Data	15	62	
$\langle E \rangle$	-13.505	-13.447	0.052
V	0.01	0.016	
$\langle \Delta E^2 \rangle$	0.0	0.006	0.088
$\langle d \rangle$	0.0	0.087	0.088

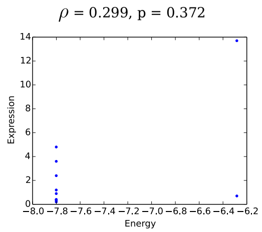


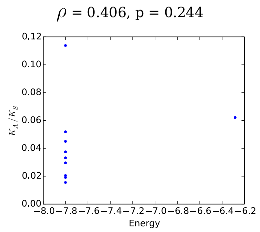
SKO1 (essential TF)

Total sites: 12
Unique sites: 2

	Essential	**Noness.**	**p**
Total Data	1	11	
Expr Data	1	10	
S. Par. Data	1	9	
$\langle E \rangle$	-7.801	-7.525	1.000
V	0.0	0.343	
$\langle \Delta E^2 \rangle$	0.0	0.0	0.000
$\langle d \rangle$	0.0	0.0	0.000


ρ = 0.000, p = 1.000


ρ = 0.299, p = 0.372


ρ = 0.406, p = 0.244

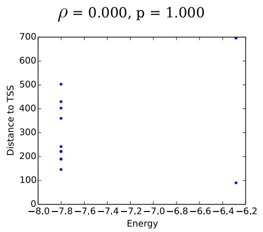
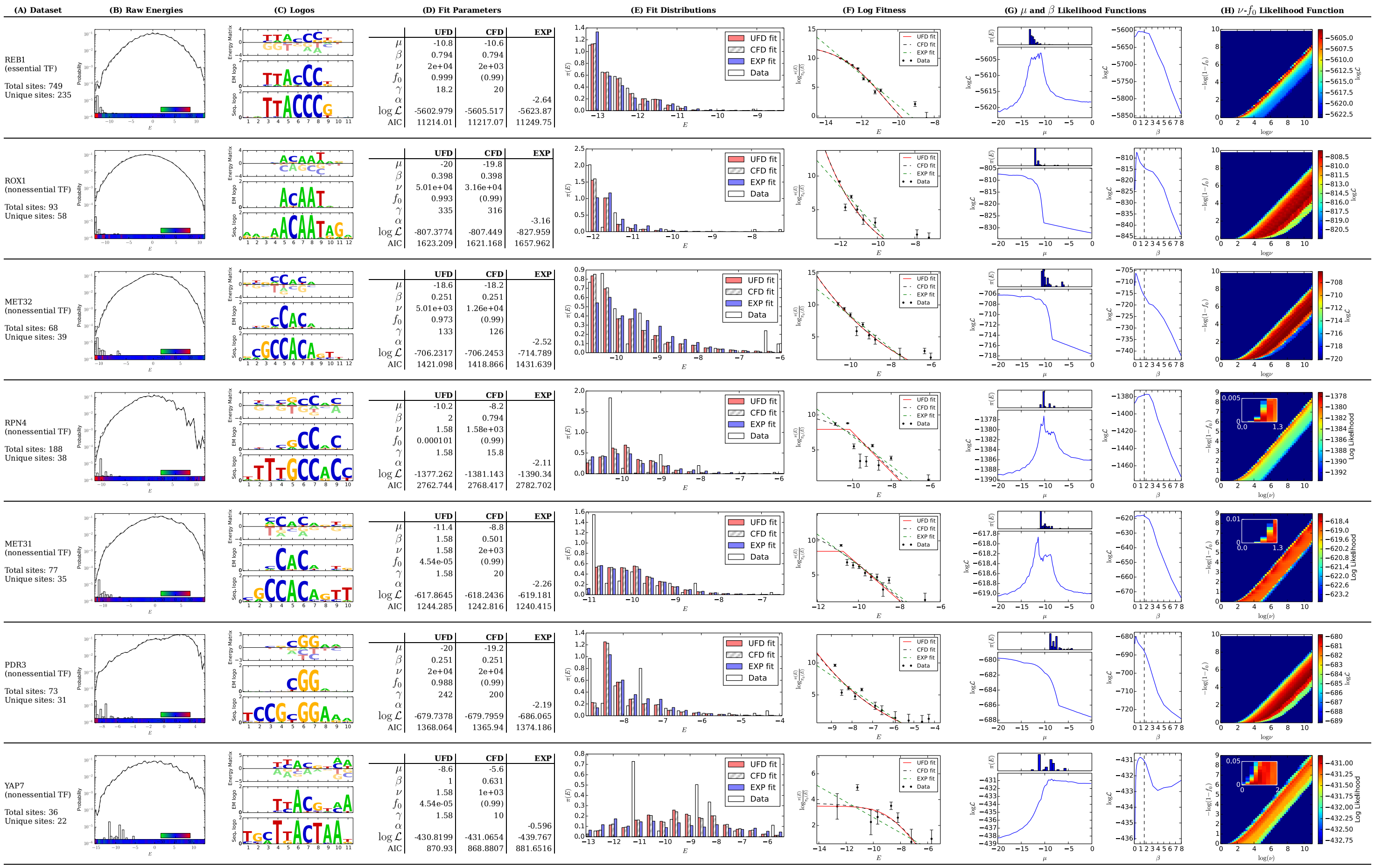

ρ = 0.000, p = 1.000

Table B.1: **Full summary of tests for site-specific selection.** For 25 TFs we compute TF-DNA interaction energies (in kcal/mol) for each site. Columns from left to right: (A) Essentiality of the TF according to the Yeast Deletion Database; total number of binding sites for each TF; total number of sites with unique sequences. The table lists how many essential and nonessential genes are regulated by each TF, and how many of these genes have gene expression and *S. paradoxus* K_A/K_S ratio data. We also report the mean energy $\bar{E}$ and the variance V of essential and nonessential sites, and mean Hamming distance $\bar{d}$ and mean squared energy difference between *S. cerevisiae* and *S. paradoxus* sites regulating essential and nonessential genes. (B) Growth rate in strains with nonessential gene knockouts versus energy of TF binding sites regulating the knockout genes. (C) Gene expression versus energy of TF sites regulating the genes. (D) Ratio of nonsynonymous to synonymous substitutions (K_A/K_S) in genes versus energy of their TF regulatory sites. (E) Distance between each binding site and the closest transcription start site (TSS) versus the energy of the site. For (B)–(E) we report the Spearman rank correlation ρ between each property and site energy, along with the p-value.



(A) Dataset

(B) Raw Energies

(C) Logos

(D) Fit Parameters

(E) Fit Distributions

(F) Log Fitness

(G) μ and β Likelihood Functions

(H) ν-f_0 Likelihood Function

REB1
(essential TF)

Total sites: 749
Unique sites: 235

	UFD	CFD	EXP
μ	-10.8	-10.6	
β	0.794	0.794	
ν	2e+04	2e+03	
f_0	0.999	(0.99)	
γ	18.2	20	
α			-2.64
$\log\mathcal{L}$	-5602.979	-5605.517	-5623.87
AIC	11214.01	11217.07	11249.75

ROX1
(nonessential TF)

Total sites: 93
Unique sites: 58

	UFD	CFD	EXP
μ	-20	-19.8	
β	0.398	0.398	
ν	5.01e+04	3.16e+04	
f_0	0.993	(0.99)	
γ	335	316	
α			-3.16
$\log\mathcal{L}$	-807.3774	-807.449	-827.959
AIC	1623.209	1621.168	1657.962

MET32
(nonessential TF)

Total sites: 68
Unique sites: 39

	UFD	CFD	EXP
μ	-18.6	-18.2	
β	0.251	0.251	
ν	5.01e+03	1.26e+04	
f_0	0.973	(0.99)	
γ	133	126	
α			-2.52
$\log\mathcal{L}$	-706.2317	-706.2453	-714.789
AIC	1421.098	1418.866	1431.639

RPN4
(nonessential TF)

Total sites: 188
Unique sites: 38

	UFD	CFD	EXP
μ	-10.2	-8.2	
β	2	0.794	
ν	1.58	1.58e+03	
f_0	0.000101	(0.99)	
γ	1.58	15.8	
α			-2.11
$\log\mathcal{L}$	-1377.262	-1381.143	-1390.34
AIC	2762.744	2768.417	2782.702

MET31
(nonessential TF)

Total sites: 77
Unique sites: 35

	UFD	CFD	EXP
μ	-11.4	-8.8	
β	1.58	0.501	
ν	1.58	2e+03	
f_0	4.54e-05	(0.99)	
γ	1.58	20	
α			-2.26
$\log\mathcal{L}$	-617.8645	-618.2436	-619.181
AIC	1244.285	1242.816	1240.415

PDR3
(nonessential TF)

Total sites: 73
Unique sites: 31

	UFD	CFD	EXP
μ	-20	-19.2	
β	0.251	0.251	
ν	2e+04	2e+04	
f_0	0.988	(0.99)	
γ	242	200	
α			-2.19
$\log\mathcal{L}$	-679.7378	-679.7959	-686.065
AIC	1368.064	1365.94	1374.186

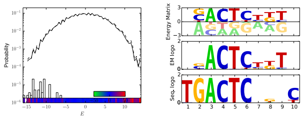
YAP7
(nonessential TF)

Total sites: 36
Unique sites: 22

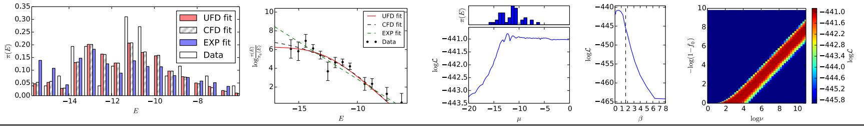
	UFD	CFD	EXP
μ	-8.6	-5.6	
β	1	0.631	
ν	1.58	1e+03	
f_0	4.54e-05	(0.99)	
γ	1.58	10	
α			-0.596
$\log\mathcal{L}$	-430.8199	-431.0654	-439.767
AIC	870.93	868.8807	881.6516

BAS1
(nonessential TF)

Total sites: 41
Unique sites: 21

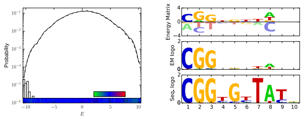


	UFD	**CFD**	**EXP**
μ	-12.2	-6.6	
β	0.501	0.316	
ν	2.51	1.58e+03	
f_0	4.54e-05	(0.99)	
γ	2.51	15.8	
α			-0.793
$\log \mathcal{L}$	-440.7891	-440.9669	-444.824
AIC	890.6893	888.5825	891.7506

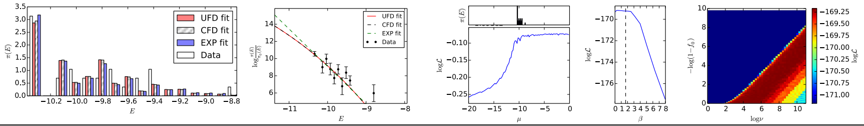


STB5
(nonessential TF)

Total sites: 28
Unique sites: 19

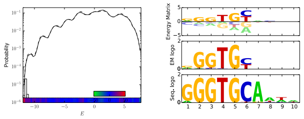


	UFD	**CFD**	**EXP**
μ	-2.6	-1.8	
β	0.316	0.316	
ν	251	2e+04	
f_0	0.401	(0.99)	
γ	150	200	
α			-4.36
$\log \mathcal{L}$	-169.1727	-169.1763	-169.393
AIC	348.0845	345.3525	340.9398

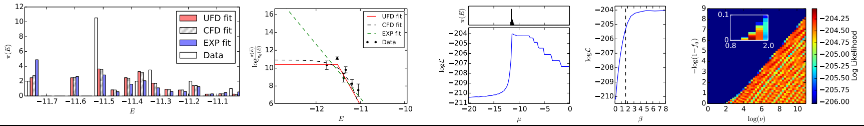


AFT1
(nonessential TF)

Total sites: 42
Unique sites: 18

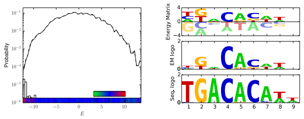


	UFD	**CFD**	**EXP**
μ	-11.2	-9.6	
β	5.01	3.16	
ν	3.98	5.01e+04	
f_0	4.54e-05	(0.99)	
γ	3.98	501	
α			-6.02
$\log \mathcal{L}$	-204.0313	-204.2217	-210.57
AIC	417.1437	415.075	423.24

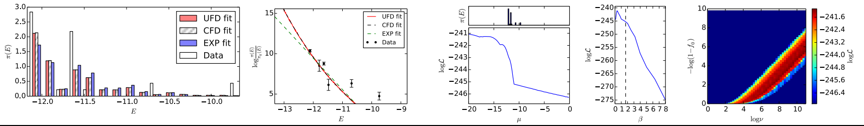


CUP9
(nonessential TF)

Total sites: 58
Unique sites: 13

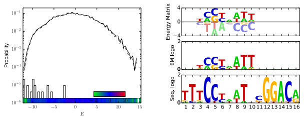


	UFD	**CFD**	**EXP**
μ	-20	-19.6	
β	0.316	0.316	
ν	1e+04	2e+04	
f_0	0.978	(0.99)	
γ	219	200	
α			-3.89
$\log \mathcal{L}$	-241.0682	-241.0979	-245.19
AIC	490.8911	488.6403	492.4514



MCM1
(essential TF)

Total sites: 18
Unique sites: 13



	UFD	**CFD**	**EXP**
μ	-20	-19.8	
β	0.251	0.251	
ν	5.01e+04	7.94e+03	
f_0	0.998	(0.99)	
γ	83.1	79.4	
α			-0.861
$\log \mathcal{L}$	-348.8152	-348.8252	-353.141
AIC	708.7073	705.3646	708.532

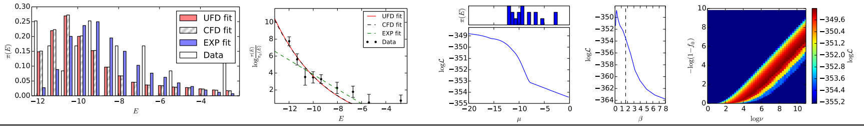
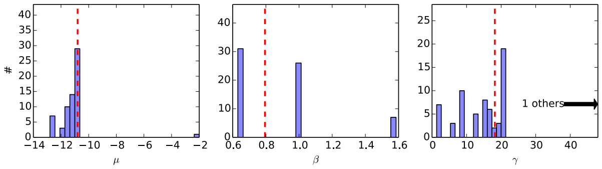


Table B.2: **Summary of fitness landscape fits to TF binding site data.** We consider 12 TFs which have more than 12 unique binding site sequences. Each row corresponds to a TF, ranked in the decreasing order of the number of unique binding site sequences. Columns, from left to right: (A) Summary of TF binding site data. (B) Same as Fig. 5A. (C) Same as Fig. 5B. (D) Fitted values of fitness landscape parameters and maximized log-likelihoods for the unconstrained fit to the Fermi-Dirac function of Eq. 6 ("UFD"), constrained fit to the Eq. 6 with $f_0 = 0.99$ ("CFD"), and fit to an exponential fitness function ("EXP"). (E) Same as Fig. 5C. (F) Same as Fig. 5D. (G) Left panel: Log-likelihood of the unconstrained Fermi-Dirac model as a function of the effective chemical potential μ. For reference, the distribution of functional binding site energies (same as in (B)) is shown on top. Right panel: Log-likelihood as a function of the effective inverse temperature β. For reference, the inverse room temperature 1.69 $(\text{kcal/mol})^{-1}$ is shown as the vertical dashed line. To generate the log-likelihood plots, μ or β were scanned across a range of values while all the other parameters were re-optimized for each new value of μ or β. (H) Heatmap of log-likelihood as a function of $\log \nu$ and $-\log(1 - f_0)$ (note that $\nu(1 - f_0) = \gamma =$ constant corresponds to a straight line with slope 1 in these coordinates). For likelihoods that have a maximum near $f_0 = 0$, insets show a zoomed-in view. To generate the log-likelihood heatmaps, ν and f_0 were scanned across the region shown while the other parameters (β and μ) were re-optimized at each point separately.

(A) Hessian Eigenvectors **(B) Subsample Fits**

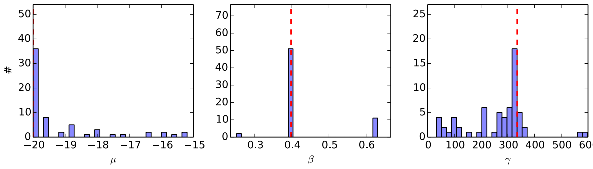
REB1

λ	$\log\nu$	μ	$\log\beta$	$\mathcal{T}(f_0)$
-0.0201	0.894	0.001	-0	0.447
-102	-0.097	-0.81	0.545	0.195
-179	0.062	-0.578	-0.804	-0.122
-1.29e+04	-0.432	0.101	-0.237	0.864



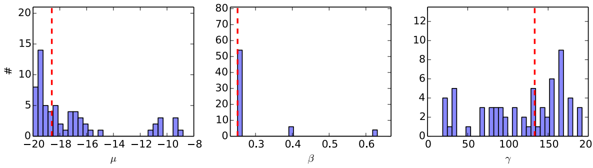
ROX1

λ	$\log\nu$	μ	$\log\beta$	$\mathcal{T}(f_0)$
0.00498	0.891	0.04	0	0.452
-0.0333	-0.1	0.989	0.008	0.111
-34.8	0.42	0.134	0.321	-0.838
-2.32e+04	0.141	0.054	-0.947	-0.283



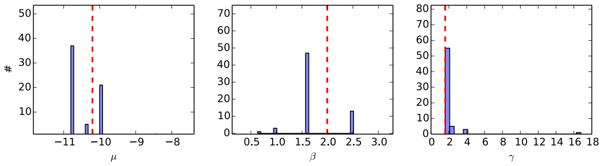
MET32

λ	$\log\nu$	μ	$\log\beta$	$\mathcal{T}(f_0)$
0.00356	0.885	0.1	0	0.454
-0.00198	-0.126	0.992	0.006	0.027
-169	0.403	0.071	0.436	-0.802
-1.46e+04	0.194	0.041	-0.9	-0.388



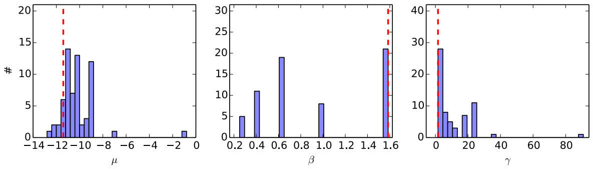
RPN4

λ	$\log\nu$	μ	$\log\beta$	$\mathcal{T}(f_0)$
0.938	-0.235	-0.007	0.42	-0.877
-7.09	0.265	-0.273	-0.805	-0.454
-58.9	0.371	0.91	-0.1	-0.155
-705	0.859	-0.311	0.406	-0.033



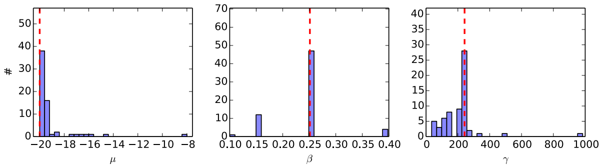
MET31

λ	$\log\nu$	μ	$\log\beta$	$\mathcal{T}(f_0)$
-0.375	-0.186	0.052	0.254	-0.948
-3.23	0.132	-0.893	-0.39	-0.179
-26.5	-0.501	-0.433	0.702	0.263
-358	0.835	-0.107	0.539	-0.025



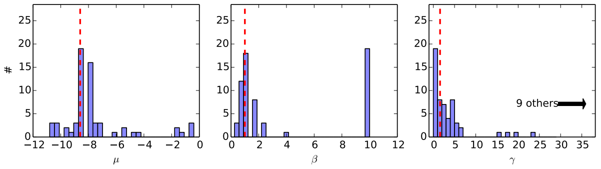
PDR3

λ	$\log\nu$	μ	$\log\beta$	$\mathcal{T}(f_0)$
0.00138	0.889	0.066	-0	0.453
-0.0051	-0.107	0.992	-0.001	0.066
-20.7	0.43	0.104	0.262	-0.858
-2.78e+04	0.117	0.027	-0.965	-0.233



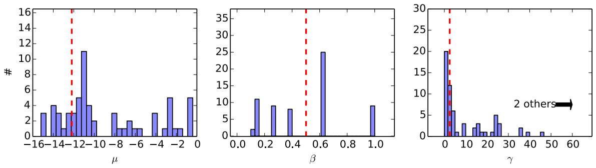
YAP7

λ	$\log\nu$	μ	$\log\beta$	$\mathcal{T}(f_0)$
-0.000577	0.048	0.045	-0.033	0.997
-0.0749	-0.65	-0.602	0.459	0.074
-2.54	-0.282	0.755	0.592	-0.001
-80.3	0.705	-0.256	0.662	-0.001



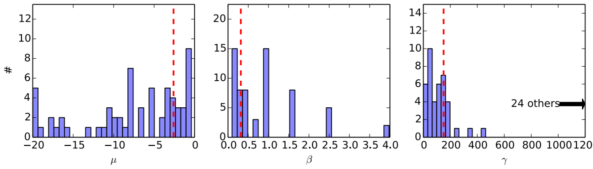
BAS1

λ	$\log\nu$	μ	$\log\beta$	$\mathcal{T}(f_0)$
-0.00334	0.01	0.017	-0.003	1
-0.215	-0.443	-0.886	0.138	0.02
-2.55	0.804	-0.46	-0.377	-0.001
-708	-0.397	0.056	-0.916	0



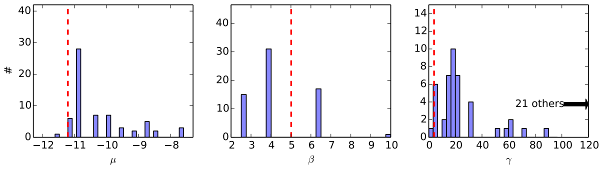
STB5

λ	$\log\nu$	μ	$\log\beta$	$\mathcal{T}(f_0)$
0.000312	-0.221	-0.975	0.014	0.002
-0.014	0.647	-0.145	0.016	0.749
-1.31	0.664	-0.146	0.407	-0.61
-817	-0.304	0.083	0.913	0.259



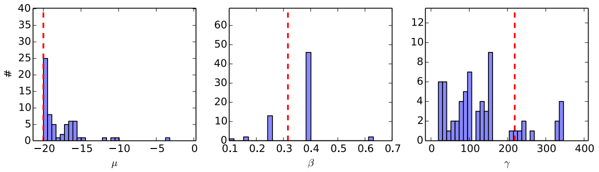
AFT1

λ	$\log\nu$	μ	$\log\beta$	$\mathcal{T}(f_0)$
0.0437	0.447	0.116	-0.887	0.002
-0.000489	-0.001	-0	0.002	1
-3.11	0.842	0.281	0.461	-0
-451	-0.302	0.953	-0.028	0



CUP9

λ	$\log\nu$	μ	$\log\beta$	$\mathcal{T}(f_0)$
-0.00113	0.893	0.014	-0	0.45
-0.0189	-0.066	0.993	0.003	0.1
-36.2	0.404	0.107	0.421	-0.805
-8.14e+03	0.187	0.052	-0.907	-0.374



MCM1

λ	$\log\nu$	μ	$\log\beta$	$\mathcal{T}(f_0)$
-0.000392	0.895	-0.005	-0	0.447
-0.00922	-0.036	0.996	0.002	0.084
-26.1	0.421	0.086	0.326	-0.842
-6.13e+03	0.145	0.032	-0.945	-0.29

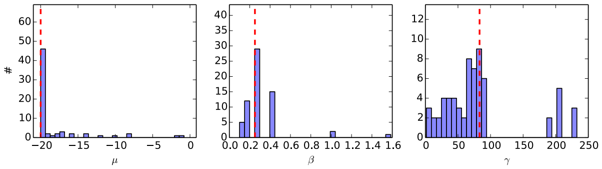


Table B.3: **Estimates of fitting error.** For the 12 TFs in Table B.2, we analyze the quality of fit. Columns, from left to right: (A) Eigenvalues and eigenvectors of the Hessian of the likelihood function around the fit maxima. Eigenvectors of the Hessian represent principal directions and the corresponding eigenvalues represent the curvature in those directions, which should be negative at a local maximum. Positive eigenvalues occur if the maximizer did not reach a maximum. Here, the degeneracy represented by $\gamma = \nu(1 - f_0)$ is apparent as many fits have an eigenvalue close to zero (flat) or even slightly positive in the direction (ν, f_0). For fits subject to the μ-γ degeneracy, one can see a second low eigenvalue corresponding to the μ direction. For computational reasons the Hessian is evaluated using transformed variables $\log \nu$, μ, $\log \beta$, and $\mathcal{T}(f_0) = \text{atanh}(2f_0 - 1)$. (B) For each TF, 64 subsets of the full data set were generated by randomly selecting half of the binding sites in the full data set. Maximum likelihood fits were carried out as for the full data set, except that to reduce computation time the grid spacing in the initial four dimensional parameter search was doubled. Shown here are histograms of the resulting parameters. Red dashed lines indicate the maximum likelihood value of each parameter obtained from the full data set.

Appendix C

Proof of the Path Expansion for Markov Processes

Here we explicitly show that the path expansion in the fully-Markov case actually sums to the total propagator (Eq. 4.12). First note that the Laplace transform of the propagator is

$$\int_0^\infty dt\; e^{-st}\langle\sigma'|e^{t\mathbf{W}}|\sigma\rangle = \langle\sigma'|(s\mathbf{1}-\mathbf{W})^{-1}|\sigma\rangle, \tag{C.1}$$

where $\mathbf{1}$ is an identity matrix of the same dimension as $\mathbf{W}$. We also note that for two matrices $\mathbf{A}$ and $\mathbf{B}$,

$$(\mathbf{A}-\mathbf{B})^{-1} = \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1} + \cdots . \tag{C.2}$$

This can easily be proved by multiplying both sides by $\mathbf{A}-\mathbf{B}$.

Now we decompose the full rate matrix $\mathbf{W}$ into a component with only the diagonal elements, $\mathbf{W}_{\text{diag}}$, and a component with only the off-diagonal elements, $\mathbf{W}_{\text{off}}$, and expand:

$$\begin{aligned}(s\mathbf{1}-\mathbf{W})^{-1} &= (s\mathbf{1}-\mathbf{W}_{\text{diag}}-\mathbf{W}_{\text{off}})^{-1}\\ &= (s\mathbf{1}-\mathbf{W}_{\text{diag}})^{-1} + (s\mathbf{1}-\mathbf{W}_{\text{diag}})^{-1}\mathbf{W}_{\text{off}}(s\mathbf{1}-\mathbf{W}_{\text{diag}})^{-1}\\ &\quad + (s\mathbf{1}-\mathbf{W}_{\text{diag}})^{-1}\mathbf{W}_{\text{off}}(s\mathbf{1}-\mathbf{W}_{\text{diag}})^{-1}\mathbf{W}_{\text{off}}(s\mathbf{1}-\mathbf{W}_{\text{diag}})^{-1} + \cdots\end{aligned} \tag{C.3}$$

Since

$$\begin{aligned}\langle\sigma'|\mathbf{W}_{\text{off}}|\sigma\rangle &= (1-\delta_{\sigma',\sigma})\langle\sigma'|\mathbf{W}|\sigma\rangle,\\ \langle\sigma'|(s\mathbf{1}-\mathbf{W}_{\text{diag}})^{-1}|\sigma\rangle &= \frac{\delta_{\sigma',\sigma}}{s-\langle\sigma|\mathbf{W}|\sigma\rangle} = \delta_{\sigma',\sigma}\frac{\theta(\sigma)}{1+s\theta(\sigma)},\end{aligned} \tag{C.4}$$

we insert identities of the form $\mathbf{1} = \sum_\sigma |\sigma\rangle\langle\sigma|$ to obtain

$$\begin{aligned}\langle\sigma'|(s\mathbf{1}-\mathbf{W})^{-1}|\sigma\rangle &= \langle\sigma'|(s\mathbf{1}-\mathbf{W}_{\text{diag}})^{-1}|\sigma\rangle\\ &\quad+\sum_{\alpha,\beta}\langle\sigma'|(s\mathbf{1}-\mathbf{W}_{\text{diag}})^{-1}|\beta\rangle\langle\beta|\mathbf{W}_{\text{off}}|\alpha\rangle\langle\alpha|(s\mathbf{1}-\mathbf{W}_{\text{diag}})^{-1}|\sigma\rangle+\cdots\\ &= \delta_{\sigma',\sigma}\frac{\theta(\sigma)}{1+s\theta(\sigma)} + \sum_{\alpha,\beta}\frac{\delta_{\sigma',\beta}\theta(\sigma')}{1+s\theta(\sigma')}(1-\delta_{\alpha,\beta})\langle\beta|\mathbf{W}|\alpha\rangle\frac{\delta_{\alpha,\sigma}\theta(\sigma)}{1+s\theta(\sigma)}+\cdots\\ &= \delta_{\sigma',\sigma}\frac{\theta(\sigma)}{1+s\theta(\sigma)} + (1-\delta_{\sigma',\sigma})\frac{\theta(\sigma')\langle\sigma'|\mathbf{W}|\sigma\rangle\theta(\sigma)}{(1+s\theta(\sigma'))(1+s\theta(\sigma))}+\cdots\\ &= \delta_{\sigma',\sigma}\frac{\theta(\sigma)}{1+s\theta(\sigma)} + (1-\delta_{\sigma',\sigma})\frac{\theta(\sigma')}{(1+s\theta(\sigma'))}\frac{\langle\sigma'|\mathbf{Q}|\sigma\rangle}{(1+s\theta(\sigma))}+\cdots\,.\end{aligned} \tag{C.5}$$

This is therefore a sum over individual path propagators in s-space of the form in Eq. 4.9. Carrying out the inverse Laplace transform on both sides back to the time domain yields Eq. 4.12.

Appendix D

Protein Evolution with Independent Folding and Binding

Here we consider a model of protein evolution similar to that in Chapter 5, but in which folding and binding and structurally-independent traits of the protein. (This is based on Refs. 10, 11.) The protein thus has four possible structural states (instead of three, cf. Eq. 5.1), with the following free energies, equilibrium probabilities, and fitnesses:

State	Free energy	Probability	Fitness
folded, bound	$E_f + E_b$	$\frac{e^{-\beta(E_f+E_b)}}{1+e^{-\beta E_f}+e^{-\beta E_b}+e^{-\beta(E_f+E_b)}}$	1
folded, unbound	E_f	$\frac{e^{-\beta E_f}}{1+e^{-\beta E_f}+e^{-\beta E_b}+e^{-\beta(E_f+E_b)}}$	f_0
unfolded, bound	E_b	$\frac{e^{-\beta E_b}}{1+e^{-\beta E_f}+e^{-\beta E_b}+e^{-\beta(E_f+E_b)}}$	f_0
unfolded, unbound	0	$\frac{1}{1+e^{-\beta E_f}+e^{-\beta E_b}+e^{-\beta(E_f+E_b)}}$	f_0

(D.1)

The average fitness is therefore:

$$\mathcal{F}(E_f, E_b) = \frac{e^{-\beta(E_f+E_b)} + f_0 e^{-\beta E_f} + f_0 e^{-\beta E_b} + f_0}{e^{-\beta(E_f+E_b)} + e^{-\beta E_f} + e^{-\beta E_b} + 1}. \tag{D.2}$$

Figure D.1 shows the fitness landscape in Eq. D.2 as a function of the energy traits E_f

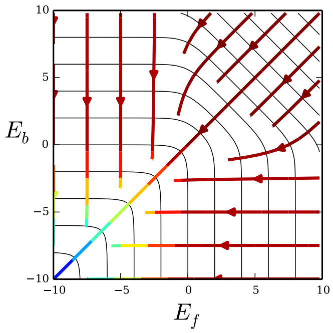


Figure D.1: **Fitness landscape and selection streamlines for independent folding and binding.** Contours are drawn as in Fig. 5.1, but with the fitness function in Eq. D.2 ($f_0 = 0$).

and E_b, along with the streamlines for the selection force (cf. Fig. 5.1). Unconstrained, populations will tend to follow these selection streamlines in trait space to maximize fitness. In this case, selection clearly drives the population toward lower E_f and E_b. The geometry of the fitness contours, or equivalently the selection streamlines, falls into three general classes depending on the overall regimes of E_f and E_b. For low E_f and high E_b (very stable but poor binding), the fitness contours are approximately horizontal lines; hence, selection drives populations toward stronger binding (lower E_b) with little regard for the effect on E_f, since the protein is already so stable. We think of adaptation in this regime as in the *binding phase*, since the need to bind the new target molecule dominates evolutionary dynamics. Figure D.2A shows an example realization of adaptation in this regime. For high E_f and low E_b (strong binding but marginally stable), we see the opposite situation: fitness contours are approximately vertical, and selection drives populations toward more stable folding (lower E_f). We refer to this as the *folding phase* of adaptation (see Fig. D.2C for an example). In between there is a crossover regime when E_f and E_b are of similar magnitude. Here the fitness contours have sharp, 90° corners, and selection tends to optimize binding and folding equally (Fig. D.2B).

By varying the overall energy offsets E_f^{ref} and E_b^{min} we can systematically shift the model between these different phases. We scan over these parameters and calculate average properties of the landscape and adaptation in Figs. D.3 and D.4. We see that in the binding

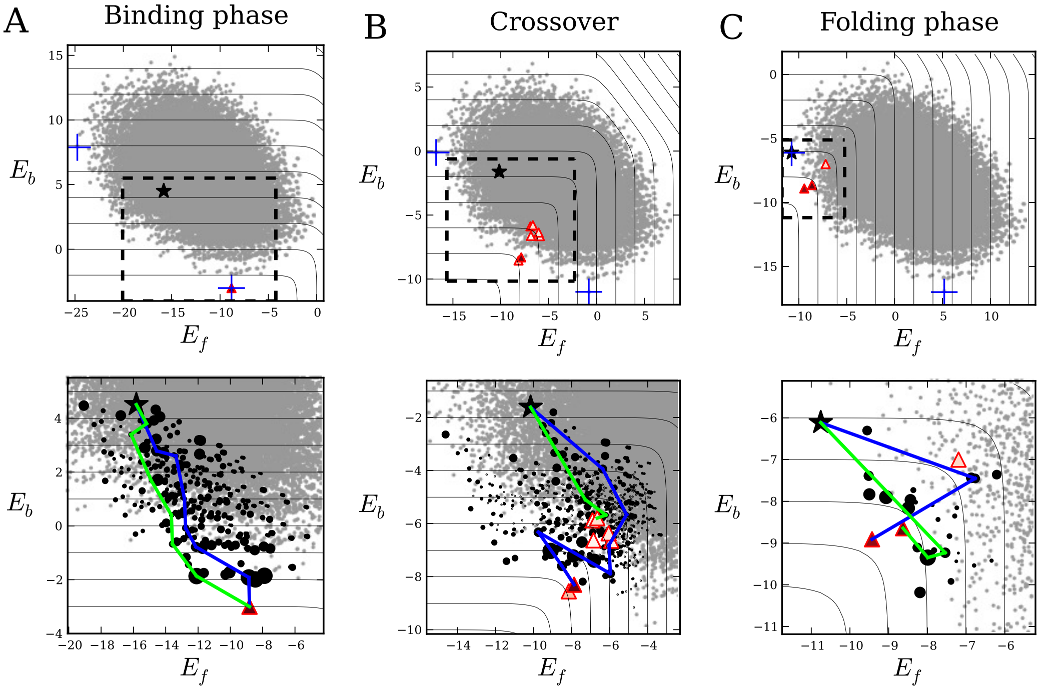


Figure D.2: **Example landscapes for independent folding and binding.** Symbols are the same as Fig. 5.3A,B. The offsets E_f^{ref} and $E_{b_1}^{\mathrm{min}} = E_{b_2}^{\mathrm{min}}$ are different for each realization, but ϵ_f's and the two sets of ϵ_b's (one for $\mathcal{F}_1$ and another for $\mathcal{F}_2$) are the same. (A) Binding phase, with $E_f^{\mathrm{ref}} = -17$ kcal/mol and $E_{b_1}^{\mathrm{min}} = E_{b_2}^{\mathrm{min}} = -3$ kcal/mol. (B) Crossover regime, with $E_f^{\mathrm{ref}} = -9$ kcal/mol and $E_{b_1}^{\mathrm{min}} = E_{b_2}^{\mathrm{min}} = -11$ kcal/mol. (C) Folding phase, with $E_f^{\mathrm{ref}} = -3$ kcal/mol and $E_{b_1}^{\mathrm{min}} = E_{b_2}^{\mathrm{min}} = -17$ kcal/mol. We use $f_0 = 0$, $L = 5$, and $k = 8$.

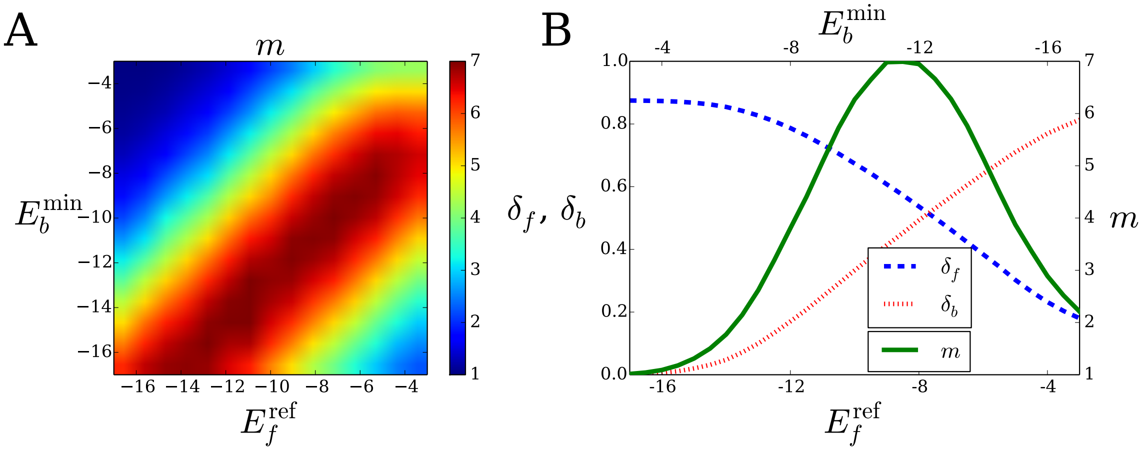


Figure D.3: **Phases of adaptation for independent folding and binding.** (A) Average number m of local fitness maxima as a function of the energy offsets E_f^{ref} and E_b^{min}. (B) Average number m of local fitness maxima (solid, green), average Hamming distance δ_f between the maxima and the best-folding sequence (dashed, blue), and average Hamming distance δ_b between the maxima and the best-binding sequence (dotted, red) for the parameter subspace $E_f^{\mathrm{ref}} + E_b^{\mathrm{min}} = -20$ kcal/mol. Note that the distance between two random sequences is $1 - 1/k = 0.875$, where $k = 8$ is the alphabet size. All data points are averages over 5×10^3 realizations; realizations with no adaptation are excluded. We use $f_0 = 0$, $L = 5$, and $k = 8$.

phase, there is typically a single local fitness maximum which coincides with the best-binding genotype (Fig. D.3). In the folding phase, there are also few local maxima but they tend to be close in genotype space to the best-folding genotype (Fig. D.3). The crossover regime, however, has the most epistasis, as indicated by the number of local maxima, the accessibility of those maxima, and the fraction of fitness landscape realizations in which the global maximum has the largest commitment probability (Figs. D.3, D.4A). The different landscape structures in the binding and folding phases lead to substantial differences in adaptive dynamics. In particular, paths are longer and take more time in the binding phase compared to the folding phase; they are also more diverse (Fig. D.4C,D). Initial and final states in the binding phase are separated by longer Hamming distances (Fig. D.4C). In the folding phase, there is an appreciable probability that no adaptation occurs at all, since the initial state may coincide with one of the local maxima (Fig. D.4B).

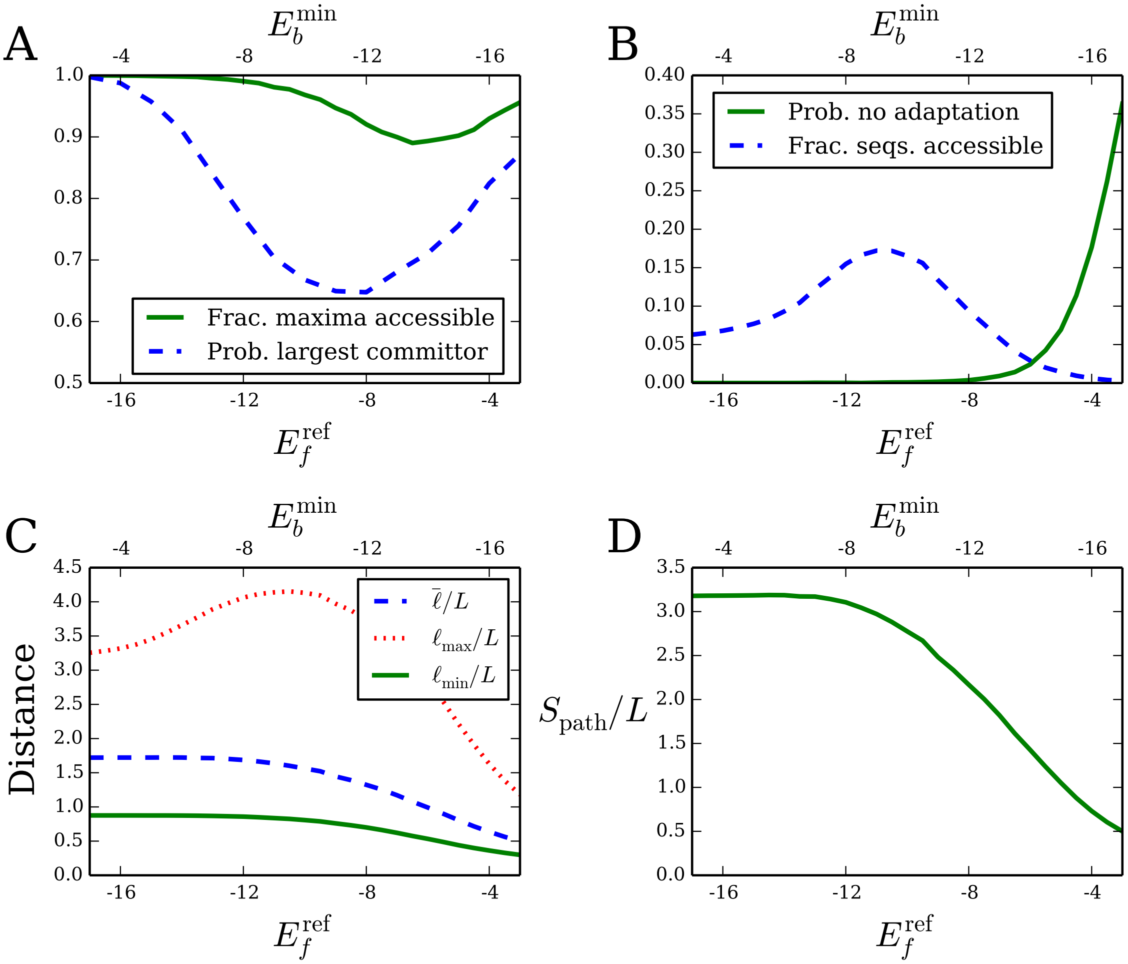


Figure D.4: **Average properties of adaptation for independent folding and binding.** (A) Fraction of local fitness maxima accessible from the initial state (solid, green), and probability that the global maximum has the largest commitment probability (committor) among all local maxima (dashed, blue). (B) Probability that the initial sequence starts at a local maximum resulting in no adaptation (solid, green), and fraction of sequence space accessible to adaptive paths (dashed, blue). (C) Mean path length $\bar{\ell}$ (dashed, blue), maximum possible path length $\ell_{\max}$ (dotted, red), and the average net distance $\ell_{\min}$ between the initial state and final states (solid, green). On average, proteins undergo twice as many substitutions as the net distance $\ell_{\min}$, and the maximum number of substitutions is three times larger than $\ell_{\min}$. (D) Path entropy S_{path}. All quantities in (C) and (D) are per-residue. The probability of no adaptation in (B) is an average over 2×10^4 landscape realizations; all other data points are averages over 5×10^3 realizations, and realizations with no adaptation are excluded.

Bibliography

[1] Schrödinger E (1944) *What is Life?* (Cambridge University Press, Cambridge).

[2] Anderson PW (1972) More is different. *Science* 177:393–396.

[3] Laughlin RB, Pines D (2000) The theory of everything. *Proc Natl Acad Sci USA* 97:28–31.

[4] Laughlin R (2005) *A Different Universe.* (Basic Books, New York).

[5] Nelson P (2007) *Biological Physics: Energy, Information, Life.* (W.H. Freeman and Company, New York, USA).

[6] Phillips R, Kondev J, Theriot J, Garcia H (2012) *Physical Biology of the Cell.* (Garland Science, New York), Second edition.

[7] Dobzhansky T (1973) Nothing in biology makes sense except in the light of evolution. *Am Biol Teach* 35.

[8] Manhart M, Haldane A, Morozov AV (2012) A universal scaling law determines time reversibility and steady state of substitutions under selection. *Theor Popul Biol* 82:66–76.

[9] Haldane A, Manhart M, Morozov AV (2014) Biophysical fitness landscapes for transcription factor binding sites. *PLoS Comput Biol* 10:e1003683.

[10] Manhart M, Morozov AV (2013) Path-based approach to random walks on networks characterizes how proteins evolve new functions. *Phys Rev Lett* 111:088102.

[11] Manhart M, Morozov AV (2014) in *First-Passage Phenomena and Their Applications*, eds. Metzler R, Oshanin G, Redner S. (World Scientific, Singapore).

[12] Manhart M, Morozov AV (2014) Protein folding and binding can emerge as evolutionary spandrels through structural coupling. arXiv:1408.3786.

[13] Darwin CR (1859) *The Origin of Species.* (J. Murray, London).

[14] Gould SJ (1990) *Wonderful Life: The Burgess Shale and the Nature of History.* (W. W. Norton and Company, New York, USA).

[15] Kimura M (1983) *The Neutral Theory of Molecular Evolution.* (Cambridge University Press, Cambridge, UK).

[16] Gillespie JH (1991) *The Causes of Molecular Evolution.* (Oxford University Press, Oxford).

[17] Crow JF, Kimura M (1970) *An Introduction to Population Genetics Theory.* (Harper and Row, New York).

[18] Kimura M, Ohta T (1971) *Theoretical Aspects of Population Genetics.* (Princeton University Press, Princeton).

[19] Gillespie J (2004) *Population Genetics: A Concise Guide.* (The Johns Hopkins University Press, Baltimore, USA).

[20] Ewens WJ (2004) *Mathematical Population Genetics.* (Springer, New York).

[21] Wakeley J (2005) The limits of theoretical population genetics. *Genetics* 169:1–7.

[22] Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc 6th Int Cong Genet* 1:356–366.

[23] Orr HA (2009) Fitness and its role in evolutionary genetics. *Nat Rev Genet* 10:531–539.

[24] Szendro IG, Schenk MF, Franke J, Krug J, de Visser JA (2013) Quantitative analyses of empirical fitness landscapes. *J Stat Mech* p. P01005.

[25] Hunter JD (2007) Matplotlib: A 2D graphics environment. *Computing in Science & Engineering* 9:90–95.

[26] van Nimwegen E, Crutchfield JP, Huynen M (1999) Neutral evolution of mutational robustness. *Proc Natl Acad Sci USA* 96:9716–9720.

[27] Bloom JD, Raval A, Wilke CO (2007) Thermodynamics of neutral protein evolution. *Genetics* 175:255–266.

[28] Poelwijk FJ, Tanase-Nicola S, Kiviet DJ, Tans SJ (2011) Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes. *J Theor Biol* 272:141–144.

[29] Weinreich DM, Lan Y, Wylie CS, Heckendorn RB (2013) Should evolutionary geneticists worry about higher-order epistasis? *Curr Opin Genet Dev* 23:700–707.

[30] Mammano F, Trouplin V, Zennou V, Clavel F (2000) Retracing the evolutionary pathways of human immunodeficiency virus type 1 resistance to protease inhibitors: virus fitness in the absence and in the presence of drug. *J Virol* 74:8524–8531.

[31] Weinreich DM, Delaney NF, DePristo MA, Hartl DL (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* 312:111–114.

[32] Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ (2007) Empirical fitness landscapes reveal accessible evolutionary paths. *Nature* 445:383–386.

[33] Carneiro M, Hartl DL (2010) Adaptive landscapes and protein evolution. *Proc Natl Acad Sci USA* 107:1747–1751.

[34] Franke J, Klozer A, de Visser JA, Krug J (2011) Evolutionary accessibility of mutational pathways. *PLoS Comput Biol* 7:e1002134.

[35] Lobkovsky AE, Wolf YI, Koonin EV (2011) Predictability of evolutionary trajectories in fitness landscapes. *PLoS Comput Biol* 7:e1002302.

[36] Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF (2011) Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* 332:1193–1196.

[37] Chou HH, Chiu HC, Delaney NF, Segrè D, Marx CJ (2011) Diminishing returns epistasis among beneficial mutations decelerates adaptation. *Science* 332:1190–1192.

[38] Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS (2006) Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* 444:929–932.

[39] Breen MS, Kemena C, Vlasov PK, Notredame C, Kondrashov FA (2012) Epistasis as the primary factor in molecular evolution. *Nature* 490:535–538.

[40] Bloom JD, Arnold FH (2009) In the light of directed evolution: Pathways of adaptive protein evolution. *Proc Natl Acad Sci USA* 106:9995–10000.

[41] Kauffman SA, Weinberger ED (1989) The NK model of rugged fitness landscapes and its application to maturation of the immune response. *J Theor Biol* 141:211–245.

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

[43] Aita T et al. (2000) Analysis of a local fitness landscape with a model of the rough Mt. Fuji-type landscape: application to prolyl endopeptidase and thermolysin. *Biopolymers* 54:64–79.

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

[45] Flyvbjerg H, Lautrup B (1992) Evolution in a rugged fitness landscape. *Phys Rev A* 46:6714–6723.

[46] Altenberg L (1997) in *Handbook of Evolutionary Computation*, eds. Bäck T, Fogel D, Michalewicz Z. (IOP Publishing Ltd and Oxford University Press), pp. B2.7:5–B2.7:10.

[47] Rokyta DR, Beisel CJ, Joyce P (2006) Properties of adaptive walks on uncorrelated landscapes under strong selection and weak mutation. *J Theor Biol* 243:114–120.

[48] Kryazhimskiy S, Tkačik G, Plotkin JB (2009) The dynamics of adaptation on correlated fitness landscaps. *Proc Natl Acad Sci USA* 106:18638–18643.

[49] Stauffer D, Aharony A (1994) *Introduction to Percolation Theory.* (Taylor and Francis, London).

[50] Wilke CO (2012) Bringing molecules back into molecular evolution. *PLoS Comput Biol* 8:e1002572.

[51] Harms MJ, Thornton JW (2013) Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat Rev Genet* 14:559–571.

[52] Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV (2002) Birth and death of protein domains: A simple model of evolution explains power law behavior. *BMC Evol Biol* 2:18.

[53] Karev GP, Wolf YI, Berezovskaya FS, Koonin EV (2004) Gene family evolution: an in-depth theoretical and simulation analysis of non-linear birth-death-innovation models. *BMC Evol Biol* 4:32.

[54] Bloom JD, Wilke CO, Arnold FH, Adami C (2004) Stability and the evolvability of function in a model protein. *Biophys J* 86:2758–2764.

[55] DePristo MA, Weinreich DM, Hartl DL (2005) Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* 6:678–687.

[56] Bloom JD et al. (2005) Thermodynamic prediction of protein neutrality. *Proc Natl Acad Sci USA* 102:606–611.

[57] Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH (2005) Why highly expressed proteins evolve slowly. *Proc Natl Acad Sci USA* 102:14338–14343.

[58] Bloom JD, Labthavikul ST, Otey CR, Arnold FH (2006) Protein stability promotes evolvability. *Proc Natl Acad Sci USA* 103:5869–5874.

[59] Zeldovich KB, Chen P, Shakhnovich EI (2007) Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc Natl Acad Sci USA* 104:16152–16157.

[60] Bloom JD, Glassman MJ (2009) Inferring stabilizing mutations from protein phylogenies: Application to influenza hemagglutinin. *PLoS Comput Biol* 5:e1000349.

[61] Lobkovsky AE, Wolf YI, Koonin EV (2010) Universal distribution of protein evolution rates as a consequence of protein folding physics. *Proc Natl Acad Sci USA* 107:2983–2988.

[62] Serohijos AWR, Rimas Z, Shakhnovich EI (2012) Protein biophysics explains why highly abundant proteins evolve slowly. *Cell Rep* 2:249–256.

[63] Serohijos AWR, Shakhnovich EI (2014) Contribution of selection for protein folding stability in shaping the patterns of polymorphisms in coding regions. *Mol Biol Evol* 31:165–176.

[64] Heo M, Kang L, Shakhnovich EI (2009) Emergence of species in evolutionary "simulated annealing". *Proc Natl Acad Sci USA* 106:1869–1874.

[65] Dill KA, Ghosh K, Schmit JD (2011) Physical limits of cells and proteomes. *Proc Natl Acad Sci USA* 108:17876–17882.

[66] Heo M, Maslov S, Shakhnovich EI (2011) Topology of protein interaction network shapes protein abundances and strengths of their function and nonspecific interactions. *Proc Natl Acad Sci USA* 108:4258–4263.

[67] Gerland U, Hwa T (2002) On the selection and evolution of regulatory DNA motifs. *J Mol Evol* 55:386–400.

[68] Sengupta AM, Djordjevic M, Shraiman BI (2002) Specificity and robustness in transcription control networks. *Proc Natl Acad Sci USA* 99:2072–2077.

[69] Berg J, Lässig M (2003) Stochastic evolution of transcription factor binding sites. *Biophysics (Moscow)* 48:S36–S44.

[70] Berg J, Willmann S, Lässig M (2004) Adaptive evolution of transcription factor binding sites. *BMC Evol Biol* 4:42.

[71] Lässig M (2007) From biophysics to evolutionary genetics: statistical aspects of gene regulation. *BMC Bioinformatics* 8:S7.

[72] Mustonen V, Kinney J, Callan CG, Lässig M (2008) Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites. *Proc Natl Acad Sci USA* 105:12376–12381.

[73] François P, Hakim V (2004) Design of genetic networks with specified functions by evolution *in silico*. *Proc Natl Acad Sci USA* 101:580–585.

[74] François P, Hakim V, Siggia ED (2007) Deriving structure from evolution: metazoan segmentation. *Mol Syst Biol* 3:154.

[75] François P, Siggia ED (2008) A case study of evolutionary computation of biochemical adaptation. *Phys Biol* 5:026009.

[76] François P, Siggia ED (2010) Predicting embryonic patterning using mutual entropy fitness and *in silico* evolution. *Development* 137:2385–2395.

[77] Warmflash A, François P, Siggia ED (2012) Pareto evolution of gene networks: an algorithm to optimize multiple fitness objectives. *Phys Biol* 9:056001.

[78] Kimura M, Ohta T (1969) The average number of generations until fixation of a mutant gene in a finite population. *Genetics* 61:763–771.

[79] Kimura M, Ohta T (1969) The average number of generations until extinction of an individual mutant gene in a finite population. *Genetics* 63:701–709.

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

[81] Champagnat N (2006) A microscopic interpretation for adaptive dynamics trait substitution sequence models. *Stochastic Process. Appl.* 116:1127–1160.

[82] Champagnat N, Ferrire R, Mlard S (2006) Unifying evolutionary dynamics: from individual stochastic processes to macroscopic models. *Theor Popul Biol* 69:297–321.

[83] Charlesworth B (2009) Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 10:195–205.

[84] Gillespie JH (1984) Molecular evolution over the mutational landscape. *Evolution* 38:1116–1129.

[85] Lobkovsky AE, Koonin EV (2012) Replaying the tape of life: quantification of the predictability of evolution. *Front Gene* 3:246.

[86] Tenaillon O et al. (2012) The molecular diversity of adaptive convergence. *Science* 335:457–461.

[87] Barrick JE, Lenski RE (2013) Genome dynamics during experimental evolution. *Nat Rev Genet* 14:827–839.

[88] Desai MM (2013) Statistical questions in experimental evolution. *J Stat Mech* 2013:P01003.

[89] Orr H (2005) The genetic theory of adaptation: A brief history. *Nat Rev Genet* 6:119–127.

[90] Fisher RA (1958) *The Genetical Theory of Natural Selection.* (Dover, New York).

[91] Ohta T, Tachida H (1990) Theoretical study of near neutrality. I. Heterozygosity and rate of mutant substitution. *Genetics* 126:219–229.

[92] Ohta T (1992) Theoretical study of near neutrality. II. Effect of subdivided population structure with local extinction and recolonization. *Genetics* 130:917–923.

[93] Wright S (1931) Evolution in Mendelian populations. *Genetics* 16:97–159.

[94] Moran PAP (1958) Random processes in genetics. *Proc Camb Philos Soc* 54:60–71.

[95] Ewens W (1967) The probability of survival of a new mutant in a fluctuating environment. *Heredity* 22:438–443.

[96] Maruyama T (1970) On the fixation probability of mutant genes in a subdivided population. *Genet Res Camb* 15:221–225.

[97] Otto SP, Whitlock MC (1997) The probability of fixation in populations of changing size. *Genetics* 146:723 –733.

[98] Möhle M (2001) Forward and backward diffusion approximations for haploid exchangeable population models. *Stoch Proc Appl* 95:133–149.

[99] Möhle M, Sagitov S (2001) A classification of coalescent processes for haploid exchangeable population models. *Ann. Prob.* 29:1547–1562.

[100] Whitlock M (2003) Fixation probability and time in subdivided populations. *Genetics* 164:767–779.

[101] Kimura M (1955) Solution of a process of random genetic drift with a continuous model. *Proc Natl Acad Sci USA* 41:144–150.

[102] Cannings C (1974) The latent roots of certain Markov chains arising in genetics: A new approach, I. Haploid models. *Adv Appl Prob* 6:260–290.

[103] Wichman H, Badgett M, Scott L, Boulianne C, Bull J (1999) Different trajectories of parallel evolution during viral adaptation. *Science* 285:422–424.

[104] Bull JJ, Badgett MR, Wichman HA (2000) Big-benefit mutations in a bacteriophage inhibited with heat. *Mol Biol Evol* 17:942–950.

[105] Holder KK, Bull JJ (2001) Profiles of adaptation in two similar viruses. *Genetics* 159:1393–1404.

[106] Barrett RDH, M'Gonigle LK, Otto SP (2006) The distribution of beneficial mutant effects under strong selection. *Genetics* 174:2071–2079.

[107] Orr H (2001) The genetics of species differences. *Trends Ecol Evol* 16:343–350.

[108] Eyre-Walker A, Keightley P (2007) The distribution of fitness effects of new mutations. *Nat Rev Genet* 8:610–618.

[109] Morjan C, Rieseberg L (2004) How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Mol Ecol* 13:1341–1356.

[110] Barrett RDH, MacLean RC, Bell G (2006) Mutations of intermediate effect are responsible for adaptation in evolving *Pseudomonas fluorescens* populations. *Biol Lett* 2:236–238.

[111] Eigen M, McCaskill J, Schuster P (1989) The molecular quasi-species. *Adv Chem Phys* 75:149–263.

[112] Bürger R (2000) *The Mathematical Theory of Selection, Recombination, and Mutation.* (Wiley, New York).

[113] Proulx SR (2000) The ESS under spatial variation with applications to sex allocation. *Theor Popul Biol* 58:33–47.

[114] Shpak M (2007) Selection against demographic stochasticity in age-structured populations. *Genetics* 177:2181 –2194.

[115] Parsons TL, Quince C, Plotkin JB (2010) Some consequences of demographic stochasticity in population genetics. *Genetics* 185:1345 –1354.

[116] Ochman H, Selander RK (1984) Evidence for clonal population structure in *Escherichia coli. Proc Natl Acad Sci USA* 81:198–201.

[117] Wick LM, Weilenmann H, Egli T (2002) The apparent clock-like evolution of *Escherichia coli* in glucose-limited chemostats is reproducible at large but not at small population sizes and can be explained with Monod kinetics. *Microbiology* 148:2889–2902.

[118] Dos Vultos T et al. (2008) Evolution and diversity of clonal bacteria: The paradigm of *Mycobacterium tuberculosis. PLoS ONE* 3:e1538EP.

[119] Achtman M (2008) Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu Rev Microbiol* 62:53–70.

[120] McVean GAT, Vieira J (2001) Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila. Genetics* 157:245 –257.

[121] Sella G, Hirsh AE (2005) The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci USA* 102:9541–9546.

[122] Yang Z (2006) *Computational Molecular Evolution.* (Oxford University Press, Oxford).

[123] Mustonen V, Lässig M (2010) Fitness flux and the ubiquity of adaptive evolution. *Proc Natl Acad Sci USA* 107:4248–4253.

[124] Allen LJS (2011) *An Introduction to Stochastic Processes with Applications to Biology.* (Chapman and Hall, CRC, Boca Raton), Second edition.

[125] Roberts FS (1979) *Measurement Theory with Applications to Decisionmaking, Utility, and the Social Sciences.* (Addison-Wesley, Reading).

[126] Iwasa Y (1988) Free fitness that always increases in evolution. *J Theor Biol* 135:265–281.

[127] Li WH (1987) Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *J Mol Evol* 24:337–345.

[128] Gillespie JH (1975) Natural selection for within-generation variance in offspring number II. Discrete haploid models. *Genetics* 81:403 –413.

[129] Rouzine IM, Rodrigo A, Coffin JM (2001) Transition between stochastic evolution and deterministic evolution in the presence of selection: General theory and application to virology. *Microbiol Mol Biol Rev* 65(1):151–185.

[130] Kimura M (1957) Some problems of stochastic processes in genetics. *Ann Math Stat* 28:882–901.

[131] van Kampen N (2007) *Stochastic Processes in Physics and Chemistry.* (Elsevier, Amsterdam).

[132] Sjödin P, Kaj I, Krone S, Lascoux M, Nordborg M (2005) On the meaning and existence of an effective population size. *Genetics* 169:1061–1070.

[133] Gillespie JH (1974) Natural selection for within-generation variance in offspring number. *Genetics* 76:601 –606.

[134] Gillespie JH (1977) Natural selection for variances in offspring numbers: A new evolutionary principle. *Am Nat* 111:1010–1014.

[135] Watterson GA (1977) Reversibility and the age of an allele. II. *Theor Popul Biol* 12:179–196.

[136] Levikson B (1977) The age distribution of Markov processes. *J Appl Prob* 14:492–506.

[137] Kurtz TG (1981) *Approximation of Population Processes*, CBMS-NSF Regional Conference Series in Applied Mathematics. (Society for Industrial and Applied Mathematics, Philadelphia).

[138] Ewens WJ (1990) in *Mathematical and Statistical Developments of Evolutionary Theory*, ed. Lessard S. (Kluwer Academic Publishers, Amsterdam), pp. 177–227.

[139] Kingman JFC (1982) The coalescent. *Stoch Proc Appl* 13:235–248.

[140] Watterson GA (1976) Reversibility and the age of an allele. I. *Theor Popul Biol* 10:239–253.

[141] Bulmer MG (1991) The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897–907.

[142] McVean GA, Charlesworth B (1999) A population genetics model for the evolution of synonymous codon usage: patterns and predictions. *Genet Res* 74:145–158.

[143] McVean GA, Vieira J (1999) The evolution of codon preferences in *Drosophila*: a maximum-likelihood approach to parameter estimation and hypothesis testing. *J Mol Evol* 49:63–75.

[144] Nielsen R, DuMont VLB, Hubisz MJ, Aquadro CF (2007) Maximum likelihood estimation of ancestral codon usage bias parameters in *Drosophila*. *Mol Biol Evol* 24:228–235.

[145] Yang Z, Nielsen R (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. *Mol Biol Evol* 25:568–579.

[146] Jukes TH, Cantor CR (1969) in *Mammalian protein metabolism*, ed. Munro HN. (Academic Press, New York), pp. 21–123.

[147] Kimura M (1980) A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. *J Mol Evol* 16:111–120.

[148] Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 10:512–526.

[149] Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. *J Mol Evol* 17:368–376.

[150] Felsenstein J (2011) PHYLIP (Phylogeny Inference Package) version 3.69.

[151] Rodríguez F, Oliver J, Marín A, Medina J (1990) The general stochastic model of nucleotide substitution. *J Theor Biol* 142:485–501.

[152] Barry D, Hartigan JA (1987) Asynchronous distance between homologous DNA sequences. *Biometrics* 43:261–276.

[153] Levin DA, Peres Y, Wilmer EL (2009) *Markov Chains and Mixing Times.* (American Mathematical Society, Providence, RI).

[154] Sarai A, Takeda Y (1989) Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically. *Proc Natl Acad Sci USA* 86:6513–6517.

[155] Lehming N, Sartorius J, Kisters-Woike B, von Wilcken-Bergmann B, Muller-Hill B (1990) Mutant lac repressors with new specificities hint at rules for protein-DNA recognition. *EMBO J* 9:615–621.

[156] Bershtein S, Goldin K, Tawfik DS (2008) Intense neutral drifts yield robust and evolvable consensus proteins. *J Mol Biol* 379:1029–1044.

[157] Ptashne M, Gann A (2002) *Genes and Signals.* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor).

[158] Rhee HS, Pugh BF (2011) Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* 147:1408–1419.

[159] Lee TI et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298:799–804.

[160] Harbison CT et al. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature* 431:99–104.

[161] MacIsaac K et al. (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* 7:113.

[162] Chen K, van Nimwegen E, Rajewsky N, Siegal ML (2010) Correlating gene expression variation with cis-regulatory polymorphism in *Saccharomyces cerevisiae*. *Genome Biol Evol* 2:697–707.

[163] Stormo GD, Fields DS (1998) Specificity, free energy and information content in protein-DNA interactions. *TIBS* 23:109–113.

[164] Berger MF et al. (2006) Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. *Nat Biotech* 24:1429–1435.

[165] Foat BC, Morozov AV, Bussemaker HJ (2006) Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* 22:e141–e149.

[166] Fordyce PM et al. (2010) *De novo* identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis. *Nature Biotech* 28:970–975.

[167] Winzeler EA et al. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285:901–906.

[168] Berg OG, von Hippel PH (1987) Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters. *J Mol Biol* 193:723–743.

[169] Tsai IJ, Bensasson D, Burt A, Koufopanou V (2008) Population genomics of the wild yeast *Saccharomyces paradoxus*: Quantifying the life cycle. *Proc Natl Acad Sci USA* 105:4957–4962.

[170] Dujon B (2010) Yeast evolutionary genomics. *Nat Rev Genet* 11:512–524.

[171] Liti G et al. (2009) Population genomics of domestic and wild yeasts. *Nature* 458:337–341.

[172] Doniger SW et al. (2008) A catalog of neutral and deleterious polymorphism in yeast. *PLoS Genet* 4:e1000183.

[173] Fay JC, Benavides JA (2005) Evidence for domesticated and wild populations of *Saccharomyces cerevisiae*. *PLoS Genet* 1:66–71.

[174] Replansky T, Koufopanou V, Greig D, Bell G (2008) *Saccharomyces sensu stricto* as a model system for evolution and ecology. *Trends Ecol Evol (Amst)* 23:494–501.

[175] Giaever G et al. (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418:387–391.

[176] Holstege FCP et al. (1998) Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95:717–728.

[177] Hahn MW, Stajich JE, Wray GA (2003) The effects of selection against spurious transcription factor binding sites. *Mol Biol Evol* 20:901–906.

[178] Djordjevic M, Sengupta AM, Shraiman BI (2003) A biophysical approach to transcription factor binding site discovery. *Genome Res* 13:2381–2390.

[179] Burnham KP, Anderson DR (2002) *Model Selection and Multimodal Inference: A Practical Information-Theoretic Approach.* (Springer-Verlag, New York), Second edition.

[180] Jordan IK, Rogozin IB, Wolf YI, Koonin EV (2002) Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. *Genome Res* 12:962–968.

[181] Pal C, Papp B, Hurst LD (2003) Genomic function (communication arising): Rate of evolution and gene dispensability. *Nature* 421:496–497.

[182] Zhang J, He X (2005) Significant impact of protein dispensability on the instantaneous rate of protein evolution. *Mol Biol Evol* 22:1147–1155.

[183] Choi SC, Hobolth A, Robinson DM, Kishino H, Thorne JL (2007) Quantifying the impact of protein tertiary structure on molecular evolution. *Mol Biol Evol* 24:1769–1782.

[184] Krylov DM, Wolf YI, Rogozin IB, Koonin EV (2003) Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Res* 13:2229–2235.

[185] Wang Z, Zhang J (2009) Why is the correlation between gene importance and gene evolutionary rate so weak? *PLoS Genet* 5:e1000329.

[186] Fang G, Rocha E, Danchin A (2005) How essential are nonessential genes? *Mol Biol Evol* 22:2147–2156.

[187] Hirsh AE, Fraser HB (2001) Protein dispensability and rate of evolution. *Nature* 411:1046–1049.

[188] Doniger SW, Fay JC (2007) Frequent gain and loss of functional transcription factor binding sites. *PLoS Comput Biol* 3:e99.

[189] Raijman D, Shamir R, Tanay A (2008) Evolution and selection in yeast promoters: Analyzing the combined effect of diverse transcription factor binding sites. *PLoS Comput Biol* 4:e7.

[190] Tirosh I, Weinberger A, Bezalel D, Kaganovich M, Barkai N (2008) On the relation between promoter divergence and gene expression evolution. *Mol Syst Biol* 4.

[191] Tuch BB, Galgoczy DJ, Hernday AD, Li H, Johnson AD (2008) The evolution of combinatorial gene regulation in fungi. *PLoS Biol* 6:e38.

[192] Jovelin R, Phillips P (2009) Evolutionary rates and centrality in the yeast gene regulatory network. *Genome Biol* 10:R35.

[193] Wuchty S, Oltvai ZN, Barabasi AL (2003) Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nat Genet* 35:176–179.

[194] van Dijk ADJ, van Mourik S, van Ham RCHJ (2012) Mutational robustness of gene regulatory networks. *PLoS ONE* 7:e30591.

[195] He X, Duque TSPC, Sinha S (2012) Evolutionary origins of transcription factor binding site clusters. *Mol Biol Evol* 29:1059–1070.

[196] He BZ, Holloway AK, Maerkl SJ, Kreitman M (2011) Does positive selection drive transcription factor binding site turnover? A test with *Drosophila* cis-regulatory modules. *PLoS Genet* 7:e1002053.

[197] Habib N, Wapinski I, Margalit H, Regev A, Friedman N (2012) A functional selection model explains evolutionary robustness despite plasticity in regulatory networks. *Mol Syst Biol* 8.

[198] Weiss GH (1994) *Aspects and Applications of the Random Walk.* (North Holland, Amsterdam).

[199] Noé F, Schütte C, Vanden-Eijnden E, Reich L, Weikl TR (2009) Constructing the equilibrium ensemble of folding pathways from short off-equilibrium simulations. *Proc Natl Acad Sci USA* 106:19011–19016.

[200] Bolhuis PG, Chandler D, Dellago C, Geissler PL (2002) Transition path sampling: Throwing ropes over rough mountain passes, in the dark. *Ann Rev Phys Chem* 53:291–318.

[201] ben-Avraham D, Havlin S (2000) *Diffusion and Reactions in Fractals and Disordered Systems.* (Cambridge University Press, Cambridge).

[202] Condamin S, Bénichou O, Tejedor V, Voituriez R, Klafter J (2007) First-passage times in complex scale-invariant media. *Nature* 450:77–80.

[203] Roma DM, O'Flanagan RA, Ruckenstein AE, Sengupta AM (2005) Optimal path to epigenetic switching. *Phys Rev E* 71:011902.

[204] Waddington CH (1957) *The Strategy of the Genes. A Discussion of Some Aspects of Theoretical Biology.* (Allen and Unwin, London).

[205] Enver T, Pera M, Peterson C, Andrews PW (2009) Stem cell states, fates, and the rules of attraction. *Cell Stem Cell* 4:387.

[206] Flomenbom O, Klafter J (2005) Closed-form solutions for continuous time random walks on finite chains. *Phys Rev Lett* 95:098105.

[207] Sun SX (2006) Path summation formulation of the master equation. *Phys Rev Lett* 96:210602.

[208] Flomenbom O, Silbey RJ (2007) Properties of the generalized master equation: Green's functions and probability density functions in the path representation. *J Chem Phys* 127:034103.

[209] Flomenbom O, Silbey RJ (2007) Path-probability density functions for semi-Markovian random walks. *Phys Rev E* 76:041101.

[210] Harland B, Sun SX (2007) Path ensembles and path sampling in nonequilibrium stochastic systems. *J Chem Phys* 127:104103.

[211] Klafter J, Silbey R (1980) Derivation of the continuous-time random-walk equation. *Phys Rev Lett* 44:55.

[212] Redner S (2001) *A Guide to First-Passage Processes.* (Cambridge University Press, Cambridge).

[213] Maes C, Netočný K, Wynants B (2009) Dynamical fluctuations for semi-Markov processes. *J Phys A: Math Theor* 42:365002.

[214] Cox DR (1962) *Renewal Theory.* (Methuen, London).

[215] Abramowitz M, Stegun IA (1964) *Handbook of Mathematical Functions.* (Dover, New York), Tenth edition.

[216] Majid I, ben-Avraham D, Havlin S, Stanley HE (1984) Exact-enumeration approach to random walks on percolation clusters in two dimensions. *Phys Rev B* 30:1626–1628.

[217] Bollt EM, ben-Avraham D (2005) What is special about diffusion on scale-free nets? *New J Phys* 7:26–47.

[218] Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) *Numerical Recipes in C: The Art of Scientific Computing.* (Cambridge, Cambridge), Second edition.

[219] Metzner P, Schütte C, Vanden-Eijnden E (2009) Transition path theory for Markov jump processes. *Multiscale Model Simul* 7:1192–1219.

[220] Hänggi P, Talkner P, Borkovec M (1990) Reaction rate theory: fifty years after Kramers. *Rev Mod Phys* 62:251–341.

[221] Finkelstein AV, Ptitsyn O (2002) *Protein Physics: A Course of Lectures.* (Academic Press, London, UK).

[222] Dellago C, Bolhuis PG, Csajka FS, Chandler D (1998) Transition path sampling and the calculation of rate constants. *J Chem Phys* 108:1964–1977.

[223] Dellago C, Bolhuis PG, Geissler PL (2003) Transition path sampling. *Adv Chem Phys* 123:1.

[224] Hummer G (2004) From transition paths to transition states and rate coefficients. *J Chem Phys* 120:516–523.

[225] More T, Walczak A, Zamponi F (2012) Transition path sampling algorithm for discrete many-body systems. *Phys Rev E* 85:036710.

[226] Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JK (2003) Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol* 20:1692–1704.

[227] Rodrigue N, Lartillot N, Bryant D, Philippe H (2005) Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene* 347:207–217.

[228] Rodrigue N, Philippe H, Lartillot N (2006) Assessing site-interdependent phylogenetic models of sequence evolution. *Mol Biol Evol* 23:1762–1775.

[229] Rodrigue N, Kleinman CL, Philippe H, Lartillot N (2009) Computational methods for evaluating phylogenetic models of coding sequence evolution with dependence between codons. *Mol Biol Evol* 26:1663–1676.

[230] E W, Vanden-Eijnden E (2006) Toward a theory of transition paths. *J Stat Phys* 123:503–523.

[231] Metzner P, Schütte C, Vanden-Eijnden E (2006) Illustration of transition path theory on a collection of simple examples. *J Chem Phys* 125:084110.

[232] E W, Vanden-Eijnden E (2010) Transition-path theory and path-finding algorithms for the study of rare events. *Ann Rev Phys Chem* 61:391–420.

[233] Creighton TE (1992) *Proteins: Structures and Molecular Properties.* (W.H. Freeman and Company, New York, USA).

[234] Chen K, Arnold FH (1993) Tuning the activity of an enzyme for unusual environments: sequential random mutagenesis of subtilisin E for catalysis in dimethylformamide. *Proc Natl Acad Sci USA* 90:5618–5622.

[235] Campbell RE et al. (2002) A monomeric red fluorescent protein. *Proc Natl Acad Sci USA* 99:7877–7882.

[236] Eijsink VG, Gaseidnes S, Borchert TV, van den Burg B (2005) Directed evolution of enzyme stability. *Biomol Eng* 22:21–30.

[237] Jackel C, Kast P, Hilvert D (2008) Protein design by directed evolution. *Annu Rev Biophys* pp. 153–173.

[238] Bucciantini M et al. (2002) Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* 416:507–511.

[239] Drummond DA, Wilke CO (2008) Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–352.

[240] Geiler-Samerotte KA et al. (2011) Misfolded proteins impose a dosage-dependent fitness cost and trigger a cytosolic unfolded protein response in yeast. *Proc Natl Acad Sci USA* 108:680–685.

[241] Bershtein S, Mu W, Serohijos AWR, Zhou J, Shakhnovich EI (2013) Protein quality control acts on folding intermediates to shape the effects of mutations on organismal fitness. *Mol Cell* 49:133–144.

[242] Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS (2007) The stability effects of protein mutations appear to be universally distributed. *J Mol Biol* 369:1318–1332.

[243] Tokuriki N, Stricher F, Serrano L, Tawfik DS (2008) How protein stability and new functions trade off. *PLoS Comput Biol* 4:e1000002.

[244] Wang X, Minasov G, Shoichet BK (2002) Evolution of an antibiotic resistance enzyme constrained by stability and activity trade-offs. *J Mol Biol* 320:85–95.

[245] Sun SB et al. (2013) Mutational analysis of 48g7 reveals that somatic hypermutation affects both antibody stability and binding affinity. *J Am Chem Soc* 135:9980–9983.

[246] Taverna DM, Goldstein RA (2002) Why are proteins marginally stable? *Proteins* 46:105–109.

[247] Johnson ME, Hummer G (2011) Nonspecific binding limits the number of proteins in a cell and shapes their interaction networks. *Proc Natl Acad Sci USA* 108:603–608.

[248] Gould SJ, Lewontin RC (1979) The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proc R Soc Lond B* 205:581–598.

[249] Pigliucci M, Kaplan J (2000) The fall and rise of Dr Pangloss: adaptationism and the Spandrels paper 20 years later. *Trends Ecol Evol* 15:66–77.

[250] Weiss MA et al. (2002) Protein structure and the spandrels of San Marco: Insulin's receptor-binding surface is buttressed by an invariant leucine essential for its stability. *Biochemistry* 41:809–819.

[251] Fall S et al. (2007) Horizontal gene transfer regulation in bacteria as a 'spandrel' of DNA repair mechanisms. *PLoS ONE* 2:e1055.

[252] Hane F (2013) Are amyloid fibrils molecular spandrels? *FEBS Lett* 587:3617–3619.

[253] Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the genetic level. *Nat Rev Genet* 12:767–780.

[254] Soskine M, Tawfik DS (2010) Mutational effects and the evolution of new protein functions. *Nat Rev Genet* 11:572–582.

[255] Bershtein S, Mu W, Shakhnovich EI (2012) Soluble oligomerization provides a beneficial fitness effect on destabilizing mutations. *Proc Natl Acad Sci USA* 109:4857–4862.

[256] Weinreich DM, Watson RA, Chao L (2005) Sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* 59:1165–1174.

[257] Clackson T, Wells JA (1995) A hot spot of binding energy in a hormone-receptor interface. *Science* 267:383–386.

[258] Moreira IS, Fernandes PA, Ramos MJ (2007) Hot spots — a review of the protein-protein interface determinant amino-acid residues. *Proteins* 68:803–812.

[259] Wells JA (1990) Additivity of mutational effects in proteins. *Biochemistry* 29:8509–8517.

[260] Serrano L, Day AG, Fersht AR (1993) Step-wise mutation of barnase to binase. A procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. *J Mol Biol* 233:305–312.

[261] Thorn KS, Bogan AA (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics* 17:284–285.

[262] Walsh C (2003) *Antibiotics: Actions, Origins, Resistance.* (American Society for Microbiology, Washington, DC).

[263] Lynch M (2007) *The Origins of Genome Architecture.* (Sinauer, Sunderland).

[264] Kumar MD et al. (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nuleic Acids Res* 34:D204–D206.

[265] Shoval O et al. (2012) Evolutionary trade-offs, Pareto optimality, and the geometry of phenotype space. *Science* 336:1157–1160.

[266] Lynch M (2007) The evolution of genetic networks by non-adaptive processes. *Nat Rev Genet* 8:803–813.

[267] Wright PE, Dyson HJ (2009) Linking folding and binding. *Curr Opin Struct Biol* 19:31–38.

[268] Brown CJ, Johnson AK, Dunker AK, Daughdrill GW (2011) Evolution and disorder. *Curr Opin Struct Biol* 21:441–446.

[269] Dixit PD, Maslov S (2013) Evolutionary capacitance and control of protein stability in protein-protein interaction networks. *PLoS Comput Biol* 9:e1003023.

[270] Istomin AY, Gromiha MM, Vorov OK, Jacobs DJ, Livesay DR (2008) New insight into long-range nonadditivity within protein double-mutant cycles. *Proteins* 70:915–924.

[271] Reetz MT (2013) The importance of additive and non-additive mutational effects in protein engineering. *Angew Chem Int Ed* 52:2658–2666.

[272] Metzker ML (2010) Sequencing technologies — the next generation. *Nat Rev Genet* 11:31–46.

[273] Berman HM, Henrick K, Nakamura H (2003) Announcing the worldwide Protein Data Bank. *Nat Struct Biol* 10:980.

[274] Stark C et al. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34(Database issue):D535–D539.

[275] Cherry JM et al. (2012) Saccharomyces genome database: the genomics resource of budding yeast. *Nucl Acids Res* 40:D700–D705.

[276] Lenski R, Elena SF (2003) Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. *Nat Rev Genet* 4:457–469.

[277] Kemeny JG, Snell JL (1960) *Finite Markov Chains.* (Van Nostrand, New York).

[278] Larkin MA et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* 23:2947–2948.

[279] Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 24:1586–1591.