BIOPHYSICAL MODELS OF EVOLUTION

By

ALLAN M. HALDANE

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
Alexandre V Morozov
and approved by

New Brunswick, New Jersey
January, 2014

ABSTRACT OF THE DISSERTATION

Biophysical Models of Evolution

By ALLAN M. HALDANE

Dissertation Director:
Alexandre V Morozov

The recent emergence of quantitative high-throughput experimental technology and new biophysical knowledge may finally enable significant empirical and quantitative understanding of adaptive evolution, which has been elusive for almost a century. The modern aim is to unite classical population genetics with biophysical molecular models, and to connect physical properties of biological molecules such as DNA, RNA and proteins with evolutionary parameters. In this vein, I have studied such population models theoretically, and applied one such model to yeast evolution.

In Chapters 2 and 3, I will discuss "universality" in population genetics, in particular the universal applicability of a formula for the steady state distribution of phenotypes in a population evolving in the "monomorphic regime", which describes most organisms. I show that this formula applies far outside the "weak selection" context it was originally developed in, and that it is a universal feature of evolution in this regime. Such universal features will be important components of any grand theory of adaptive evolution, and are essential for studies of real populations where the microscopic population dynamics are generally unknown.

I then apply this model to a particular molecular system in yeast, Transcription Factor binding sites, which are short DNA sequences which play an important role in gene regulation. Using the functional relationship between evolutionary fitness and the phenotypic steady state distribution, I infer the form of the selective pressure the sites experience, and find it is consistent with a simple thermodynamic model of two-state TF-DNA binding. I also show that the selection pressure a site experiences is decoupled from the selection pressure on the gene it regulates. This suggests that binding sites for a given TF evolve over a universal fitness landscape derived from simple physical interactions.

Acknowledgments

I would like to deeply thank Alex Morozov, my advisor, who has always been positive, supportive, and patient, and always made science exciting. Many thanks also to Michael Manhart with whom I collaborated on the projects presented here, and whom it has been a pleasure to work with. Thanks to my graduate committee for their time and effort, and in particular Ron Levy, for his support in this last semester, for an exciting new project, and for his push to make me a better scientist. I would also like to thank the fellow students I've known during my time here, especially George, Manjul, August, Deepa, Eliane, Aatish, Kshitij, Razvan, Liyang, Mike, Bill, Mohammed, Dave, Ted, and especially Julie Tsitron, who has become a great friend through these years, and who always brightens my day.

Thanks to all!

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Imagine an population of cells as they grow, die, and reproduce over time. One cell, though, has an advantage: A single mutation somewhere along its genome causes a regulatory protein to bind more strongly to the DNA strand, causing a cascade of effects which enable the cell and its descendants to grow faster than other cells. What is the probability that its descendants will thrive and dominate the population rather than die out, given the change in binding strength? What mutants, with what binding strengths, do we expect to find in the population after millions of years?

Questions similar to these were first considered almost a century ago by the evolutionary geneticists Fischer, Wright, and Haldane, but the limited biophysical knowledge at the time hampered empirical investigation. Even today, despite much progress in understanding "neutral" or non-adaptive evolution, relatively little is known quantitatively about adaptive evolution[4]: We have a poor understanding of the mathematical form of selective forces[5], of the evolutionary dynamics of populations subject to these selective forces[6], and of the nature of phenotypes and novel traits that may evolve over time. This is not for lack of mathematical models, but because the multilocus nature and environmental dependence of most phenotypes makes it difficult to measure the selective benefit of particular genotypes.

However, it may be possible to make progress as a result of the recent emergence of quantitative high-throughput experimental technology, new biophysical knowledge, and the growth of computing power and bioinformatics. In particular new data about living systems at the molecular level may enable greater quantitative understanding of adaptive evolution. At the molecular level, phenotypes are readily quantified and the genotype-phenotype relationship is clearer, and sometimes direct: The phenotype of interest may be a physical property of the genetic molecule (DNA) itself! The selection pressures due to such a phenotype is often easily deduced, and may be incorporated into an evolutionary model.

The inverse is also true: Knowing how evolutionary forces shaped these molecular systems may give insight into their biophysics, and evolutionary data such as phylogenies and sequence alignments can be used to infer biophysical information[7–9]. For example, correlated mutations in the genome can signal the existence of physical interactions in the encoded molecules. The field of "directed evolution" [10–14], in which evolutionary forces are harnessed to create new molecular functions, also benefits from improvements to molecular evolutionary theory.

There has thus been growing interest in uniting biophysical molecular models with classical population genetics. Two particular molecular processes which have received significant recent attention along this line are protein-DNA binding [15–20], and protein folding and stability [21–29]. Proteins, the polypeptide chains which act as the machinery of the cell, are generally only functional if they fold into a specific shape. Similarly, "Transcription Factor" (TF) proteins must bind to their target DNA binding sites to fulfill their role in the cell's gene regulatory network. Both of these functions are necessary for the organisms's surival, and importantly, such selective pressure can be quantified in terms of physical parameters such as the folding energy or binding free energy, which in turn may be parametrized in terms of the underlying protein or DNA sequence making them amenable to quantitative evolutionary analysis. A major goal of such research has been to connect biophysical quantities, such as binding energies, to evolutionary parameters, such as selection strengths.

In this thesis, I contribute to this field first through further development of population genetics in light of biophysics, and second by applying one such evolutionary model to the evolution of TF binding in the yeast *S. cerevisiae*, a model organism.

1.0.1 Thesis Organization and Results

In Chapter 2 I provide a global view of molecular evolutionary models. I describe how two important regimes of evolution, the monomorphic and quasispecies regimes, emerge from many basic models of molecular evolution. I discuss the steady state distribution of molecular phenotypes that arise in each regime given particular selection pressures, and the boundary between the regimes.

In chapter 3, we focus on the particular case of monomorphic evolution in the presence of strong selection, and generalize results previously obtained only in the nearly-neutral limit (when selection is weak or nonexistent) for particular population models. In the neutral limit, it is known that many population genetics models exhibit shared "universal" properties. I discuss the extent of universality outside the nearly neutral limit, and show that it applies even far from neutrality. I also show that the monomorphic steady state derived in Chapter 2 is often valid without a-priori assumptions of near-neutrality.

In chapter 4 we turn to TF binding sites in yeast. Using an evolutionary model for monomorphic populations evolving over a fitness landscape, we infer the fitness landscape as a function of TF-DNA binding energy for a collection of 12 yeast TFs given sequence information, and show that the shape of the predicted fitness function is in broad agreement with a simple thermodynamic model of two-state TF-DNA binding. By correlating TF-DNA binding energies with biological properties of the sites or the genes they regulate, we are able to rule out several scenarios of site-specific selection, under which binding sites of the same TF would experience a spectrum of selection pressures depending on their position in the genome. These findings argue for the existence of universal fitness landscapes which shape evolution of all sites for a given TF, and whose properties are determined in part by the physics of protein-DNA interactions.

Much of this work has been published in [30], or is shortly to be published in [31], and has been developed in collaboration with Alex Morozov and Michael Manhart.

In the remainder of this chapter, I will review relevant aspects of population genetics and biophysical models, and show how these can be united to provide a quantitative model of molecular evolution. The idea of connecting biophysics to evolutionary models was first proposed by Berg at al in 1987 [7], in the context of TF binding site evolution. Following in those steps, I will first describe TF binding site biophysics, which will serve as the biophysical system of interest throughout this thesis.

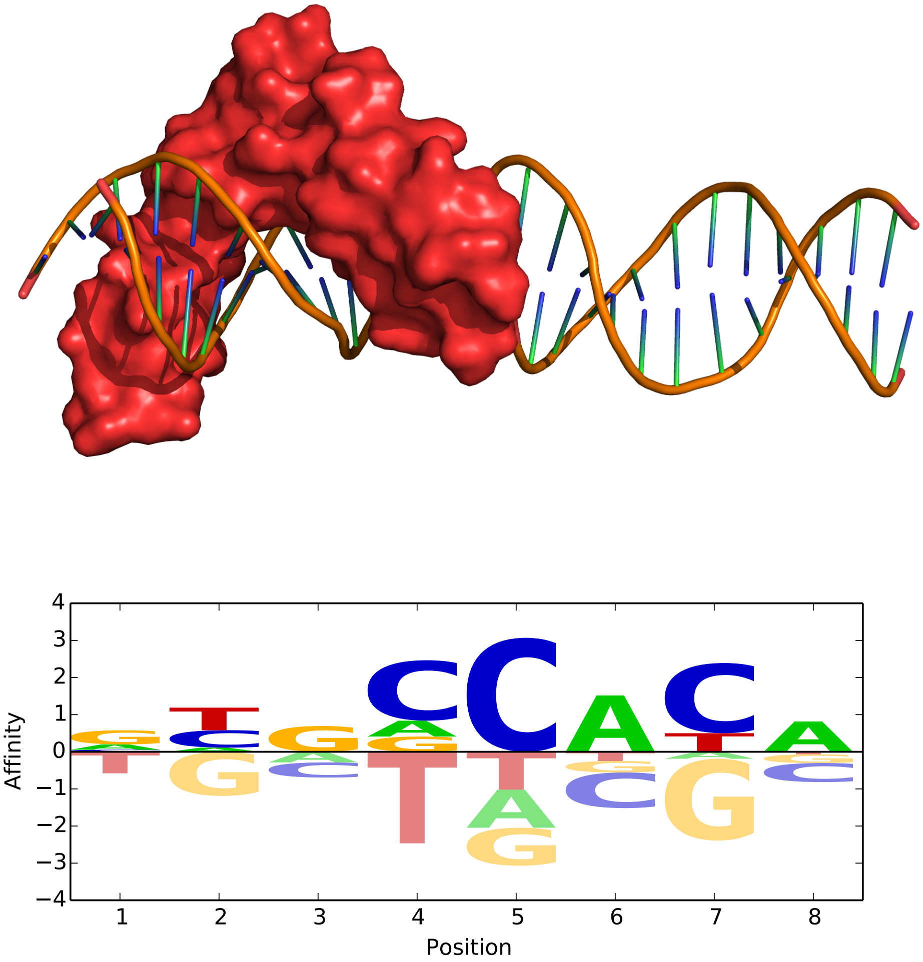


Figure 1.1: (Top) Crystal structure of a homolog of the TF MET32 in *S. cerevissiae* bound to a segment of DNA (PDB-ID 1MEY). The DNA cartoon and "Connolly surface" of the protein are shown. Only the protein's binding site was crystallized, corresponding to about 50% of the full protein. (Bottom) Affinity logo for MET32, from data provided by [1]. The height of each letter corresponds to $\epsilon_i^{\sigma_i}$ in equation 1.2. Nucleotides above the midline have positive affinity (more negative contribution to the binding energy), and vice versa. The consensus sequence is GTGCCACA.

1.1 Transcription Factors and Gene Regulation

Genetic information is stored in the form of genes, which (roughly) are segments of DNA which encode other molecules known as proteins through a 4 letter alphabet of nucleotides “A C G T”. Each gene’s nucleotide sequence is converted by molecular machinery into a protein polypeptide sequence through two processes known as ‘transcription’ and ‘translation’. Proteins act as the main functional entities in the cell, and carry out the many cellular processes which determine our higher level traits.

However, the cells in our bodies produce a variety of different tissues despite despite sharing identical genes, related species sharing many genes in common may vary greatly in morphology, and a single individual’s traits may vary depending on its environment. This can be explained by *gene regulation*: Not all genes are ‘expressed’ at any time, and are instead toggled on and off by the cell through a system known as the gene regulatory network. Special proteins, *transcription factors* (TFs), bind to DNA at specific nucleotide sequences in the promoter regions of genes known as TF binding sites. These promoters are arranged such that TF binding can inhibit or enhance the expression of a nearby gene, often by physically impeding transcriptional machinery. To control the expression of a gene, the cell produces more or less of a TF regulating that gene. TFs often regulate the genes encoding other TFs, thus creating a complex network of regulatory interactions. Understanding TF-mediated regulation is key to understanding the complex regulatory networks within cells — one of the main challenges facing molecular biology.

The TF binding sites are short sequences, typically 10 base pairs (bp) in length in both eukaryotes and prokaryotes but varying from 5 to 30bp [32]. In the eukaryote *S. cerevisiae*, each TF can have 1000s of binding sites in the genome and each gene is regulated by up to about 15 TFs which may interact cooperatively, although we shall largely ignore this cooperativity here [33, 34]. Prokaryotes, in contrast, generally have simpler regulatory networks, and each gene is regulated by a much smaller number of TFs [35]. Here we shall focus on eukaryotes.

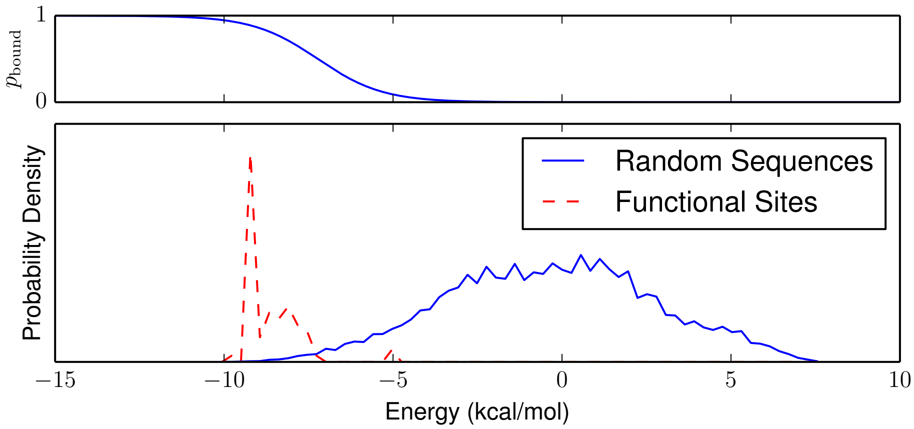


Figure 1.2: Distribution of binding energies for binding sites of the TF MET31 in *S. cerevisiae* based on data provided by [1]. (Top) Hypothetical binding probability as a function of energy (equation 4.1) using biophysical parameters determined in Chapter 3. Here, the chemical potential $\mu = -7.2$ kcal/mol. (Bottom) Red curve: Distribution of binding site energies for MET31 binding sites in the *S. cerevisiae* genome, cataloged by [2]. Blue Curve: Distribution of binding energies for randomly generated sequences. The vast majority of random sequences would be unbound, while the functional site sequences are generally bound. The functional sequences make up a small fraction of all possible sequences, corresponding to the low energy tail of the random site distribution. The energy here is normalized so the mean sequence energy is 0.

1.1.1 TF Biophysics: Two State Kinetics

Each TF binds most strongly to one particular target sequence (often called the “consensus sequence”), but it will also bind to other similar sequences with varying binding affinity. TF binding may be understood physically, and is well approximated by a two state kinetics model, which we shall now describe. Typically, many copies of a TF protein are present in the cytoplasm of the cell and some proportion of them will be bound to binding sites. A TF binding site may be either in a ‘bound’ state with free energy E or in the ‘unbound’ state with a reference free energy of 0. In thermodynamic equilibrium, the Fermi-Dirac function gives the probability that the site is bound to a TF[7]:

$$p_{\text{bound}}(E) = \frac{1}{1 + e^{\beta(E-\mu)}}, \tag{1.1}$$

where $\beta = 1/kT$ is the inverse temperature (≈ 1.7 (kcal/mol)$^{-1}$ at room temperature), and μ is the (concentration-dependent) chemical potential of the TF protein in solution, which appears because there is an entropic cost to removing a TF protein from the cytoplasm of the cell. This function is shown in the top of figure 1.2, for $\mu = -7.2$ kcal/mol. The chemical potential acts as a threshold in equation 4.1: Binding sites with binding energy below the chemical potential tend to be bound, while sites with energies above it are unbound. The chemical potential of TFs is typically on the order of 1-10 kcal/mol [16].

In our an evolutionary analysis, it will be important to quantify the binding strength in terms of the genetic sequence of the binding site. Indeed, the binding energy $E = E(\sigma)$ of a site can be considered a function of its nucleotide sequence, $\sigma = (\sigma_1, \ldots, \sigma_L)$, where L is the length of the site and $\sigma_i \in \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$. While the function $E(s)$ might in principle be a complicated function of genotype, it is often well approximated by a simple "additive energy" model. In this mean-field approximation, each nucleotide makes an additive contribution to the total energy of the site [8]. These contributions are parametrized by an energy matrix (EM), whose $(L \times 4)$ entries $\epsilon_i^{\sigma_i}$ give the contribution to the total energy from the nucleotide σ_i at position i:

$$E(\sigma) = \sum_{i=1}^{L} \epsilon_i^{\sigma_i}. \tag{1.2}$$

EMs can be readily generalized to more complex models of sequence-dependent energetics, such as those with contributions from dinucleotides, or which account for alternate binding modes or nonspecific binding [36–39]. A variety high-throughput techniques have been developed to determine the locations of the binding sites and the specificity (and EMs) of the TFs, including SELEX[40], ChIP-chip, ChIP-seq and DIP-seq[2, 41–43], protein binding microarrays (PBMs)[44], and microfluidics [1, 45, 46]. See [47] for a recent overview of these methods.

The EM encodes information about the specificity of the TF: If one particular sequence is highly favored for binding over others, the $\epsilon_i^{\sigma_i}$ corresponding to that sequence will be much lower, corresponding to a higher TF binding probability. This specificity can be visualized using an affinity logo of the EM, as shown in figure 1.1[37].

Functional binding sites tend to have low binding energy (ie, a high binding affinity), while all other regions of the genome (which consists of effectively random sequences, with respect to TF binding) generally do not bind to the TF. This is illustrated in figure 1.2. The distribution of random sequence energies is approximately Gaussian in the limit of large sequence length, which is a consequence of the central limit theorem as the binding energy may be thought of a the sum of many randomly chosen energetic contributions. This proven more carefully in [15]. The variance of this distribution can be shown to be [15]

$$\chi^2 = \sum_i^L \sum_\alpha p_\alpha (\epsilon_i^\alpha - \bar{\epsilon}_i)^2 \tag{1.3}$$

where ϵ_i^α is the matrix element for base α at position i, $\bar{\epsilon_i} = \sum_\alpha p_\alpha \epsilon_i^\alpha$, and p_α is the background probability of nucleotide α, that is, the probability that a random nucleotide in the genome is α. This distribution of random sequences will become important again in our evolutionary analysis, as it is the "neutral distribution", which we shall note as $\pi_0(E)$: In a neutral evolutionary process without selection, all sequences are equally likely to evolve, and at large times the expected distribution of evolved sequences is simply this random distribution. It is also worth noting the large number of possible genotypes which make up this distribution: For a sequence of length 10, there are 4^{10} or 10^6 possible sequences.

Specificity is much lower in eukaryotes than in prokaryotes, and in eukaryotes a significant number of high affinity but nonfunctional sequences appear in the genome by chance. In figure 1.2, this can be observed through the overlap of the tail of the random sequence distribution and the functional site distribution of energies. In multicellular eukaryotes, a binding site for a TF (with $E < \mu$) is expected to appear by chance approximately once every 4kb, and an estimated $10^4 - 10^6$ spurious binding sites are expected in the genome although only 10^2 to 10^4 may be accessible due to chromatinization[48]. For this reason, scoring sequences using an EM is not enough to know whether a sequence is a functional binding site. Other information, such as cross-species conservation of the site, must be used to verify functionality.

TF binding sites provide an ideal dataset for evolutionary analysis because the set of binding sites in the genome, for a particular TF, can be seen as independent populations evolving in parallel, each binding site realizing one possible evolutionary pathway through time. Because of the effect of recombination (described briefly below) each binding site is evolutionarily unlinked from other binding sites, and the sites are evolving independently under similar selective pressures to bind to the TF, as shall be further justified in Chapter 4.

Why does the distribution of TF binding site energies take the form shown in figure

1.2? A naive observer might expect that evolution would drive all binding sites to the most energetically favorable sequence, yet this is evidently not the case. What determines this variance in energy? To explain this, we turn to population genetics.

1.2 Population Genetics

Quantitative models of evolving populations were first developed as part of the "neo-Darwinian" or "Modern Synthesis" of Wright, Fisher and Haldane starting in the 1930s [49–51], which united long-term, large-scale Darwinian evolution and shorter term, incremental Mendelian genetics. These models describe a population composed of individuals with varied genotypes, which reproduce, mutate and die from generation to generation. As such a population evolves, the fraction of individuals with a particular genotype may grow or shrink as more fit individuals outcompete less fit ones, as individuals die and reproduce stochastically, and as their offpring "mutate", or experience random genetic change. These three processes are known as *natural selection*, *genetic drift*, and *mutation* respectively, and they form the underlying driving forces of evolution.

Wright envisioned that populations would explore a "fitness landscape" in which individuals with higher fitness would outcompete others and the population would gradually move towards the peaks of the landscape[50]. In the case of TF binding sites, binding must occur in order to carry out gene regulation. We may therefore suppose that the fitness of a binding site is proportional to the probability of TF binding, and define the fitness landscape

$$f(E) = (1 - f_0)\frac{1}{1 + e^{\beta(E-\mu)}} + f_0 \tag{1.4}$$

where f_0 represents the fitness cost of not binding, and $f(E)$ then ranges from f_0 to 1. f_0 accounts for the fact that many genes are non-essential, which implies that the fitness cost of non-binding can be small. Natural selection on this fitness landscape drives the population to evolve towards lower energy sequences, which explains why functional sites are biased towards the low energy tail in figure 1.1.

However, in the 1960s the "Neutral" and "Nearly Neutral" theories of evolution were developed by Kimura, Ohta and collaborators, who explained that most genetic change occurs as a result of genetic drift rather than selection[57, 76]. A individual in the population may have a lucky year and reproduce more than others, despite having no fitness advantage or even being slightly unfit. Since a large fraction of possible genetic changes are neutral, new (but phenotypically equivalent) genotypes will arise over time due to these stochastic effects. The widepsread neutrality at the molecular level can be seen in figure 1.1 which shows that many different genotypes share the same energy, for example the degeneracy is highest near $E = 0$. The Neutral Theory of evolution is responsible for some of the most important quantitative evolutionary tools today, including the 'molecular clock', which allows evolutionary history to be inferred from the buildup of neutral mutations, and it provides null models used to test for the existence of selective pressure in DNA and protein sequences.

Genetic drift is the origin of the variance in binding energy seen in figure 1.1. The fact that even deleterious mutants may survive through stochastic effects, combined with the fact that there are many more possible mutants near $E = 0$, means that there is a statistical mutational bias towards sequences with $E = 0$. Thus, we see that the distribution of binding energies in figure 1.1 is the result of two opposing forces: Selection, which drives the population towards low energy, and drift, which drives the population towards the most degenerate phenotypes near $E = 0$. This is often referred to as the "mutation-selection" balance, since in many cases mutations carry the population away from the fitness peaks, opposing the selective pressure.

Other important driving forces which I shall not discuss, but which are frequently considered in population genetics, are *recombination*, which is a mixing of genes which occurs during sexual reproduction, and *migration* or geographic effects, which may limit the spatial spread of genotypes. Sexual organisms are generally *diploid* (contain 2 copies of every gene), as opposed to *haploid* (with only one copy). We shall limit our attention to haploid asexual organisms, which approximately describes many microorganisms. Many of the results discussed here can nevertheless be extended to other cases.

With this example in mind, we shall discuss various quantitative models of such evolving populations and driving forces, for arbitrary fitness landscapes. We begin here with "microscopic" population models, which are the foundation of population genetics modeling.

1.2.1 Wright-Fisher Model

The prototypical population model is the Wright-Fischer model, which describes a population of fixed size N. In this model, the population at generation t is generated by taking N random samples with replacement from generation $t-1$, weighted by the 'fitness' of the individuals. That is, the probability of sampling genotype x from the previous generation for the current generation is

$$p(x) = \frac{1}{Z} n(x) f(x) = \frac{n(x)}{N} \frac{f(x)}{\bar{f}} \tag{1.5}$$

where $f(x)$ is the fitness and $n(x)$ is the number of individuals with genotype x, Z is a normalization constant and $\bar{f} = \sum_x \frac{n(x)}{N} f(x)$ is the mean population fitness. The fitness defined implicitly here can be though of as the logarithm of the growth rate of individuals with genotype x, as the sampling probability implies $n'(x) \propto Np(x) \propto n(x)f(x)$ and therefore $n(x,t) \propto n(x,0)f(x)^t \propto n(x,0)e^{t \log f(x)}$ in the limit that $n(x)$ is small relative to N. That is, a small subpopulation of genotype x will grow exponentially at a rate $\log(f(x))$. The Wright-Fisher model accounts for natural selection through the fitness weighting and for the stochastic fluctuations in $n(x)$ (genetic drift) through the random sampling procedure.

Mutation can then be incorporated by letting each individual mutate to other genotypes with probability u each generation, in a way specified by a separate mutational model. In the case of molecular evolution, we shall assume evolution proceeds by point mutations, and the mutational network of genotypes takes the form of a fully connected hypercube in which each genotype of length L is connected to $4L$ neighbors[53].

A special case of the Wright-Fisher model is the 'two-allele' population where only two genotypes are present, which we will call 'wild type' and 'mutant'. This situation commonly arises in many populations, as shall be described in section 1.2.3. Given that the number of mutants is n at time t, the probability of ending up with n' mutants in generation $t+1$ due to the random sampling procedure is a binomial distribution

$$\Pi(n'|n) = \binom{N}{n'} q^{n'} (1-q)^{N-n'}, \quad \text{where} \quad q \equiv \frac{n}{N} \frac{f_m}{\bar{f}} \tag{1.6}$$

where f_m is the fitness of the mutant type. Iterating this transition probability gives the fraction of mutants over time, and eventually the mutant population will reach an absorbing state either at $n = 0$ (the mutant becomes extinct) or $n = N$ (the mutant 'fixes'), after which only one genotype is present and the population no longer changes, assuming new mutations do not occur. Calculating the fixation probability and fixation time (given an an initial mutant fraction n_0) is not trivial, and is achieved using diffusion theory, described below.

1.2.2 Moran and Cannings Models

In an alternate model, the Moran model, a population of size N is updated through replacement of a single individual at a time. At each step, one individual is chosen to die, and another is chosen to replicate. An individual's fitness will weight the probability for that individual to be chosen for either of these substeps.

In the two-allele case, the single time-step transition probabilities of the Moran model are [54, 55]

$$\begin{aligned} \Pi(n+1|n) =& \frac{f_m}{\bar{f}}\frac{n}{N}\left(1-\frac{n}{N}\right) \\ \Pi(n-1|n) =& \frac{f_w}{\bar{f}}\frac{n}{N}\left(1-\frac{n}{N}\right) \\ \Pi(n|n) =& \ 1-\Pi(n+1|n)-\Pi(n-1|n), \end{aligned} \tag{1.7}$$

for mutant count n. The fixation probability of the mutant is exactly solvable in the Moran model, and can be shown to be [55]

$$\phi(r) = \frac{1 - f_w/f_m}{1-(f_w/f_m)^N} \tag{1.8}$$

Like the Wright Fisher model, the Moran model captures the effects of selection and drift, and is also easily extended to include mutation.

Generalization of the Wright-Fisher and Moran models leads to a class of population models collectively known as the Cannings model. Like the Wright Fisher and Moran models, the population is generated from a random sample of the previous generation for fixed population size N, but the distribution of offspring is arbitrary (as opposed to the binomial distribution of the WF model, for example), subject only to the condition that the population size stays constant and that all individuals (in the neutral case) have equal probabilities of reproducing.

1.2.3 Regimes of Haploid Evolution

These "microscopic" population models give rise to a variety of higher-level dynamics in common. In each model, three regimes emerge: The monomorphic, polymorphic, and quasispecies regimes, which are illustrated in figure 1.3.

The **monomorphic** regime results from very low mutation rates u and small populations N. It has been called the "succesional mutation" regime [56]. Here, mutants are so rare that the population is usually composed of individuals with exactly the same genotype. Rarely, a mutant appears in the population, and its descendants may either go on to fix or go extinct (as described in the two-allele case above), long before another mutant appears. Such fixation events are known as 'selective sweeps'. In this regime, evolution takes the form of a 'substitution process', in which the population jumps as a whole from genotype to genotype. It is believed that many higher eukaryotes [57] and some microorganisms contain loci that can be adequately described as monomorphic [58–61]. As a result, this regime has been applied in settings as diverse as the evolution of transcription factor (TF) binding sites in yeast [19, 20], viral protein evolution [25, 27], and codon usage bias (e.g., [62, 63]).

The **polymorphic** regime occurs for high mutation rates and large populations. Here the population is composed of many competing genotypes, none achieving fixation before new mutants appear. This has also been called the "concurrent mutations" regime[56].

The **quasispecies** regime is the large population limit of the polymorphic regime. For very large populations composed of individuals spanning a huge number of genotypes, stochastic effects and genetic drift become negligible relative to the subpopulation sizes of each genotype, and evolution becomes deterministic. The quasispecies regime was first studied in the context of populations of self-replicating RNA molecules, as might have existed near the origin of life [53]. Virus populations are also often though to evolve in the quasispecies regime[64].

At large times, in all three regimes of evolution, the population reaches a steady state distribution of phenotypes, which we shall write as $\pi(E)$ in the case of the TF phenotype E. In the monomorphic limit, the population jumps back and forth between a small subset of genotypes over time, and the steady state gives the probability of finding the population

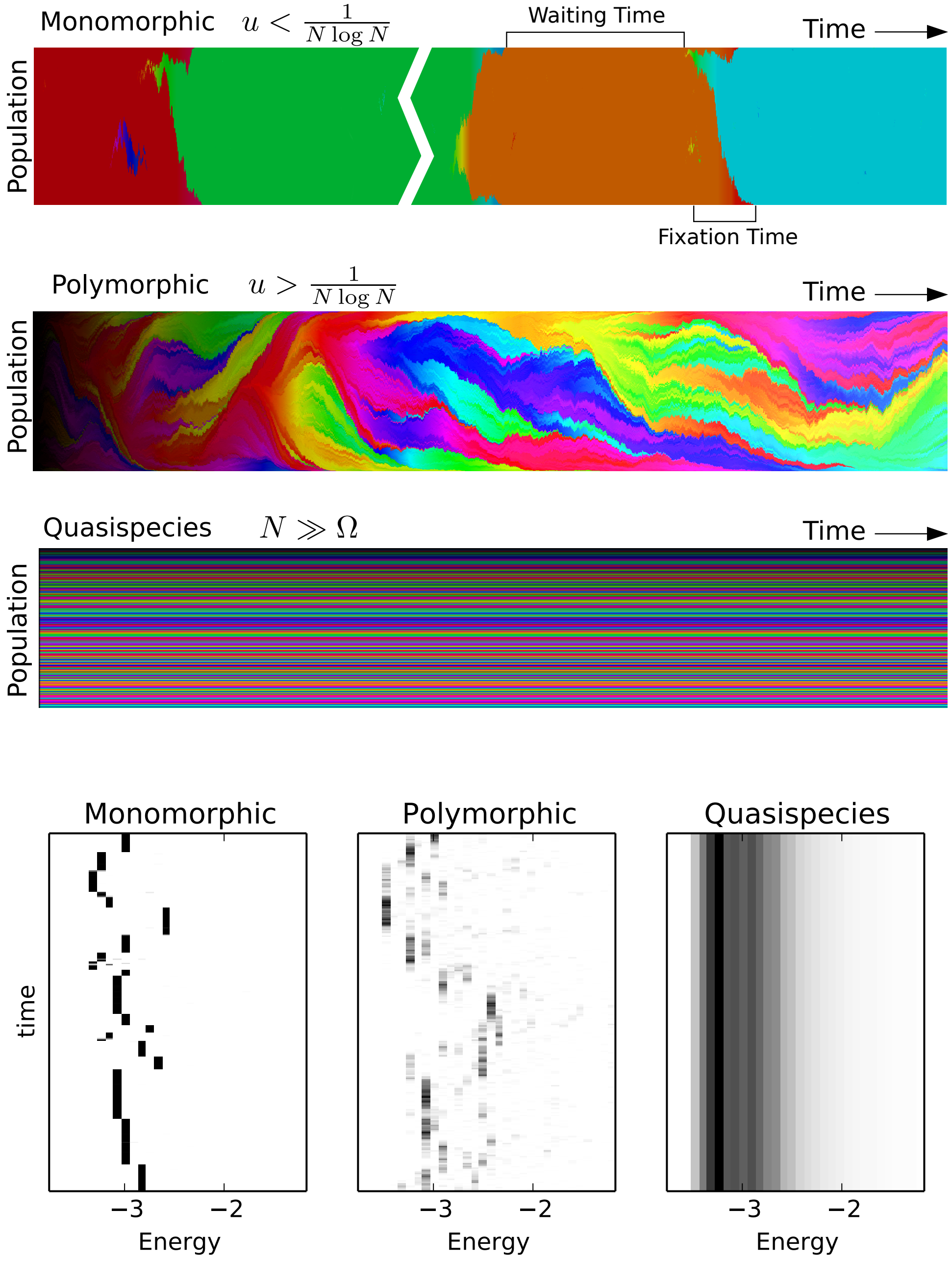


Figure 1.3: Simulations of a Wright-Fischer process in three different regimes. (Top) Here, colors represent different genotypes, brightness correponds to fitness, and time increases towards the right. Each vertical slice of each image represents the distribution of genotypes in the population at one point in time. These regimes are obtained for different conditions on the mutation rate u and population size N, and the size of the genotype space Ω, as discussed in chapter 2. Colors for one genotype are allowed to vary slightly from one generation to the next, for visualization reasons. The break in the monomorphic plot represents a long period in which no mutants appeared. (Bottom) Evolution in energy-space in the three regimes for a fitness function in the form of equation 1.4, showing populations which have reached steady state. Each horizontal slice represents a histogram of the energy distribution at that point in time.

in that state at one point in time. Empirically determining this steady state distribution requires either a time or ensemble average. In the case of TF binding sites, which we will argue evolve in the monomorphic regime, the set of binding sites in the yeast genome provides an ensemble average, and the steady state distribution is simply the distribution shown in figure 1.2. In the quasispecies limit, on the other hand, the distribution of genotypes in the population stays constant over time and no averaging is necessary, but rather many samples from the same population are required.

In Chapter 2 we shall see how, given information about the genotype-pheotype mapping (ie, equations 1.4, 1.2 and an EM), using population genetics one can compute the expected steady state distribution of phenotypes in each regime given a fitness landscape, and vice versa. In the case of TF binding sites, we can measure the steady state distribution from binding site in the genome, but we also know the expected fitness landscape 1.4 from biophysical modeling. We can therefore test whether they are consistent with each other, as shall be discussed in Chapter 4.

This idea of using biophysical models of TFs to predict evolutionary properties was considered using a quasispecies model in 2002 by Gerland and Hwa [16], and independently by Sengupta, Djordjevic and Shraiman [15]. Starting in 2003 Berg and Lässig studied TF evolution using a monomorphic model [17], followed by a number of related studies[18–20] which aim to infer a fitness landscape from sequence data. The work on TF binding site evolution presented in this thesis builds on these studies.

1.2.4 Universality

Before we can embark on this program, there seems to be a problem: From among the variety of microscopic population models described above, which of these models accurately describes real populations? Each model may predict a different steady state distribution. Remarkably, however, in many cases all of these models converge to a single unified quantitative model in what is known as the *diffusion limit*, which gives formulae for the fixation probability and fixation time which are independent of the underlying population model and mutational model[65, 66], for a large class of models [65, 67–72]. The Wright-Fisher and Moran models are no doubt gross simplifications of true population dynamics, but universality in the diffusion limit reassures us that predictions based on population genetics theory may apply even in complex, unknown populations. Understanding which properties are universal will be a key initial step towards a general theory of adaptive evolution. Universality is the subject of Chapter 3, where we will also show show that many universal results obtained in the diffusion limit apply even far outside this limit.

Here, we shall introduce the diffusion limit and begin to derive a simple but illustrative (and universal), result: The fixation probability of a single mutant in a population otherwise composed of many identical individuals, corresponding to the two-allele case described above.

1.2.5 Diffusion Theory

Diffusion theory has been used extensively in population genetics starting with Wright, but the most significant progress is due to Kimura [73, 74]. Besides leading to a universal picture of population genetics, diffusion theory also makes many computations mathematically tractable, such as the aforementioned fixation probability in the Wright-Fisher two-allele model.

Here we focus on the two-allele case, and start from discrete population dynamics. For any population model described above, let us model the time evolution of the number of mutants n in a population of size N, starting from an initial mutant population n_0. It is convenient to work in terms of the population fraction of mutants $x = n/N$, rather than n itself. To trace the growth of the mutant population, we seek to calculate the population fraction x after t generations, denoted $\theta(x, x_0; t)$, starting from a fraction $x_0 = n_0/N$.

We start by writing the 'retrospective' update equation in the discrete case,

$$\theta(x, x_0; t+1) = \sum_{\delta_x} \phi(x, x_0 + \delta_x, t)\Pi(x, \delta_x) \tag{1.9}$$

where $\Pi(x, \delta_x)$ is the (model-dependent) probability of a jump in population fraction by δ_x given that the mutant fraction is x. $\Pi(x, \delta_x)$ corresponds to $\Pi(n'|n)$ defined above. In words, this equation says that we can find the probability of having a population fraction x at time $t+1$, given the probability of having a fraction x at time t after starting from intermediate initial positions $x_0+\delta_x$ in one generation less, multiplied by the probability of jumping from the starting point x_0 to these intermediate starting points in one generation.

In the diffusion approach, we take the continuum limit of this update equation in time and population fraction, giving

$$\theta(x, x_0; t+\delta t) = \int \theta(x, x_0+\delta_x, t)\Pi(x, \delta_x, \delta t) d\delta_x \tag{1.10}$$

where the time step is no longer 1 generation, but an infinitesimal parameter δt.

If we now treat x as a continuous variable (taking the infinite population limit) and expand $\phi(x, x_0+\delta_x, x, t)$ in the integral in terms of δ_x to second order, we obtain

$$\theta(x, x_0; t+\delta t) = \int \Pi(x, \delta_x, \delta t)\left(\theta(x, x_0, t) + \frac{\partial\theta(x, x_0, t)}{\partial x_0}\delta_x + \frac{\partial^2\theta(x, x_0, t)}{\partial^2 x_0}\delta_x^2\right) d\delta_x \tag{1.11}$$

$$\theta(x, x_0; t+\delta t) = \theta(x, x_0, t) \tag{1.12}$$

$$+\frac{\partial\theta(x, x_0, t)}{\partial x_0}\left(\int \Pi(x, \delta_x, \delta t)\delta_x d\delta_x\right) \tag{1.13}$$

$$+\frac{\partial^2\theta(x, x_0, t)}{\partial^2 x_0}\left(\int \Pi(x, \delta_x, \delta t)\delta_x^2 d\delta_x\right) \tag{1.14}$$

which is then easily rearranged into the “backwards Kolmogorov equation”,

$$\frac{\partial\theta(x, x_0; t+\delta t)}{\partial t} = M(x)\frac{\partial\theta(x, x_0, t)}{\partial x_0} + \frac{V(x)}{2}\frac{\partial^2\theta(x, x_0, t)}{\partial^2 x_0} \tag{1.15}$$

where the moments $M(x)$ and $V(x)$ are given by

$$M(x) = \lim_{\delta t\to 0}\frac{1}{\delta t}\int \Pi(x, \delta_x, \delta t)\delta_x d\delta_x \tag{1.16}$$

$$V(x) = \lim_{\delta t\to 0}\frac{1}{\delta t}\int \Pi(x, \delta_x, \delta t)\delta_x^2 d\delta_x \tag{1.17}$$

We can now solve for $\theta(x, x_0, t)$ and many related quantities by solving the Kolmogorov equation. Note that $M(x)$ and $V(x)$ are simply the moments (the mean and variance) of the Π distribution, with some scaling for time.

Remarkably, the Wright-Fisher, Moran, and generalized Cannings models have the same moments, up to scaling factors, and thus they obey the same diffusion equation, 1.15. For the Wright-Fisher process, these moments are easily found from $\Pi(n'|n)$ to be

$$M(x) = sx(1-x) \tag{1.18}$$

$$V(x) = \frac{x(1-x)}{N} \tag{1.19}$$

up to $\mathcal{O}(s)$,where $s = \frac{f_m}{f_w} - 1$ is the 'selection coefficient' for the mutant relative to the population. s ranges from -1 to ∞ and $s = 0$ in the neutral case. This choice of moments is often called the "WF diffusion limit". The moments of the Moran model are the same but multiplied by a factor of two, which may be absorbed in equation 1.15 by rescaling the time. Thus, through diffusion theory the universality of population genetics models becomes apparent.

Manipulation of equation 1.15 leads to the s-dependent fixation probability and fixation time. More complete derivations are given in appendix A, here we shall simple quote the memorable results:

- The probability that a single mutant in an otherwise homogeneous population of size N fixes is
$$\phi(s) = \frac{1 - e^{-s}}{1 - e^{-Ns}} \tag{1.20}$$
In the neutral limit when $s \to 0$, this fixation probability is $\phi = 1/N$.
- The extinction time (given that the single initial mutant becomes extinct) in the neutral case is (in number of generations)
$$\bar{t}_{\text{ext}} = 2 \log N \tag{1.21}$$
- The fixation time (given that the single initial mutant fixes) in the neutral case is (in number of generations)
$$\bar{t}_{\text{fix}} = 2N \tag{1.22}$$

The full s-dependent fixation times, too complex to quote here, are plotted in figure 1.4. The interplay of genetic drift and selection are apparent: As shown in the inset, even deleterious mutation can fix because of genetic drift. For large (positive and negative) s both fixation and extinction happen quickly due to selection pressure, but they occur much more slowly in the neutral case when only genetic drift is present.

Note that the moments used here were only obtained to $\mathcal{O}(s)$, the nearly neutral limit. This is why the exact Moran fixation probability, equation 1.8, appears slightly different from equation 1.20: In the small s limit, the they are the same.

It is important to understand the nature of the approximations involved in the diffusion limit. By taking the infinite population limit we have assumed that the change in population fraction of the mutant, in some small unit of time, is infinitesimal. The diffusion limit will be a poor approximation anytime the population fraction may jump by a large amount in one generation. This can occur when the population size is very small, or when the selection coefficient is far from 0. Thus, even without the $\mathcal{O}(s)$ cutoff used in calculating the moments, diffusion theory should be understood to be valid mainly in the neutral limit.

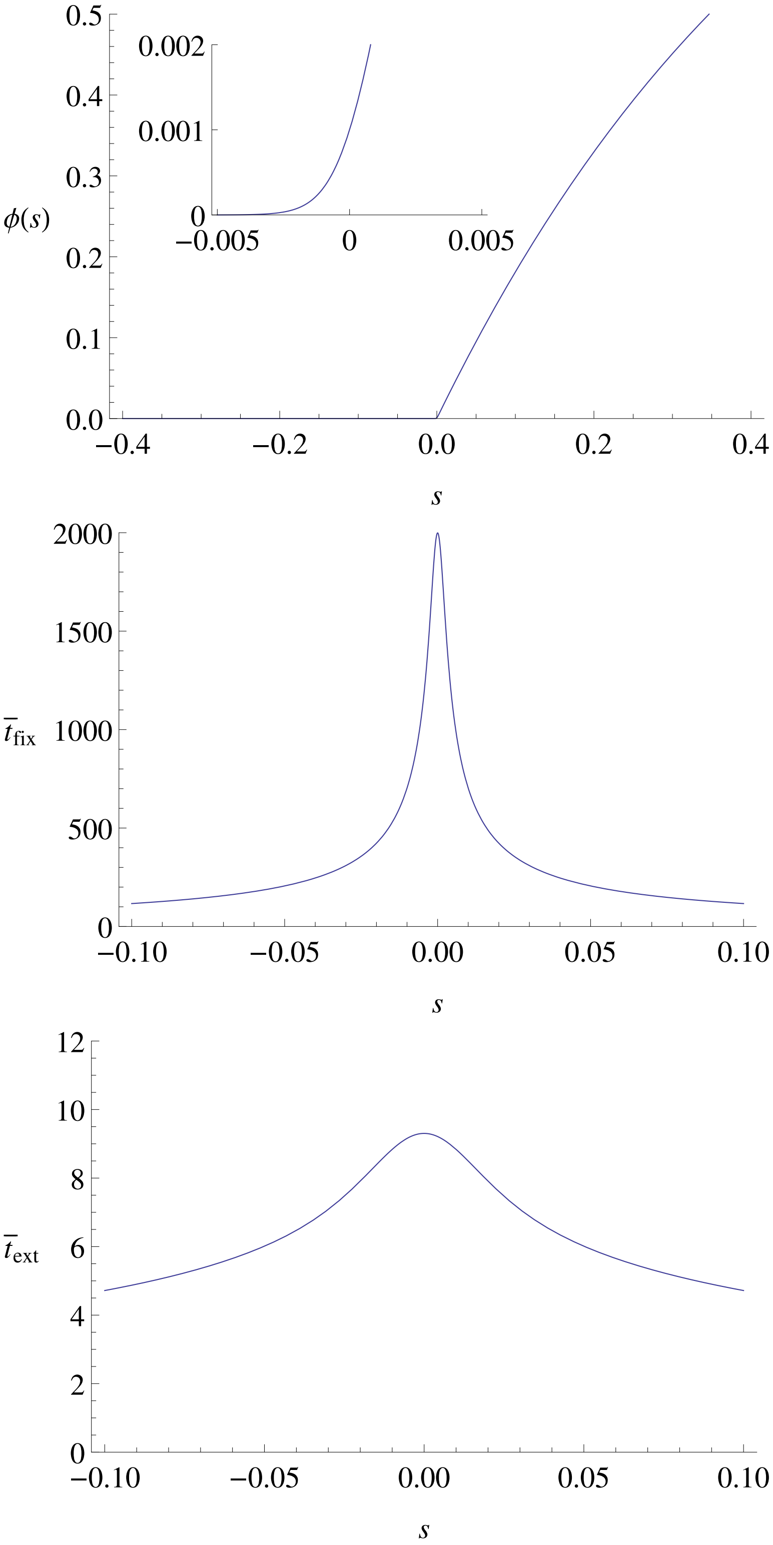


Figure 1.4: Properties of the two allele system with one initial mutant in the diffusion limit, for $N = 1000$. (Top) The Fixation Probability of the mutant. The region near $s = 0$ is inset. (Middle) The fixation time, assuming the mutant fixed. Very beneficial mutants (high s) fix quickly, but so do very deleterious mutants if they fix at all: The very few deleterious mutants which do fix must do so through large and immediate stochastic fluctuations, before they are driven to extinction by natural selection. (Bottom) The extinction time, assuming the mutant became extinct.

Chapter 2

Regimes of Haploid Evolution

In this short chapter we shall give a more quantitative analysis of the population genetics described in Chapter 1, and derive the steady state distribution of phenotypes in the monomorphic and quasispecies regimes in the diffusion limit. These results in the monomorphic regime will be used in chapters 3 and 4. Previously, the steady has been derived for particular population models, here we extend these results, and give a simple derivation in the monomorphic case showing that it applies to all 'reversible' population models. This derivation shows that the monomorphic steady state we derive is "universal", as shall be further described in Chapter 3.

2.0.6 Phenomenology

We shall begin with a phenomenological picture of asexual haploid evolution. In figure 2.1, we show the results of a large set of simulations of a Wright-Fisher process for varying population size N and mutation rate u, with a biophysically inspired fitness function from equation 1.4 with the energy phenotype and mutational model as described in Chapter 1. As can be seen from figure 2.1B, the monomorphic regime, in which the population contains only one genotype, occurs for low mutation rates and population sizes, while the polymorphic regime occurs in the opposite conditions. A boundary, which we shall derive shortly, is superimposed in white.

In figure 2.1 A we show the mean fitness in steady state for each pair of parameters. In these simulations, high fitness generally means low energy, as illustrated in figure 2.1 C. A number of interesting features appear:

- The mean fitness in the monomorphic regime is u-independent. Although it is not apparent from the plot, the steady state distribution itself is also u-independent.

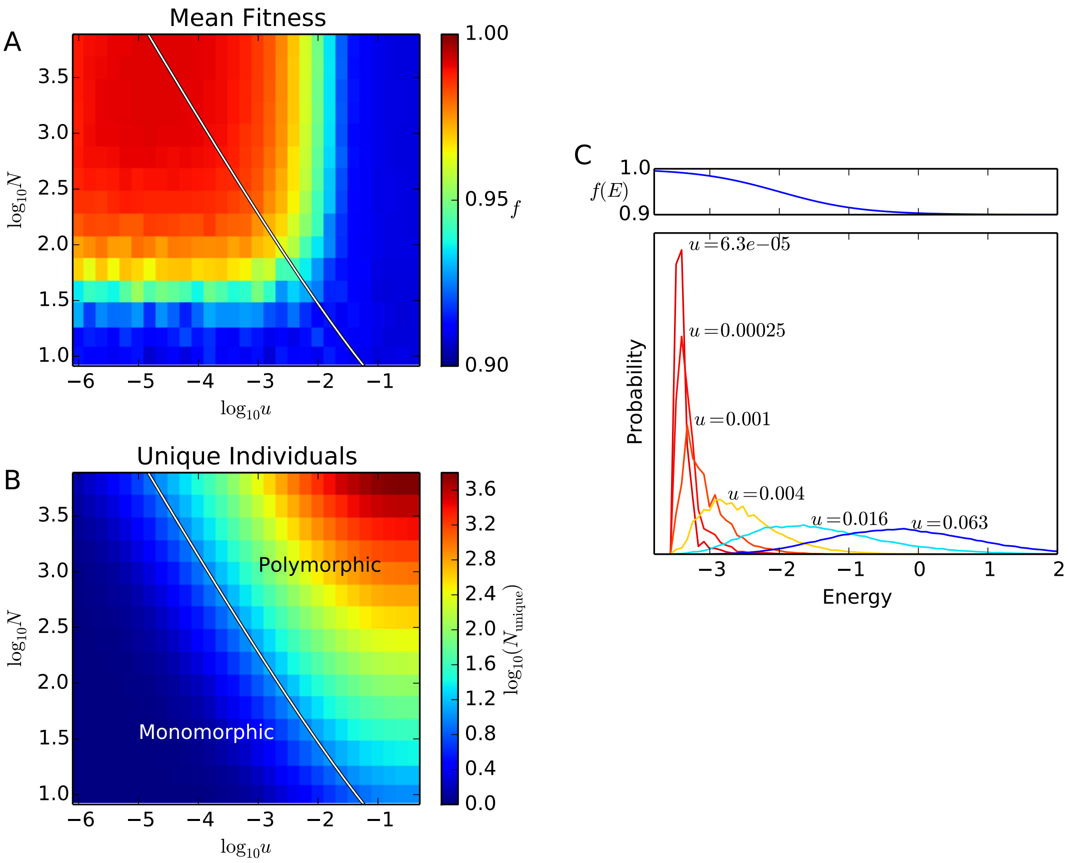


Figure 2.1: For varying population size N and mutation rate u, results from simulations of a Wright-Fisher process, evolved until steady state is reached, under a Fermi-Dirac fitness function following 1.4 with $f_0 = 0.9$, $\mu = -2$, $\beta = 1.688$. (A) The mean population fitness in steady state. The population is largely unfit when the mutation rate is high or the population size is low, as explained in the text. The white line shows the theoretical boundary between the monomorphic and polymorphic regimes, equation 2.26 (B) For the same simulations, the mean number of unique genotypes in the population. (C) For $N = 1000$, steady state distributions for various u. The fitness function is shown above. Color indicates the mean fitness of the population, corresponding to the color in panel A, and the steady states shown here correspond to points along the line $\log_{10} N = 3$ in panel A.

- Conversely, the steady state in the polymorphic regime is N-independent
- An 'error catastrophe' occurs at very high u. Random (unfit) mutants accumulate faster than selection can weed them out, and they dominate the population. This effect has been widely discussed in quasispecies literature[53, 75].
- A 'fluctuation catastrophe' occurs at very low N. Stochastic fluctuations become very large, and one of the many possible unfit genotypes becomes likely to take over the population by chance.

Because of the first two properties, the monomorphic and quasispecies limits are the two important cases of asexual evolution: Knowing the steady state in the monomorphic limit (as $u \to 0$) gives the steady state everywhere in the monomorphic regime. Likewise, the steady state in the quasispecies limit (as $N \to \infty$) gives the steady state everywhere in the polymorphic regime. Thus the steady state can be predicted for nearly all parameter combinations, except near the boundary between regimes, where either limit may still provide a good approximation.

We shall now describe, in turn, the monomorphic steady state, the boundary between the monomorphic and polymorphic regimes, and the quasispecies steady state.

2.1 Monomorphic Evolution

2.1.1 The Substitution Process

In the monomorphic limit, the mutation rate is sufficiently low that the vast majority of single mutations either fix or become extinct before a second mutation on the locus arises [57]. Thus we can describe evolution of this population as a series of substitution events in which the entire population switches from genotype σ to genotype σ', known as the "substitution process". As shown in figure 1.3, two timescales are important: The waiting time until a new mutant appears, and the fixation or extinction time of a new mutant. The waiting time is $1/uN$ generations on average, as uN is the probability of a mutant appearing per generation in the population of size N. In the monomorphic limit, $u \to 0$, the u-independent fixation time thus becomes negligible relative to the waiting time, and we therefore approximate fixation as instantaneous.

Let σ and σ' be two genotypes (sequences of L nucleotides or amino acids) at the locus of interest. The substitution rate from σ to σ' can be approximated by the rate of producing a single mutant times the probability that the mutation fixes [57, 76]:

$$W(\sigma'|\sigma) \approx N\mu(\sigma'|\sigma) \cdot \phi(\sigma'|\sigma), \tag{2.1}$$

where N is an effective population size, $\mu(\sigma'|\sigma)$ is the single nucleotide or amino acid mutation rate from σ to σ' (not to be confused with the chemical potential in TF biophysics), which in the simplest case is related to the per-locus mutation rate as $\mu = u/L$, and $\phi(\sigma'|\sigma)$ is the probability that a single σ' mutant fixes in a population of wild-type σ. We will assume that μ is nonzero only for genotypes σ and σ' differing by a single nucleotide or amino acid.

Given an ensemble of populations evolving with these rates, we can define $\pi(\sigma, t)$ to be the probability that a population is monomorphic at the locus with genotype σ at time t. This probability evolves over time via the master equation

$$\frac{d}{dt}\pi(\sigma', t) = \sum_{\sigma \in \mathcal{S}} [W(\sigma'|\sigma)\ \pi(\sigma, t) - W(\sigma|\sigma')\ \pi(\sigma', t)], \tag{2.2}$$

where $\mathcal{S}$ is the set of all possible genotypes at the locus of interest. 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 $\tilde{\pi}(\sigma)$ [77] satisfying

$$\sum_{\sigma \in \mathcal{S}} [W(\sigma'|\sigma)\ \tilde{\pi}(\sigma) - W(\sigma|\sigma')\ \tilde{\pi}(\sigma')] = 0. \tag{2.3}$$

The form of this steady-state distribution depends on the underlying population genetics model that gives the fixation probability ϕ. We note that the steady state is independent of u, as observed above, as it merely rescales W. That is, it may affect the time to reach equilibrium, but not the equilibrium itself.

2.1.2 The Steady State of the Substitution Process

In order to solve for the steady state in equation 2.3, we now make an important assumption: That in steady state the probability of transitioning from genotype σ to genotype σ' is equal the the probability of the reverse transition. More formally, we assume that the steady state satisfies detailed balance, also known as time reversibility (or simply 'reversibility'), such that

$$W(\sigma'|\sigma)\ \tilde{\pi}(\sigma) = W(\sigma|\sigma')\ \tilde{\pi}(\sigma'), \tag{2.4}$$

where $\tilde{\pi}(\sigma)$ denotes the steady-state distribution. 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.

As we shall see in Chapter 3, the assumption that the underlying microscopic population model is reversible in steady state is not entirely justified. However, the Moran model is known to be reversible, and in the diffusion limit (ie, in nearly-neutral evolution) population models generally will also be reversible [55, 113, 114]. As shall be described in Chapter 3, however, this assumption is well supported in most cases.

We will also 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. 2.1 shows that the neutral substitution rates are just the mutation rates [57]: $W(\sigma'|\sigma) = \mu(\sigma'|\sigma)$. Let the steady-state distribution of the neutral substitution process be $\tilde{\pi}_0(\sigma)$. Then reversibility of the neutral model is expressed by

$$\mu(\sigma'|\sigma)\ \tilde{\pi}_0(\sigma) = \mu(\sigma|\sigma')\ \tilde{\pi}_0(\sigma'). \tag{2.5}$$

Many popular neutral mutational models are reversible (see [78] for a summary), although this condition is not guaranteed. 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. However, reversible mutation 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., [78–83]). Moreover, [84] 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 (e.g., [85]).

Under the assumptions of reversibility of both the mutational model and more generally the population model, let us proceed. The condition of reversibility implies that

$$\begin{aligned}\frac{\tilde{\pi}(\sigma')}{\tilde{\pi}(\sigma)} &= \frac{W(\sigma'|\sigma)}{W(\sigma|\sigma')} && (2.6)\\ &= \frac{\mu(\sigma'|\sigma)}{\mu(\sigma|\sigma')} \cdot \frac{\phi\left(\frac{f(\sigma')}{f(\sigma)}\right)}{\phi\left(\frac{f(\sigma)}{f(\sigma')}\right)} && (2.7)\\ &= \frac{\tilde{\pi}_0(\sigma')}{\tilde{\pi}_0(\sigma)} \cdot \psi\left(\frac{f(\sigma')}{f(\sigma)}\right) && (2.8)\end{aligned}$$

where we have invoked the reversibility of the neutral rates (Eq. 2.5) and we have defined a new function

$$\psi(r) \equiv \frac{\phi(r)}{\phi(1/r)}. \tag{2.9}$$

with $r = \frac{f(\sigma)}{f(\sigma')}$. By substituting this relation into the trivial identity $(\tilde{\pi}(\sigma'')/\tilde{\pi}(\sigma')) \cdot (\tilde{\pi}(\sigma')/\tilde{\pi}(\sigma)) = \tilde{\pi}(\sigma'')/\tilde{\pi}(\sigma)$, it follows that

$$\psi\left(\frac{f(\sigma'')}{f(\sigma')}\right) \cdot \psi\left(\frac{f(\sigma')}{f(\sigma)}\right) = \psi\left(\frac{f(\sigma'')}{f(\sigma)}\right), \tag{2.10}$$

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:

$$\psi(r) = r^{\nu}, \tag{2.11}$$

for some constant ν [86]. The constant ν can only depend on the population size N, since that is the only other parameter in our population model. Now rewriting Eq. 2.8 with our explicit form of ψ,

$$\frac{\tilde{\pi}(\sigma')}{\tilde{\pi}(\sigma)} = \frac{\tilde{\pi}_0(\sigma')}{\tilde{\pi}_0(\sigma)} \left(\frac{f(\sigma')}{f(\sigma)}\right)^{\nu}, \tag{2.12}$$

we can deduce the steady state:

$$\tilde{\pi}(\sigma) = \frac{1}{Z}\ \tilde{\pi}_0(\sigma)\ (f(\sigma))^\nu, \tag{2.13}$$

where Z is a constant chosen for normalization. We finally project this equation from genotypes to phenotypes, by summing over all genotypes with shared phenotype:

$$\sum_{\substack{\sigma \\ \mathcal{P}(\sigma)=E}} \tilde{\pi}(\sigma) = \sum_{\substack{\sigma \\ \mathcal{P}(\sigma)=E}} \frac{1}{Z}\tilde{\pi}_0(\sigma) f(\sigma)^\nu \tag{2.14}$$

$$\tilde{\pi}(E) = \frac{1}{Z}\tilde{\pi}_0(E) f(E)^\nu \tag{2.15}$$

This is the promised relationship from Chapter 1, and the goal of this section: A relationship between the steady state distribution $\tilde{\pi}(E)$ and the fitness landscape $f(E)$.

This steady-state formula was derived in the special case of the Moran model by [87]. We generalize this earlier result by showing that reversibility implies the power law for ψ and the steady-state formula of Eq. 2.13, and thus the steady-state behavior of *any* reversible substitution process, not just the Moran model as studied in [87], is given by Eq. 2.13. Note that this result, derived in the monomorphic limit, requires no additional assumptions, such as the weak-selection diffusion approximation.

2.1.3 ν is an Effective Population Size

Before discussing the consequences of this relation, there is one more issue: The value of ν. While here we have derived the steady state without any reference to the form of the fixation probability, had we done so in equation 2.7, we would find that $\nu = 2(N-1)$ if we substitute the WF diffusion limit fixation probability (equation 1.20), or $\nu = N - 1$ if we substitute the exact Moran model fixation probability (equation 1.8). In these models, ν appears to be linearly related to the census population size N.

Here we shall give an argument that ν is always linear in N in the diffusion limit, and is a “scaling” effective population size that is of the same order as the census population size for fixed-size models. We also show that it must be nonzero, and therefore the steady state formula always applies in the diffusion limit. Using the definition $\psi(r) = \phi(r)/\phi(1/r)$, one can show that

$$\nu = \frac{2\phi'(1)}{\phi(1)} = 2N\phi'(1), \tag{2.16}$$

where $\phi'(1) = d\phi(r)/dr|_{r=1}$ and the neutral fixation probability must be $\phi(1) = 1/N$. The fixation probability in the diffusion limit is given by [73] (see appendix A)

$$\phi(r) = \frac{\int_0^{1/N} dx\ G(x,r)}{\int_0^1 dx\ G(x,r)}, \quad G(x,r) = \exp\left(-2\int_0^x dy\ \frac{M(y,r)}{V(y,r)}\right), \tag{2.17}$$

where $M(x,r)$ and $V(x,r)$ are the first two moments of the change in mutant fraction x per unit time. Here we shall expand the moments in terms of r,

$$\begin{aligned} M(x,r) &= M_0(x) + (r-1)M_1(x) + \mathcal{O}((r-1)^2) \\ V(x,r) &= V_0(x) + (r-1)V_1(x) + \mathcal{O}((r-1)^2). \end{aligned} \tag{2.18}$$

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. 2.17 and expanding to lowest order in $r-1$, we obtain

$$\begin{aligned} \phi(r) = \frac{1}{N} + 2(r-1)\Bigg(\frac{1}{N}\int_0^1 dx \int_0^x dy\ \frac{M_1(y)}{V_0(y)} \\ - \int_0^{1/N} dx \int_0^x dy\ \frac{M_1(y)}{V_0(y)}\Bigg) + \mathcal{O}((r-1)^2). \end{aligned} \tag{2.19}$$

Therefore

$$\phi'(1) = 2\left(\frac{1}{N}\int_0^1 dx \int_0^x dy\ \frac{M_1(y)}{V_0(y)} - \int_0^{1/N} dx \int_0^x dy\ \frac{M_1(y)}{V_0(y)}\right), \tag{2.20}$$

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. 2.20, $\phi'(1)$ 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. 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 [88], 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

$$\frac{d}{dt}\mathbb{E}[X] = \mathbb{E}[M(X,r)], \tag{2.21}$$

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

$$\frac{d}{dt}\mathbb{E}[X] \approx (r-1)\mathbb{E}[M_1(X)] \propto (r-1)\mathbb{E}[X], \tag{2.22}$$

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.

Through equation 2.20 we can now understand the linear relationship between ν and N. 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. 2.20 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 [55, 89].

2.1.4 Discussion

The steady state 2.13 has a clear analogy in statistical mechanics, as it can be written in the form of the Boltzmann distribution[87, 90]:

$$\pi(E) = \frac{1}{Z}\pi_0(E)e^{-\nu \log 1/f(E)} \qquad p_i = \frac{1}{Z}g_i e^{-\beta V_i} \tag{2.23}$$

where on the right we have quoted the thermodynamic Boltzmann distribution, and one sees that ν plays the role of the inverse temperature β, π_0 plays the role of the degeneracy or entropy g, and $\log 1/f$ plays the role of the potential V. In the 'mutation-selection' balance, the opposing driving forces of natural selection versus drift are thus analogous to the opposing driving forces of energy and entropy in many physical systems, and the steady state distribution of phenotypes may be thought of as an equilibrium distribution. Just as in thermodynamic equilibrium, this distribution satisfies detailed balance, it is independent of the underlying dynamics, and the microstates are well-mixed within the macrostates. That is, the equation is valid for any underlying (reversible) mutational model and for any arrangement of mutational connectivity of the genotype space, and the sequences which share the same phenotype have relative probabilities in proportion to their neutral probabilities.

One can see that the population size ν tunes the strength of the fitness $f(E)$ relative to stochastic forces, thereby affecting the strength of the 'selection' driving force. One can also see that the steady state is mutation rate independent, as observed in the introductory phenomenology of this chapter.

All of these properties will serve as a contrast to the quasispecies steady state, which is not in equilibrium.

An appealing aspect of the steady state equation is that it may be inverted, in order to infer a fitness landscape in terms of the steady state distribution of phenotypes, which may be measured experimentally. That is we may write [19, 20]

$$f(\sigma) = \left(Z \frac{\tilde{\pi}(\sigma)}{\tilde{\pi}_0(\sigma)} \right)^{1/\nu} \tag{2.24}$$

Such inference will be the subject of chapter 4.

Here we will note one complication in the practical application of equation 2.24: Calculating the neutral distribution $\tilde{\pi}_0(E)$ for a particular TF (and EM). In chapter 1 we noted that this distribution is approximately Gaussian in the limit of large L, however, in many cases such as TF binding sites the sequences are quite short. Thus, $\tilde{\pi}_0(E)$ must be calculated by explicitly binning the energies of all possible 4^L sequences, thus invoking the genotype-phenotype mapping of the additive energy model. For short enough sequences ($L \lesssim 15$ for a computer in 2013) explicit computation of $\tilde{\pi}_0$ is feasible. For longer sequences we have also developed an approximate algorithm to calculate $\tilde{\pi}_0(E)$ quite accurately and efficiently, which shall not be described here but is available upon request.

Given $\tilde{\pi}_0(E)$ and the distribution of site energies $\tilde{\pi}(E)$ (for which there is no computational obstacle) it is then straightforward to compute $f(E)$ using equation 2.24.

2.2 The Boundary between Monomorphic and Polymorphic Regimes

We now turn to the boundary between the monomorphic and polymorphic regimes. In the monomorphic regime in the neutral limit, the expected fixation/extinction time (number of generations for a single initial mutant to either fix or go extinct) is $2 \log N$ generations: The mutant may fix with probability $1/N$ in time $2N$ (equation 1.21), or it may become extinct with probability $(N-1)/N$ in time $2 \log N$ (equation 1.22), and therefore the net absorption time is

$$\frac{1}{N}(2N) + \frac{N-1}{N}(2 \log N) \propto 2 \log N \tag{2.25}$$

in the large N limit. However, the waiting time, as discussed above, is $1/uN$, and the monomorphic regime occurs as no new mutants occur while a previous mutant is still fixing, which leads to the condition $1/uN < 2 \log N$. Dropping the factor of two for simplicity, we get that the boundary between the regimes must be around

$$u < \frac{1}{N \log N}. \tag{2.26}$$

This result has been obtained by other means by [91] and [92].

In the non-neutral case this boundary gains a complicated dependence on the relative fitness of the mutants, which has not been previously well described. If strong selection is present, for example, mutants may fix more quickly, which decreases the fixation time and may push the population into the monomorphic regime. In appendix B we show that in the non-neutral case, beneficial mutants will generally fix or become extinct in between $2 \log N$ and $4 \log N$ generations, while deleterious mutants fix in less than $2 \log N$ generations. In steady state (which involves a balance of beneficial and deleterious mutants), the deleterious mutants dominate the average, and we argue that the fixation/extinction time in this case will typically be slightly less than the neutral $2 \log N$ result, in a way that depends on the exact distribution of mutant fitnesses. Thus, even with selection, in steady state we expect the boundary of equation 2.26 to be a good approximation of the true boundary between the regimes.

2.3 Quasispecies Evolution

In contrast to the monomorphic case, in which the population as a whole jumps in phenotype space, in the quasispecies regime the population is composed of many different individuals with varied phenotypes and fitnesses. At least at the genotype level, the time-evolution of this population may be modeled as a Master equation, however unlike the monomorphic time-evolution master equation (equation 2.2), which describes the time evolution of a probability density, here we must model the evolution of the population density itself.

One important consequence of this is the presence of 'sources' and 'sinks' in the quasispecies case: Subpopulations with low fitness will tend to die, while those with high fitness grow, and thus there may be net mutational currents within the population from high fitness to low. In the monomorphic case, in contrast, the population never dies, and there are no absorbing states. Unlike the monomorphic case, a population in the quasispecies regime may reach steady state but it does not reach equilibrium, and many of the important properties of equilibrium do not apply.

As the quasispecies regime is not the focus of this thesis, we shall only give a brief description of the steady state. In traditional quasispecies theory, first developed by [53], one writes a Master equation for the distribution of genotypes

$$\vec{x}(t+1) = M \cdot \vec{x}(t) \tag{2.27}$$

where $\vec{x}$ is a vector representing the distribution of genotypes, and M is a transition matrix. Off diagonal elements in M represent mutation rates from one genotype to another, and M will typically be sparse if we consider genetic sequences evolving through point mutations, as each genotype can only mutate to $4L$ other genotypes. The diagonal elements of M represent growth rates, minus the overall mutation rate away from that genotype. The eponymous 'quasispecies' refers to the eigenvectors of this matrix, which represent combinations of genotypes that grow together at the same rate (the eigenvalue), and might then be though of as a single growing entity. The steady state distribution is given by the leading eigenvector of this matrix, that is the eigenvector with the largest eigenvalue. Already from this equation one can see that the form of the steady state is independent of population size N as mentioned at the beginning of this chapter, since the equation is linear in x.

2.3.1 Modeling using PDEs

Equation 2.27 is unwieldy in the context of molecular evolution because of the huge number of possible genotypes, growing as 4^L for nucleotide sequences of length L. We would like a description in *phenotype space* rather than *genotype space.* That is, an equation describing $\eta(E,t)$, the distribution of sequences with energy E in the population at time t. Such an approach was first developed using a diffusion approximation by [15], which we now introduce.

2.3.2 Neutral Evolution

We begin with the simpler case of neutral evolution. The exact master equation describing neutral evolution is

$$\frac{d}{dt}\eta(x,t) = u \sum_{x' \in \mathcal{N}(x)} [\eta(x',t) - \eta(x,t)] \tag{2.28}$$

where $\eta(x,t)$ represents the number of individuals with genotype x at time t, and the sum is over the genotypes x' which are one mutation away from genotype x (corresponding the point mutations). As described in detail in [15], one may project this equation from genotypes to energy in the limit that $E(x) - E(x')$ is small and L is large, and assuming that sequences with the same energy are well mixed. That is, while the distribution $\eta(E)$ might be biased towards certain energies, sequences with the same energy are distributed randomly among themselves. This assumption is not true in general, but may be approximately true in some cases. Under this assumption, one obtains the Fokker-Planck equation

$$\frac{d}{dt}n(E,t) = u\left[\frac{\partial}{\partial E}En(E,t) + \frac{\partial^2}{\partial E^2}D(E)n(E,t)\right] \tag{2.29}$$

The drift term is proportional to E, and represents the fact that sequences will drift towards $E = 0$, as a result of the fact that the expected 'jump' from a sequence x with energy E_x in one generation is linear in E, as

$$\langle \Delta E \rangle(x) = \langle E'_x - E_x \rangle_{x' \in \mathcal{N}(x)} = \sum_{i\alpha} (\epsilon_{i\alpha} - \epsilon_{ix_i}) = -E_x/L \tag{2.30}$$

The drift term $D(E)$ is more complicated. The approach in [15] is to approximate it near $E = 0$, which gives $D(E) = \chi^2$, the variance of the neutral distribution. With this choice, the steady state solution to equation 2.29 is the Gaussian neutral distribution, as expected.

2.3.3 Evolution with Selection

We now wish to model quasispecies evolution with selection. In [15], the steady state with selection was found for one particular choice of fitness function (A 0-K Fermi-Dirac fitness) which could equivalently be modeled by imposing certain boundary conditions on the neutral equation. However, we wish to find steady states for arbitrary fitness functions. This may be achieved by adding a selection term to the neutral evolution equation proportional to $\log f(E)$, in order to cause multiplicative growth/death proportional to the fitness.

$$\frac{d}{dt}n(E,t) = u\left[\frac{\partial}{\partial E}En(E,t) + \frac{\partial^2}{\partial E^2}D(E)n(E,t)\right] + \log(f(E))n(E,t) \tag{2.31}$$

Analytical solutions to this equation, for some choice of fitness function, are difficult. Fokker-Planck equations are also notoriously difficult to solve numerically. However, we find a method known as Moving Finite Elements allows us to solve this equation efficiently and accurately for arbitrary fitness functions [93]. Unlike in the monomorphic case, there is no simple analytical formula.

As we shall not use the quasispecies steady state in later chapters, we shall stop at this result. However, as noted above, this result assumes that the sequences with the same energy are 'well mixed', which is not always the case due to the non-equilibrium nature of the quasispecies steady state, and this result mainly applies in the large L limit. Correcting these issues may be the subject of future work.

Chapter 3

Strong Selection in the Monomorphic Limit

Population genetics models of the substitution process have traditionally focused on the weak-selection regime, which is accurately described by diffusion theory. Predictions in this regime can be considered universal in the sense that many population models exhibit equivalent behavior in the diffusion limit, as discussed in the preceding chapters. However, a growing number of experimental studies suggest that strong selection plays a key role in some systems, and thus there is a need to understand universal properties of models without a priori assumptions about selection strength.

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, limit. 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 [57]. 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 [95–98], 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 (reviewed in [99–101]). 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 [72, 102, 103].

Models attempting to include a wider range of selection strengths are often deterministic [53, 104] and therefore exclude populations with non-negligible genetic drift, while stochastic theories typically demonstrate model-dependent behavior when selection becomes too strong

[66, 105, 106], 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 about the strength of selection are made.

In this chapter we investigate such properties, focusing on time reversibility (i.e., detailed balance) and the steady state of the substitution process. In Chapter 2 we introduced the function

$$\psi(r) \equiv \frac{\phi(r)}{\phi(1/r)}. \tag{3.1}$$

For *any* time-reversible population model, such as the Moran process, we have shown in Chapter 2 that the substitution rates obey a simple scaling law,

$$\psi(r) = r^{\nu}. \tag{3.2}$$

This result is exact in the monomorphic limit and requires no diffusion or weak-selection approximation. Here we shall extend this analysis to population models which are irreversible. 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 [62].

We have also shown that the scaling law leads to a simple steady state distribution,

$$\tilde{\pi}(\sigma) = \frac{1}{Z}\ \tilde{\pi}_0(\sigma)\ (f(\sigma))^{\nu}. \tag{3.3}$$

Here, we show that strong selection plays little role in steady state, which we find to be dominated by genetic drift and weak selection. Since evolutionary behavior in the weak regime is known to be universal through established results based on the diffusion approximation, the steady-state formula is accurate with 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 [78]. 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 [20], as will be discussed in Chapter 4.

Here, we shall consider reversibility in a number of population models. First, we consider Wright-Fisher and Moran models that describe populations of fixed size N. Next, we extend our treatment to a model in which population size varies periodically with time, and finally consider more general models proposed by [107] which allow for different variances in offspring number.

The function ψ defined in equation 3.1 will be of central importance in 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 as well as its form for specific models. In Chapter 2 we showed that reversibility implied the scaling law 3.2 for ψ, and we now continue this analysis.

3.0.4 The Scaling Law Implies Reversibility

We begin by showing that population models for which ψ follows the scaling law are reversible. Thus, we may investigate the reversibility for different models by examining ψ in that model. Let us assume Eq. 3.2 without assuming reversibility, in which case

$$\frac{W(\sigma'|\sigma)}{W(\sigma|\sigma')} = \frac{\tilde{\pi}_0(\sigma')}{\tilde{\pi}_0(\sigma)} \left(\frac{f(\sigma')}{f(\sigma)}\right)^{\nu} \tag{3.4}$$

based on the definition of $W(\sigma'|\sigma)$ (equation 2.1). We can combine this with the steady-state condition for the substitution process (Eq. 2.3) to show that

$$\begin{aligned} 0 &= \sum_{\sigma\in\mathcal{S}} [W(\sigma'|\sigma)\ \tilde{\pi}(\sigma) - W(\sigma|\sigma')\ \tilde{\pi}(\sigma')] \\ &= \sum_{\sigma\in\mathcal{S}} W(\sigma|\sigma') \left[\frac{\tilde{\pi}_0(\sigma')}{\tilde{\pi}_0(\sigma)} \left(\frac{f(\sigma')}{f(\sigma)}\right)^{\nu}\ \tilde{\pi}(\sigma) - \tilde{\pi}(\sigma')\right]. \end{aligned} \tag{3.5}$$

Clearly the distribution in Eq. 3.3 satisfies this condition, so it must be the unique steady state. Then it is trivial to show that this steady state and Eq. 3.4 automatically satisfy the reversibility condition (Eq. 2.4). Hence the power 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. 3.3. We will refer to these collective results as the *scaling law* of the substitution process.

3.1 The Limits of Reversibility

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. 3.3 hold. Obviously not all models are reversible, so the scaling law will not hold exactly in these cases. However, we demonstrate here 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. 3.3) 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 ($\log r = 0$):

$$\begin{aligned} \log \psi(r) &= \sum_{j=0}^{\infty} \frac{c_{2j+1}}{(2j+1)!} \, (\log r)^{2j+1} \\ &= c_1 (\log r) \left[1 + \frac{1}{c_1} \sum_{j=1}^{\infty} \frac{c_{2j+1}}{(2j+1)!} \, (\log r)^{2j} \right] , \end{aligned} \tag{3.6}$$

where

$$c_i = \left(\frac{d^i}{d(\log r)^i} \log \psi(r) \right) \bigg|_{r=1} . \tag{3.7}$$

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

$$\log \psi(r) = \nu (\log r) \left[1 + \frac{1}{\nu} \sum_{j=1}^{\infty} \frac{c_{2j+1}}{(2j+1)!} \, (\log r)^{2j} \right] . \tag{3.8}$$

Now we see that 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 2.1.3.

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., [65, 67–72]). The diffusion approximation is valid for weak-selection strengths: $r - 1 = s \sim \mathcal{O}(N^{-1})$ [55]. Since the scaling behavior of ψ appears in the diffusion regime, it will be shared by a large class of models.

3.1.1 The Range of Validity of the Scaling Law

There is a range of selection strengths in which the scaling approximation is 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

$$\nu(1 \mp \epsilon)\log r < \log\psi(r) < \nu(1 \pm \epsilon)\log r, \tag{3.9}$$

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. 3.8,

$$\frac{c_3}{6\nu} = \frac{\nu^3 - 3\nu^2 N^2 + 2\nu N^3\left(N - 3\phi''(1)\right) + 4N^5\left(3\phi''(1) + \phi^{(3)}(1)\right)}{12\nu N^6}, \tag{3.10}$$

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,

$$\frac{|c_3|}{6\nu}(\log r_0)^2 < \epsilon \quad \longrightarrow \quad r_0 = \exp\left(\sqrt{\frac{6\nu\epsilon}{|c_3|}}\right). \tag{3.11}$$

For any particular model, we need only compute 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 populations in 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, computed from Eq. 3.11, is (r_0^{-1}, r_0). Suppose that genotype σ_1 has fitness f_1 and genotype σ_2 has fitness less than 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 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:

$$\tilde{\pi}(\sigma_2) < \frac{1}{Z}\tilde{\pi}_0(\sigma_3)\ r_0^{-\nu} f_1^{\nu}. \tag{3.12}$$

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

$$\frac{\tilde{\pi}(\sigma_2)}{\tilde{\pi}(\sigma_1)} < \frac{\tilde{\pi}_0(\sigma_3)\ r_0^{-\nu} f_1^{\nu}}{\tilde{\pi}_0(\sigma_1)\ f_1^{\nu}} \simeq r_0^{-\nu}, \tag{3.13}$$

where the last relation holds because the neutral probabilities $\tilde{\pi}_0(\sigma_1)$ and $\tilde{\pi}_0(\sigma_3)$ 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.

3.1.2 The Power Law is Always Valid in Steady State

Furthermore, we can explicitly show that the selection strengths of the dominant substitutions in steady state are precisely those described by the diffusion approximation. Assume that we only want 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

$$r < 1 - \frac{1}{\nu}\log\delta + \mathcal{O}(\nu^{-2}). \tag{3.14}$$

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 [55]. Therefore the steady state of substitutions is adequately described by the diffusion approximation and thus by the scaling law (Eqs.

3.2 and 3.3). 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. 3.3 was previously derived for the special cases of the Moran process by [87] and for the diffusion limit of the Wright-Fisher model by [87], [19], and [108], among others. Indeed, some form of this formula can even be found in [49]. In Chapter 2 we 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, here we have shown how this result arises as an approximation and how weak selection dominates steady-state behavior in a wide class of population models, justifying the application of the steady-state formula to systems which may include mutations with large fitness effects.

3.2 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.

3.2.1 The Moran model

Consider a haploid population of fixed size N with two alleles, A and B, and let n denote the number of B alleles. As discussed in Chapter 1, the fixation probability of a single mutant in this model is given by

$$\phi(r) = \frac{1 - f_A/f_B}{1 - (f_A/f_B)^N} = \frac{1 - r^{-1}}{1 - r^{-N}}, \tag{3.15}$$

where $r = f_B/f_A$. A straightforward calculation shows that $\psi(r) = \phi(r)/\phi(1/r) = r^{N-1}$ [87]. Hence $\nu = N-1$ for Moran, and the scaling law holds exactly if the neutral substitution rates are reversible (Fig. 3.1A).

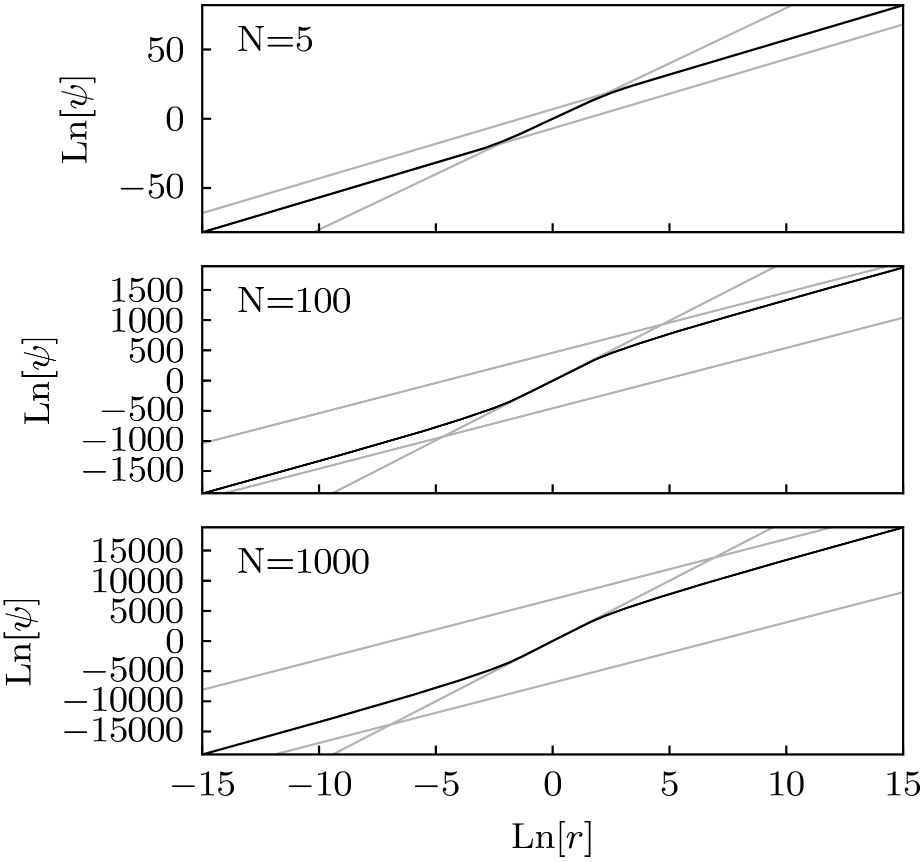


Figure 3.1: Plot of $\log \psi(r)$ as a function of $\log r$ for several population models. The scaling law appears as the straight line $\log \psi(r) = \nu \log r$. (A) The Moran model with $N = 1000$. Here the scaling law is exact with $\nu = N - 1$. (B) The simple Wright-Fisher model for $N = 1000$, calculated using the numerical procedure from 3.2.2. 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. 3.27, 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 ranges of selection strengths. (D) A model based on those in [3], 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.

3.2.2 The Wright-Fisher model

Next we consider the simple Wright-Fisher model for a haploid population of fixed size N with two alleles A and B, also introduced in Chapter 1 [49, 51]. 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. This yields many results in the neutral and weak-selection regimes [73, 94, 109], such as the formula for the fixation probability, as discussed in Chapter 1.

However, there are two problems with the classical diffusion approach. The first is that the moment functions $M(x, r)$ and $V(x, r)$ are typically expanded to the lowest order in $r-1$ for the weak-selection regime (as in 2.1.3), and so all subsequent calculations, including those leading to the fixation probability in eq. A.1, 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. 1.6. 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: computing fixation probabilities directly from the discrete Markov chain defined in Eq. 1.6. Here we pause for a moment to explain how this is achieved.

Exact Wright-Fisher fixation probability from discrete Markov chain

Studying discrete Markov chain properties of the Wright-Fisher model is not new [55]. 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.

We can represent the transition probabilities $\Pi(n'|n)$ from Eq. 1.6 as 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 matrix elements of powers of $\mathbf{P}$. That is, the probability of transitioning from n to n' in m generations is given by $(\mathbf{P}^m)_{n',n}$. Therefore the probability of fixation in m generations from initial state n is given by $(\mathbf{P}^m)_{N,n}$, and the probability of fixing a single mutant in the infinite time limit is given by

$$\lim_{m\to\infty}(\mathbf{P}^m)_{N,1} = \phi(r). \tag{3.16}$$

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 $\mathbf{A}_{ij} = \Pi(i|j)$ 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 $\mathbf{B}_{\alpha i} = \Pi(\alpha|i)$ 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 [110]:

$$\mathbf{P} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{B} & \mathbf{1}_2 \end{bmatrix}, \tag{3.17}$$

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:

$$\begin{aligned} \lim_{m\to\infty}\mathbf{P}^m = \lim_{m\to\infty}\begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{B} & \mathbf{1}_2 \end{bmatrix}^m &= \lim_{m\to\infty}\begin{bmatrix} \mathbf{A}^m & \mathbf{0} \\ \mathbf{B}(\mathbf{1}_{N-1}+\mathbf{A}+\cdots+\mathbf{A}^{m-1}) & \mathbf{1}_2 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{B}(\mathbf{1}_{N-1}-\mathbf{A})^{-1} & \mathbf{1}_2 \end{bmatrix}, \end{aligned} \tag{3.18}$$

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

$$(\mathbf{1}_{N-1} - \mathbf{A})^{-1} = \sum_{j=0}^{\infty} \mathbf{A}^j. \tag{3.19}$$

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 $n = 1$):

$$\phi(r) = (\mathbf{B}(\mathbf{1}_{N-1} - \mathbf{A})^{-1})_{2,1}. \tag{3.20}$$

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

$$\phi(r) = \mathbf{B}_{2,1} + \sum_{i=1}^{N-1} \mathbf{B}_{2,i}\mathbf{A}_{i,1} + \sum_{i,j=1}^{N-1} \mathbf{B}_{2,i}\mathbf{A}_{i,j}\mathbf{A}_{j,1} + \cdots . \tag{3.21}$$

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. 3.20 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	$\frac{r^N a_N(r)}{b_N(r)}$

(3.22)

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 $b_N(r)$ seems to be palindromic: $b_N(r) = r^{N(N-1)}b_N(1/r)$. In any case, the polynomials in these exact expressions grow increasingly intractable with N, making a numerical computation of $\phi(r)$ the only option. Eq. 3.20 can be rewritten as

$$(\mathbf{1}_{N-1} - \mathbf{A})^T\mathbf{u}^T = \mathbf{B}^T, \tag{3.23}$$

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

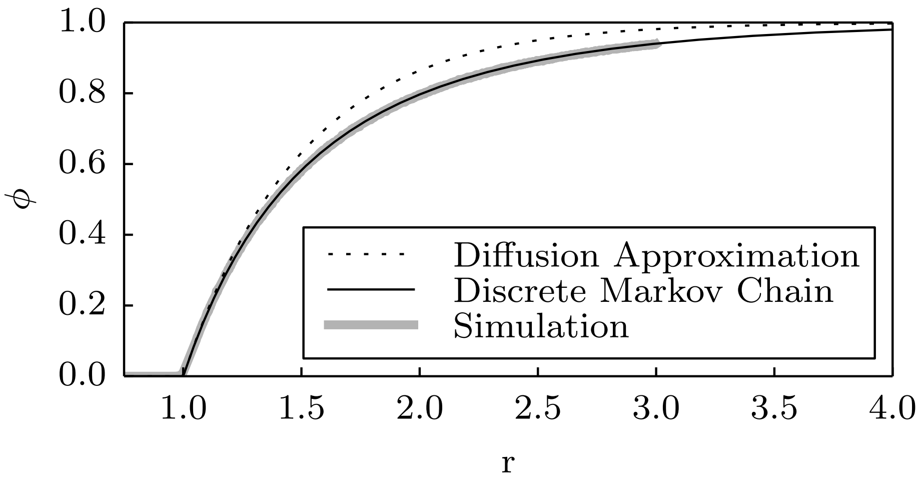


Figure 3.2: Plot of $\phi(r)$, the probability that a single mutant fixes as a function of its fitness ratio with the wild-type. For $N = 1000$, we compare an explicit simulation of the Wright-Fisher model with our discrete Markov chain approach (Eq. 3.23) and Kimura's diffusion approximation (Eq. 1.20). The agreement between the discrete Markov chain and the simulation is excellent, but there is noticeable disagreement with the diffusion approximation at larger selection strengths.

find $\mathbf{u}$ for the arbitrary fitness ratio r. The solution agrees extremely well with explicit simulations (Fig. 3.2).

Reversibility in the Wright-Fischer Model

This gives us an efficient numerical procedure for accurate computation of the fixation probability, and hence the ψ function, for any N and r. Figure 3.2 compares a simulation of $\phi(r)$ with this numerical approach along with the diffusion approximation (Eq. 1.20). 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 $\nu = 2N\phi'(1) = 2(N-1)$ [73]. Hence the expansion of $\psi(r)$ has the form

$$\log \psi(r) = 2(N-1)\log r + \mathcal{O}((\log r)^3). \tag{3.24}$$

Thus the power law and the steady state in Eq. 2.13 hold approximately with $\nu = 2(N-1)$.

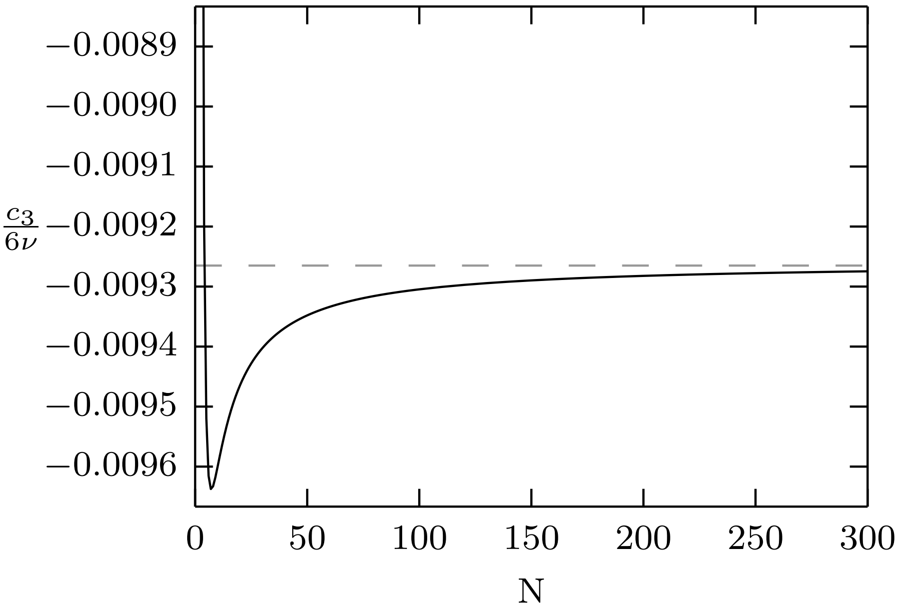


Figure 3.3: Plot of $c_3/6\nu$ as a function of N for the simple Wright-Fisher model, obtained numerically from $\phi(r)$ using the procedure described in 3.2.2. For realistic N values it rapidly converges to the constant ≈ -0.0093. This small value means that the scaling-law approximation is valid for 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.

As 3.2.2 shows, the form of the exact fixation probability is too complex to be useful for analytical calculations, such as computing c_3 in Eq. 3.10 to determine the range of selection strengths for which the power-law approximation is valid. However, we can numerically compute this next-order coefficient for a range of N using the method in 3.2.2, and Fig. 3.3 shows, remarkably, that it is N-independent for large N. Indeed, as N increases to realistic values the coefficient converges rapidly to

$$\frac{c_3}{6\nu} \approx -0.0093. \tag{3.25}$$

This value is striking both because it is small and effectively N-independent. 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. 3.11 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, which are well beyond the typical weak-selection limits of $\pm\mathcal{O}(N^{-1})$. A numerical calculation of ψ, as shown in Fig. 3.1B, confirms this large scaling region. Indeed, using the argument leading to Eq. 3.13, unfit genotypes that might lead to 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 [65]. 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. 1.20) is given by:

$$\frac{c_3}{6\nu} = -\frac{1}{6}N. \tag{3.26}$$

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.

3.2.3 Other models

Models that share the diffusion limit with the Moran and Wright-Fisher models will also share the scaling law. This encompasses a very wide class of exchangeable models [70, 71, 89]. 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 [67, 69, 111]. Other generalizations, such as incorporating the effects of subdivided populations, also lead to equivalencies [68, 72].

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 [67, 69]. Let the transition probabilities be of the Wright-Fisher form (Eq. 1.6), but now N changes over time according to

$$N(t) = N_0 + \alpha \sin\left(\frac{2\pi t}{T}\right), \tag{3.27}$$

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. 3.1C, 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 computational technique for fixation probabilities like the one used for the simple Wright-Fisher model (3.2.2), and explicit simulations prevent accurate statistics on fixation of very deleterious and extinction of very beneficial mutations. Nevertheless, deviations beyond this smaller range of selection can still be shown to be negligible in steady state. As Fig. 3.1C 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 allele types, Gillespie proposed to incorporate stochasticity at the level of selection by allowing for different variances in offspring number [3, 107, 112]. In these models fitness is characterized by $\mu - \sigma^2/N$, where μ is the mean and σ^2 is the variance in the offspring number for a given allele. Other authors have further pursued 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 [66, 105, 106].

For instance, we simulate a model described in [3]. Consider a haploid population of two allele types, 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_A^2 if i is of type A, or μ_B and σ_B^2 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 alleles. We simulate this process and obtain the ψ function in Fig. 3.1D. 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).

3.3 Discussion

3.3.1 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 [65, 66]. 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, and so the universality of the diffusion limit gives rise to a universal scaling law (Eq. 2.11) 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 for weak selection even for irreversible models. In some cases such as the simple Wright-Fisher model, it holds for such a large range of selection strengths that deviations from it are not practically important. This finding significantly generalizes previous work of [87], [19], [108], and others.

3.3.2 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 [55, 113, 114], and diffusion is known to adequately capture weak-selection behavior [115]. 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 [55]. 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 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 [116]. 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 [78, 117]. Time reversibility links the prospective and retrospective paradigms and thus has been exploited, for instance, in studies of age properties [55, 113, 118] and in phylogenetic methods [78].

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 alleles, A, B, and C, in order of decreasing fitness. If the substitution process is irreversible, there would be a net current around the loop $C \to B \to A \to C$. The net currents $C \to B$ and $B \to A$ flow from less fit to more fit alleles, but to complete the cycle, there is also a current $A \to C$ 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.

3.3.3 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 [58–61]. 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 [62]. A precise form of the steady-state distribution is critical to these applications, since this distribution is 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., [62, 63, 108, 119–122]). Results for the steady-state distribution using the standard Wright-Fisher diffusion approximation (i.e., Eq. 1.20) 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 [95–98], 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. We have provided a theoretical framework to understand the limits of this steady-state distribution, which provides a precise way to show that these big-benefit mutations are negligible in steady state.

Moreover, 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 [78]. This approximation excludes correlations among sites within a locus and the corresponding epistatic effects, whose importance is being increasingly emphasized [22, 123–125].

As noted in Chapter 2, one application of particular interest is the ability to infer an arbitrary fitness landscape from sequence data under the assumption of steady state, and eq. 2.13 can be inverted to obtain the fitness function in terms of the neutral distribution and the steady-state distribution under selection. This is this subject of chapter 4.

3.4 Appendix I: Publication Attached

Parts of this chapter were published in [30]. The publication is attached.

Theoretical Population Biology 82 (2012) 66–76

Contents lists available at SciVerse ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

A universal scaling law determines time reversibility and steady state of substitutions under selection

Michael Manhart [a], Allan Haldane [a], Alexandre V. Morozov [a,b,*]

[a] *Department of Physics and Astronomy, Rutgers University, 136 Frelinghuysen Road, Piscataway, NJ 08854, USA*

[b] *BioMaPS Institute for Quantitative Biology, Rutgers University, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA*

ARTICLE INFO

Article history:
Received 11 February 2012
Available online 10 April 2012

Keywords:
Substitutions
Reversibility
Universality
Fitness landscape

ABSTRACT

Monomorphic loci evolve through a series of substitutions on a fitness landscape. Understanding how mutation, selection, and genetic drift drive this process, and uncovering the structure of the fitness landscape from genomic data are two major goals of evolutionary theory. Population genetics models of the substitution process have traditionally focused on the weak-selection regime, which is accurately described by diffusion theory. Predictions in this regime can be considered universal in the sense that many population models exhibit equivalent behavior in the diffusion limit. However, a growing number of experimental studies suggest that strong selection plays a key role in some systems, and thus there is a need to understand universal properties of models without *a priori* assumptions about selection strength. Here we study time reversibility in a general substitution model of a monomorphic haploid population. We show that for *any* time-reversible population model, such as the Moran process, substitution rates obey an exact scaling law. For several other irreversible models, such as the simple Wright–Fisher process and its extensions, the scaling law is accurate up to selection strengths that are well outside the diffusion regime. Time reversibility gives rise to a power-law expression for the steady-state distribution of populations on an arbitrary fitness landscape. The steady-state behavior is dominated by weak selection and is thus adequately described by the diffusion approximation, which guarantees universality of the steady-state formula and its applicability to the problem of reconstructing fitness landscapes from DNA or protein sequence data.

1. Introduction

A key goal of evolutionary theory is to determine the role of natural selection in the evolution of genotypes, and to infer information about selection strength from the growing abundance of genomic data. Theoretical work on these issues takes many different forms, both because of the inherent differences among biological systems and because different simplifying assumptions are necessary for the sake of mathematical tractability. One common approximation is to consider unlinked loci in the monomorphic limit, valid for neutral evolution once sufficiently low mutation rates and effective population sizes ensure that genetic drift dominates (Crow and Kimura, 1970). Even larger populations or those with greater mutation rates can be nearly monomorphic if selection is significant.

If at any given time the population is dominated by a single genotype at the locus of interest, to a good approximation such a population evolves as a single entity on a fitness landscape (Wright, 1932) over genotype space, assuming that the evolutionary success of a genotype can be distilled into a fitness value. The movement of the entire population from one genotype to another is known as the substitution process, where each substitution event consists of a single mutation arising and then fixing instantaneously (Kimura, 1983). This picture greatly simplifies the theory, especially because it permits fixation events to be analyzed using two-allele models of population genetics (Crow and Kimura, 1970). Moreover, it is believed that many higher eukaryotes (Kimura, 1983) and some microorganisms contain loci that can be adequately described as monomorphic (Ochman and Selander, 1984; Wick et al., 2002; Dos Vultos et al., 2008; Achtman, 2008). As a result, this approach has been followed in settings as diverse as the evolution of transcription factor (TF) binding sites in yeast (Lässig, 2007; Mustonen et al., 2008), viral protein evolution (Bloom et al., 2007; Bloom and Glassman, 2009), and codon usage bias (e.g., McVean and Vieira, 2001; Yang and Nielsen, 2008). These theoretical and computational studies complement recent experimental work that has begun to reconstruct empirical fitness landscapes directly (Weinreich et al., 2006; Poelwijk et al., 2007).

Much theoretical work in population genetics has focused on gradual models of adaptation in which evolutionary change

---

[*] Corresponding author at: Department of Physics and Astronomy, Rutgers University, 136 Frelinghuysen Road, Piscataway, NJ 08854, USA.

E-mail addresses: mmanhart@physics.rutgers.edu (M. Manhart), ahalda@physics.rutgers.edu (A. Haldane), morozov@physics.rutgers.edu (A.V. Morozov).

0040-5809/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2012.03.007

proceeds through selection of alleles with very small fitness advantage (Orr, 2005). The idea of the extremely slow rate of phenotypic evolution was proposed by Darwin (1859) and subsequently made popular by Fisher (1958) in the context of the infinitesimal model. In more recent decades, experimental evidence like 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 (Kimura, 1983; Ohta and Tachida, 1990; Ohta, 1992).

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 (Wright, 1931; Fisher, 1958) and Moran (Moran, 1958) models, which share a diffusion limit with a variety of more elaborate models under the appropriate mapping of parameters (e.g., Ewens, 1967; Maruyama, 1970; Otto and Whitlock, 1997; Möhle, 2001; Möhle and Sagitov, 2001; Whitlock, 2003; Wakeley, 2005). 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 (Kimura, 1955, 1962), and more generally, the use of exchangeable models (Cannings, 1974) 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 (Kimura, 1983). 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 (Wichman et al., 1999; Bull et al., 2000; Holder and Bull, 2001; Barrett et al., 2006b), 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 (reviewed in Orr, 2001, 2005; Eyre-Walker and Keightley, 2007). 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 (Morjan and Rieseberg, 2004; Whitlock, 2003; Barrett et al., 2006a).

Models attempting to include a wider range of selection strengths are often deterministic (Eigen et al., 1989; Bürger, 2000) and therefore exclude populations with non-negligible genetic drift, while stochastic theories typically demonstrate model-dependent behavior when selection becomes too strong (Proulx, 2000; Shpak, 2007; Parsons et al., 2010), 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 about the strength of selection are made.

In this paper 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 (Ochman and Selander, 1984; Wick et al., 2002; Dos Vultos et al., 2008; Achtman, 2008). 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 (McVean and Vieira, 2001).

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 (2005), 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 (Yang, 2006). 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 (Mustonen et al., 2008).

2. Substitution model for monomorphic populations

We consider the evolution of a single locus in the monomorphic limit, where the mutation rate is sufficiently low that the vast majority of single mutations either fix or become extinct before a second mutation on the locus arises (Kimura, 1983). Thus we can describe evolution of this locus as a series of substitution events in which the entire population switches from genotype σ to genotype σ'. Since the time scale for fixation or extinction of a mutant (during which the population is actually polymorphic) is very short compared to the time scales of interest, we approximate these events as instantaneous. For a locus of length L and single-site mutation rate μ, Champagnat (2006) and Champagnat et al. (2006) have shown that the condition necessary to guarantee a monomorphic population is $\mu \leq 1/(LN \log N)$ for a population of size N. However, if most mutations introduce significant selective effects, the fixation or extinction of mutants will occur more rapidly, weakening the condition on μ. For beneficial mutations of selective advantage s (where $1 \ll Ns \ll N$), Desai and Fisher (2007) have shown that the monomorphic condition becomes $\mu \leq 1/(LN \log(Ns))$.

We will assume that the locus of interest is unlinked to the rest of the genome (linkage equilibrium) by frequent recombination with rate ρ, which satisfies $\rho \gg N\mu L$ (Mustonen and Lässig, 2010); here, recombination also includes homologous DNA transfer such as that 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, and 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.

Let σ and σ' be two genotypes (i.e., sequences of L nucleotides or amino acids) at the locus of interest. The substitution rate from σ to σ' can be approximated by the rate of producing a single mutant times the probability that the mutation fixes (Kimura and Ohta, 1971; Kimura, 1983):

$$W(\sigma'|\sigma) \approx N\mu(\sigma'|\sigma) \cdot \phi(\sigma'|\sigma), \tag{1}$$

where N is an effective population size, $\mu(\sigma'|\sigma)$ is the nucleotide or amino acid mutation rate from σ to σ', and $\phi(\sigma'|\sigma)$ is the probability that a single σ' mutant fixes in a population of wild-type σ. We will assume that μ is nonzero only for genotypes σ and σ' differing by a single nucleotide or amino acid.

Given an ensemble of populations evolving with these rates, we can define $\pi(\sigma, t)$ to be the probability that a population is monomorphic at the locus with genotype σ at time t. This probability evolves over time via the master equation

$$\frac{d}{dt}\pi(\sigma', t) = \sum_{\sigma\in\mathcal{S}}[W(\sigma'|\sigma)\,\pi(\sigma, t) - W(\sigma|\sigma')\,\pi(\sigma', t)], \tag{2}$$

where $\mathcal{S}$ is the set of all possible genotypes at the locus of interest. 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 $\tilde{\pi}(\sigma)$ (Allen, 2011) satisfying

$$\sum_{\sigma\in\mathcal{S}}[W(\sigma'|\sigma)\,\tilde{\pi}(\sigma) - W(\sigma|\sigma')\tilde{\pi}(\sigma')] = 0. \tag{3}$$

The form of this steady-state distribution depends on the underlying population genetics model that gives the fixation probability ϕ.

The monomorphic limit permits us to consider two-allele population models without mutation. First, we consider Wright–Fisher and Moran models that describe populations of fixed size N, with the selective value of each genotype σ specified by a single parameter $f(\sigma)$ which we refer to as the genotype's fitness. Next, we extend our treatment to a model in which population size varies periodically with time, and finally consider more general models proposed by Gillespie (1974) which allow for different variances in offspring number.

An important consequence of fixed size N is that only relative fitnesses matter. Relative fitness can be an arithmetic difference or a ratio, depending on the parameterization. These are equivalent under a simple exponential mapping. Note that the model is then symmetric under either a shift or rescaling of all fitnesses, a symmetry which is convenient to maintain at all stages of an approximation. For instance, in the Wright–Fisher or Moran models, it is typical to incorporate fitness as a multiplicative weight in the transition probabilities, in which case all observable quantities depend only on the ratio of the wild-type to mutant fitness. In particular, the probability that a single σ' mutant fixes in a population of wild-type σ must only depend on $r = f(\sigma')/f(\sigma)$ and, implicitly, on the population size N: $\phi(\sigma'|\sigma) \equiv \phi(r)$.

3. The scaling law and steady state

Since substitution rates depend on the fixation probability $\phi(r)$, 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:

$$W(\sigma'|\sigma)\,\tilde{\pi}(\sigma) = W(\sigma|\sigma')\,\tilde{\pi}(\sigma'), \tag{4}$$

where $\tilde{\pi}(\sigma)$ denotes the steady-state distribution. The left- and right-hand sides of this equation are the steady-state probability currents $\sigma \to \sigma'$ and $\sigma' \to \sigma$, respectively. Eq. (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 paper, 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) shows that the neutral substitution rates are just the mutation rates (Kimura, 1983): $W(\sigma'|\sigma) = \mu(\sigma'|\sigma)$. Let the steady-state distribution of the neutral substitution process be $\tilde{\pi}_0(\sigma)$. Then reversibility of the neutral model is expressed by

$$\mu(\sigma'|\sigma)\,\tilde{\pi}_0(\sigma) = \mu(\sigma|\sigma')\,\tilde{\pi}_0(\sigma'). \tag{5}$$

Many popular neutral models are reversible (see Yang, 2006, for a summary), although this condition is not guaranteed. This issue will be explored further in Section 5.

We now consider the reversibility of the substitution rates under selection, $N\mu(\sigma'|\sigma)\phi(r)$. Let us first define the function

$$\psi(r) \equiv \frac{\phi(r)}{\phi(1/r)}. \tag{6}$$

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

$$\frac{W(\sigma'|\sigma)}{W(\sigma|\sigma')} = \frac{\mu(\sigma'|\sigma)}{\mu(\sigma|\sigma')}\cdot\frac{\phi\left(\frac{f(\sigma')}{f(\sigma)}\right)}{\phi\left(\frac{f(\sigma)}{f(\sigma')}\right)} = \frac{\tilde{\pi}_0(\sigma')}{\tilde{\pi}_0(\sigma)}\cdot\psi\left(\frac{f(\sigma')}{f(\sigma)}\right), \tag{7}$$

where we have invoked the reversibility of the neutral rates (Eq. (5)). Studying the properties of the ψ function is the main focus of this paper: 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 $W(\sigma'|\sigma)$ under selection are reversible, which we will show completely constrains the form of ψ and the steady state under selection $\tilde{\pi}(\sigma)$. In this case, $W(\sigma'|\sigma)\tilde{\pi}(\sigma) = W(\sigma|\sigma')\tilde{\pi}(\sigma')$, and hence

$$\frac{\tilde{\pi}(\sigma')}{\tilde{\pi}(\sigma)} = \frac{W(\sigma'|\sigma)}{W(\sigma|\sigma')} = \frac{\tilde{\pi}_0(\sigma')}{\tilde{\pi}_0(\sigma)}\cdot\psi\left(\frac{f(\sigma')}{f(\sigma)}\right). \tag{8}$$

It follows that

$$\psi\left(\frac{f(\sigma'')}{f(\sigma')}\right)\cdot\psi\left(\frac{f(\sigma')}{f(\sigma)}\right) = \psi\left(\frac{f(\sigma'')}{f(\sigma)}\right), \tag{9}$$

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:

$$\psi(r) = r^{\nu}, \tag{10}$$

for some constant ν (Roberts, 1979). 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. (10) as the scaling law for ψ. Using the definition of $\psi(r)$ (Eq. (6)), one can show that

$$\nu = \frac{2\phi'(1)}{\phi(1)} = 2N\phi'(1), \tag{11}$$

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

Now rewriting Eq. (8) with our explicit form of ψ,

$$\frac{\tilde{\pi}(\sigma')}{\tilde{\pi}(\sigma)} = \frac{\tilde{\pi}_0(\sigma')}{\tilde{\pi}_0(\sigma)}\left(\frac{f(\sigma')}{f(\sigma)}\right)^{\nu}, \tag{12}$$

we can deduce the steady state:

$$\tilde{\pi}(\sigma) = \frac{1}{Z}\tilde{\pi}_0(\sigma)\,(f(\sigma))^{\nu}, \tag{13}$$

where Z is a normalization constant. Note that Eq. (13) can be rewritten in the form of a Boltzmann distribution, with energy replaced by the negative logarithm of fitness:

$$\tilde{\pi}(\sigma) = \frac{1}{Z}\tilde{\pi}_0(\sigma)\,e^{\nu\log f(\sigma)}. \tag{14}$$

The Boltzmann form in Eq. (14) suggests a straightforward analogy with statistical mechanics (Iwasa, 1988; Sella and Hirsh, 2005). One may think of the evolutionary model defined by Eqs. (1) and (2) 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. (13) or (14). Populations will be driven toward the peaks of the landscape by selection, which manifests itself as the f^{ν} 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 $\tilde{\pi}_0$. The corresponding entropy (defined as $-\log\tilde{\pi}_0$) 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 explored in Iwasa (1988) and Sella and Hirsh (2005).

This steady-state formula was derived in the special case of the Moran model by Sella and Hirsh (2005). We generalize this earlier result by showing that *any* reversible substitution process leads to the power law for ψ and the steady-state formula of Eq. (13). Note that this conclusion, obtained in the monomorphic limit, requires no additional assumptions, such as the weak-selection diffusion approximation.

Next, we show that the power law implies reversibility. We now assume Eq. (10) without assuming reversibility. Then

$$\frac{W(\sigma'|\sigma)}{W(\sigma|\sigma')} = \frac{\tilde{\pi}_0(\sigma')}{\tilde{\pi}_0(\sigma)}\left(\frac{f(\sigma')}{f(\sigma)}\right)^{\nu}. \tag{15}$$

We can combine this with the steady-state condition (Eq. (3)) to show that

$$\begin{aligned}0 &= \sum_{\sigma\in\mathcal{S}}[W(\sigma'|\sigma)\tilde{\pi}(\sigma) - W(\sigma|\sigma')\tilde{\pi}(\sigma')]\\ &= \sum_{\sigma\in\mathcal{S}} W(\sigma|\sigma')\left[\frac{\tilde{\pi}_0(\sigma')}{\tilde{\pi}_0(\sigma)}\left(\frac{f(\sigma')}{f(\sigma)}\right)^{\nu}\tilde{\pi}(\sigma) - \tilde{\pi}(\sigma')\right].\end{aligned} \tag{16}$$

Clearly the distribution in Eq. (13) satisfies this condition, so it must be the unique steady state. The reversibility condition (Eq. (4)) is satisfied as well, and thus the power 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. (13). We will refer to these collective results as the *scaling law* of the substitution process. 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. (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. (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 ($\log r = 0$):

$$\begin{aligned}\log\psi(r) &= \sum_{j=0}^{\infty}\frac{c_{2j+1}}{(2j+1)!}(\log r)^{2j+1}\\ &= c_1(\log r)\left[1 + \frac{1}{c_1}\sum_{j=1}^{\infty}\frac{c_{2j+1}}{(2j+1)!}(\log r)^{2j}\right],\end{aligned} \tag{17}$$

where

$$c_i = \left(\frac{d^i}{d(\log r)^i}\log\psi(r)\right)\bigg|_{r=1}. \tag{18}$$

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

$$\log\psi(r) = \nu(\log r)\left[1 + \frac{1}{\nu}\sum_{j=1}^{\infty}\frac{c_{2j+1}}{(2j+1)!}(\log r)^{2j}\right]. \tag{19}$$

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. 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., Ewens, 1967; Maruyama, 1970; Otto and Whitlock, 1997; Möhle, 2001; Möhle and Sagitov, 2001; Whitlock, 2003; Wakeley, 2005). The diffusion approximation is valid for weak-selection strengths: $r - 1 = s \sim \mathcal{O}(N^{-1})$ (Ewens, 2004). 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 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. (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

$$\nu(1 \mp \epsilon)\log r < \log\psi(r) < \nu(1 \pm \epsilon)\log r, \tag{20}$$

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. (19),

$$\frac{c_3}{6\nu} = \frac{1}{12\nu}\left(\nu^3 - 3\nu^2 + 2\nu - 6N(\nu-2)\phi''(1) + 4N\phi^{(3)}(1)\right), \tag{21}$$

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,

$$\frac{|c_3|}{6\nu}(\log r_0)^2 = \epsilon \longrightarrow r_0 = \exp\left(\sqrt{\frac{6\nu\epsilon}{|c_3|}}\right). \tag{22}$$

For any particular model, we need only compute 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, computed from Eq. (22), is (r_0^{-1}, r_0).

Suppose that genotype σ_1 has fitness f_1 and genotype σ_2 has fitness less than 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 f_1/r_0, so that its probability is given by Eq. (13). Since ψ must be monotonically increasing, the probability of the unfit σ_2 is bounded from above by the probability of σ_3:

$$\tilde{\pi}(\sigma_2) < \frac{1}{Z}\tilde{\pi}_0(\sigma_3)\, r_0^{-\nu} f_1^{\nu}. \tag{23}$$

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

$$\frac{\tilde{\pi}(\sigma_2)}{\tilde{\pi}(\sigma_1)} < \frac{\tilde{\pi}_0(\sigma_3)\, r_0^{-\nu} f_1^{\nu}}{\tilde{\pi}_0(\sigma_1)\, f_1^{\nu}} \simeq r_0^{-\nu}, \tag{24}$$

where the last relation holds because the neutral probabilities $\tilde{\pi}_0(\sigma_1)$ and $\tilde{\pi}_0(\sigma_3)$ 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

$$r < 1 - \frac{1}{\nu}\log\delta + \mathcal{O}(\nu^{-2}). \tag{25}$$

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 (Ewens, 2004). Therefore the steady state of substitutions is adequately described by the diffusion approximation and thus by the scaling law (Eqs. (10) and (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. (13) was previously derived for the special cases of the Moran process by Sella and Hirsh (2005) and for the diffusion limit of the Wright–Fisher model by Sella and Hirsh (2005), Lässig (2007), and Li (1987), among others. Indeed, some form of this formula can even be found in Wright (1931). 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.

4. 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.

4.1. The Moran model

Consider a haploid population of fixed size N with two alleles, A and B, and let n denote the number of B alleles. The single time-step transition probabilities of the Moran model are then (Moran, 1958; Ewens, 2004)

$$\begin{aligned} \Pi(n+1|n) &= \frac{f_B}{\bar{f}}\frac{n}{N}\left(1-\frac{n}{N}\right) \\ \Pi(n-1|n) &= \frac{f_A}{\bar{f}}\frac{n}{N}\left(1-\frac{n}{N}\right) \\ \Pi(n|n) &= 1-\Pi(n+1|n)-\Pi(n-1|n), \end{aligned} \tag{26}$$

where f_A, f_B are fitnesses of alleles A and B and $\bar{f} = (n/N)f_B + (1 - n/N)f_A$ is the average fitness. In this case the probability of fixing a single mutant is (Ewens, 2004)

$$\phi(r) = \frac{1-r^{-1}}{1-r^{-N}}, \tag{27}$$

where $r = f_B/f_A$. A straightforward calculation shows that $\psi(r) = \phi(r)/\phi(1/r) = r^{N-1}$ (Sella and Hirsh, 2005). Hence $\nu = N - 1$ for Moran, and the scaling law holds exactly if the neutral substitution rates are reversible (Fig. 1A).

4.2. The Wright–Fisher model

Next we define the simple Wright–Fisher model for a haploid population of fixed size N with two alleles A and B of fitness f_A and f_B, respectively (Wright, 1931; Fisher, 1958). Given that there are n alleles of type B in the current generation, the probability of having n' B alleles in the next generation is (Rouzine et al., 2001; Ewens, 2004)

$$\Pi(n'|n) = \binom{N}{n'} q^{n'}(1-q)^{N-n'}, \quad \text{where } q \equiv \frac{n}{N}\frac{f_B}{\bar{f}}. \tag{28}$$

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 (Kimura, 1955, 1957, 1962), such as the formula for the fixation probability:

$$\phi(r) = \frac{1-e^{2(1-r)}}{1-e^{2N(1-r)}}, \tag{29}$$

where $r = f_B/f_A$. However, there are two problems with the classical diffusion approach. The first is that the moment functions $M(x,r)$ and $V(x,r)$ are typically expanded to the lowest order in $r - 1$ for the weak-selection regime (as in Appendix A), and so all subsequent calculations, including those leading to the fixation probability in Eq. (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. (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: computing fixation probabilities directly from the discrete Markov chain defined in Eq. (28) (Appendix B). The end result is an efficient numerical procedure for accurate computation of the fixation probability, and hence the ψ function, for any N and r. Fig. 2 compares a

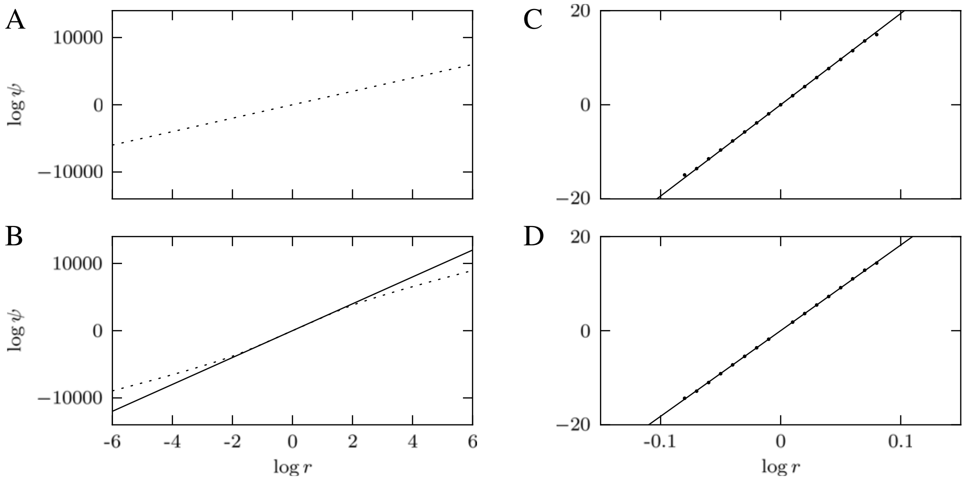


Fig. 1. Plot of $\log\psi(r)$ as a function of $\log r$ for several population models. The scaling law appears as the straight line $\log\psi(r) = \nu\log r$. (A) The Moran model with $N = 1000$. Here the scaling law is exact with $\nu = N - 1$. (B) The simple Wright–Fisher model for $N = 1000$, calculated using the numerical procedure from Appendix B. 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. (33), with $N_0 = 100$, $\alpha = 20$ and $T = 20$ generations. Simulation results are shown as dots (with each dot an average over 10^8 independent runs), 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 those in Gillespie (1975), 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.

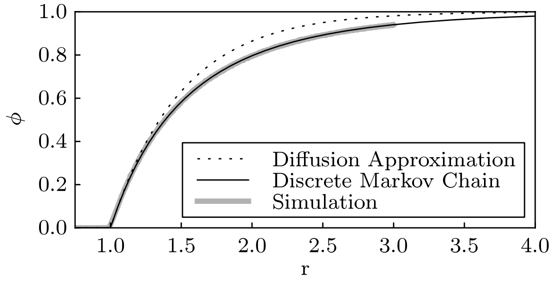


Fig. 2. Plot of $\phi(r)$, the probability that a single mutant fixes as a function of its fitness ratio with the wild-type. For $N = 1000$, we compare an explicit simulation of the Wright–Fisher model with our discrete Markov chain approach (Eq. (B.8)) and Kimura's diffusion approximation (Eq. (29)). The explicit simulation data is averaged over 10^6 independent runs. 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.

simulation of $\phi(r)$ with this numerical approach along with the diffusion approximation (Eq. (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 $\nu = 2N\phi'(1) = 2(N-1)$ (Kimura, 1962). Hence the expansion of $\psi(r)$ has the form

$$\log\psi(r) = 2(N-1)\log r + \mathcal{O}((\log r)^3). \quad (30)$$

Thus the power law and the steady state in Eq. (13) hold approximately with $\nu = 2(N-1)$. As Appendix B shows, the form of the exact fixation probability is too complex to be useful for analytical calculations, such as computing c_3 in Eq. (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 B to obtain derivatives of fixation probabilities for Eq. (21). Fig. 3 shows, 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

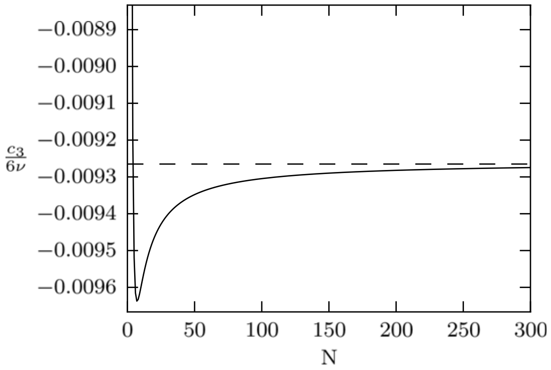


Fig. 3. Plot of $c_3/6\nu$ as a function of N for the simple Wright–Fisher model, obtained numerically from $\phi(r)$ using the procedure described in Appendix B. For realistic N values it rapidly converges to the constant ≈ -0.0093. This small value means that the scaling-law approximation is valid 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.

$$\frac{c_3}{6\nu} \approx -0.0093. \quad (31)$$

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. (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. 1B). Indeed, using the argument leading to Eq. (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 (Wakeley, 2005). 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. (29)) is given by:

$$\frac{c_3}{6\nu} = -\frac{1}{6}N. \tag{32}$$

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.

4.3. *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 (Cannings, 1974; Möhle, 2001; Möhle and Sagitov, 2001). 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 (Ewens, 1967; Otto and Whitlock, 1997; Sjödin et al., 2005). Other generalizations, such as incorporating the effects of subdivided populations, also lead to equivalencies (Maruyama, 1970; Whitlock, 2003).

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 (Ewens, 1967; Otto and Whitlock, 1997). Let the transition probabilities be of the Wright–Fisher form (Eq. (28)), with N changing over time according to

$$N(t) = N_0 + \alpha \sin\left(\frac{2\pi t}{T}\right), \tag{33}$$

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. 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 computational technique for fixation probabilities like the one used for the simple Wright–Fisher model Appendix B, 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. 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 allele types, Gillespie proposed to incorporate stochasticity at the level of selection by allowing for different variances in offspring number (Gillespie, 1974, 1975, 1977). 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 (Proulx, 2000; Shpak, 2007; Parsons et al., 2010).

Here we simulate a model described in Gillespie (1975). Consider a haploid population of two allele types, 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_A^2 if i is of type A, or μ_B and σ_B^2 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 alleles. We simulate this process to obtain the ψ function (Fig. 1D). Fitness ratios r are defined using the fitness definition $f_i = \mu_i - \sigma_i^2/N$. For each i, X_A or X_B is generated from the binomial distribution $B(n, p_A)$ or $B(n, p_B)$, respectively, where $n = 10$ and p_A and p_B are given by the desired fitness ratio r ($p_A + p_B = 1$). 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. 1D).

5. Discussion

5.1. *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 (Wakeley, 2005; Parsons et al., 2010). 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. (10)) which determines the steady-state distribution (Eq. (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 Sella and Hirsh (2005), Lässig (2007), Li (1987), and others.

5.2. *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 (Watterson, 1977; Levikson, 1977; Ewens, 2004), and diffusion is known to adequately capture weak-selection behavior (Kurtz, 1981). 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 (Ewens, 2004). 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. (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 (Ewens, 1990). 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 (Kingman, 1982; Yang, 2006). Time reversibility links the prospective and retrospective paradigms and thus has been exploited, for instance, in studies of age properties (Watterson, 1976, 1977; Ewens, 2004) and in phylogenetic methods (Yang, 2006).

An additional consequence of reversibility is the nonexistence of net probability currents in steady state, as guaranteed by Eq. (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 alleles, A, B, and C, in order of decreasing fitness. If the substitution process is irreversible, there would be a net current around the loop $C \to B \to A \to C$. The net currents $C \to B$ and $B \to A$ flow from less fit to more fit alleles, but to complete the cycle, there is also a current $A \to C$ 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.

5.3. 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 (Ochman and Selander, 1984; Wick et al., 2002; Dos Vultos et al., 2008; Achtman, 2008). 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 (McVean and Vieira, 2001). A precise form of the steady-state distribution is important in these applications, since 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., Li, 1987; Bulmer, 1991; McVean and Charlesworth, 1999; McVean and Vieira, 1999, 2001; Nielsen et al., 2007; Yang and Nielsen, 2008). Results for the steady-state distribution using the standard Wright–Fisher diffusion approximation (Eq. (29)) for individual codons have been reported that are consistent with Eq. (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 (Wichman et al., 1999; Bull et al., 2000; Holder and Bull, 2001; Barrett et al., 2006b), 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., Jukes and Cantor, 1969; Kimura, 1980; Tamura and Nei, 1993; Felsenstein, 1981; Yang, 2006; Felsenstein, 2011). Moreover, Rodríguez et al. (1990) 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 (e.g., Barry and Hartigan, 1987). 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 (Yang, 2006). This approximation excludes correlations among sites within a locus and the corresponding epistatic effects, whose importance is being increasingly emphasized (DePristo et al., 2005; Bershtein et al., 2006; Weinreich et al., 2006; Poelwijk et al., 2007).

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. (13) can be inverted to obtain the fitness function in terms of the neutral distribution and the steady-state distribution under selection (Lässig, 2007; Mustonen et al., 2008):

$$\log\left(\frac{\tilde{\pi}(\sigma)}{\tilde{\pi}_0(\sigma)}\right) = \nu \log f(\sigma) - \log Z. \tag{34}$$

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. (34) gives logarithmic fitness up to an overall scaling and shift.

The application of Eq. (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 Section 2, the monomorphic limit requires $\mu \leq 1/(LN \log N)$ for neutral evolution (Champagnat, 2006; Champagnat et al., 2006). 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. Eq. (1) 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 (Levin et al., 2009). 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 on the order of

$$(LN \log N) \times (2L \log(2L)) \text{ generations.} \tag{35}$$

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 (Berg and Lässig, 2003; Berg et al., 2004; Lässig, 2007; Mustonen et al., 2008). The steady state is expected to be reached more quickly in the one-dimensional energy space than in the high-dimensional genotype space (Mustonen et al., 2008). Mustonen et al. (2008) 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 (Sarai and Takeda, 1989; Lehming et al., 1990), 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. (34) to genomic data such as collections of TF binding sites. Current work is in progress to apply these results to evolution of regulatory sites in yeast, exploring the biophysical origins of the underlying fitness landscapes.

Acknowledgments

We wish to thank an anonymous reviewer who provided useful comments regarding the approach to evolutionary steady state. This research was supported by National Institutes of Health (R01 HG004708) and by an Alfred P. Sloan Research Fellowship to A.V.M.

Appendix A. 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 Kimura (1962):

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

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 cancelation 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. 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 (Lenski and Elena, 2003), 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.

Eq. (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 (Cannings, 1974; Ewens, 2004).

Appendix B. Exact Wright–Fisher fixation probability from discrete Markov chain

Studying discrete Markov chain properties of the Wright–Fisher model is not new (Ewens, 2004). 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.

We can represent the transition probabilities $\Pi(n'|n)$ from Eq. (28) as 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 $(\mathbf{P}^m)_{n',n}$. Therefore the probability of fixation by generation m from initial state n is given by $(\mathbf{P}^m)_{N,n}$, and the probability of fixing a single mutant in the infinite time limit is given by

$$\lim_{m\to\infty}(\mathbf{P}^m)_{N,1} = \phi(r). \tag{B.1}$$

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 $\mathbf{A}_{ij} = \Pi(i|j)$ 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 $\mathbf{B}_{\alpha i} = \Pi(\alpha|i)$ 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 (Kemeny and Snell, 1960):

$$\mathbf{P} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{B} & \mathbf{1}_2 \end{bmatrix}, \tag{B.2}$$

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:

$$\begin{aligned} \lim_{m\to\infty}\mathbf{P}^m &= \lim_{m\to\infty}\begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{B} & \mathbf{1}_2 \end{bmatrix}^m \\ &= \lim_{m\to\infty}\begin{bmatrix} \mathbf{A}^m & \mathbf{0} \\ \mathbf{B}(\mathbf{1}_{N-1} + \mathbf{A} + \cdots + \mathbf{A}^{m-1}) & \mathbf{1}_2 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{B}(\mathbf{1}_{N-1} - \mathbf{A})^{-1} & \mathbf{1}_2 \end{bmatrix}, \end{aligned} \tag{B.3}$$

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

$$(\mathbf{1}_{N-1} - \mathbf{A})^{-1} = \sum_{j=0}^{\infty} \mathbf{A}^j. \tag{B.4}$$

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 $n = 1$):

$$\phi(r) = (\mathbf{B}(\mathbf{1}_{N-1} - \mathbf{A})^{-1})_{2,1}. \tag{B.5}$$

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

$$\phi(r) = \mathbf{B}_{2,1} + \sum_{i=1}^{N-1} \mathbf{B}_{2,i}\mathbf{A}_{i,1} + \sum_{i,j=1}^{N-1} \mathbf{B}_{2,i}\mathbf{A}_{i,j}\mathbf{A}_{j,1} + \cdots. \tag{B.6}$$

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 exactly one generation, the second term is the probability of fixing in exactly two generations, etc.

For small population sizes N, Eq. (B.5) can be evaluated explicitly:

$$\begin{array}{c|c} N & \phi(r) \\ \hline 2 & \frac{r^2}{1+r^2} \\ 3 & \frac{r^3(8r^3+48r^2+6r+1)}{8r^6+48r^5+6r^4+65r^3+6r^2+48r+8} \\ \vdots & \vdots \\ N & \frac{r^N a_N(r)}{b_N(r)} \end{array} \tag{B.7}$$

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. Eq. (B.5) can be rewritten as

$$(\mathbf{1}_{N-1} - \mathbf{A})^T \mathbf{u}^T = \mathbf{B}^T, \tag{B.8}$$

where $\mathbf{u}$ is the $2 \times (N-1)$ matrix of fixation and extinction probabilities from all initial mutant fractions. 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).

References

Achtman, M., 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. Annu. Rev. Microbiol. 62, 53–70.

Allen, L.J.S., 2011. An Introduction to Stochastic Processes with Applications to Biology, second ed. Chapman and Hall, CRC, Boca Raton.

Barrett, R., MacLean, R., Bell, G., 2006a. Mutations of intermediate effect are responsible for adaptation in evolving *Pseudomonas fluorescens* populations. Biol. Lett. 2, 236–238.

Barrett, R., M'Gonigle, L., Otto, S., 2006b. The distribution of beneficial mutant effects under strong selection. Genetics 174, 2071–2079.

Barry, D., Hartigan, J.A., 1987. Asynchronous distance between homologous DNA sequences. Biometrics 43, 261–276.

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

Berg, J., Willman, S., Lässig, M., 2004. Adaptive evolution of transcription factor binding sites. BMC Evol. Biol. 4, 42.

Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N., Tawfik, D.S., 2006. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. Nature 444, 929–932.

Bloom, J.D., Glassman, M.J., 2009. Inferring stabilizing mutations from protein phylogenies: application to influenza hemagglutinin. PLoS Comput. Biol. 5, e1000349.

Bloom, J.D., Raval, A., Wilke, C.O., 2007. Thermodynamics of neutral protein evolution. Genetics 175, 255–266.

Bull, J., Badgett, M., Wichman, H., 2000. Big-benefit mutations in a bacteriophage inhibited with heat. Mol. Biol. Evol. 17, 942–950.

Bulmer, M.G., 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129, 897–907.

Bürger, R., 2000. The Mathematical Theory of Selection, Recombination, and Mutation. Wiley, New York.

Cannings, C., 1974. The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. Adv. in Appl. Probab. 6, 260–290.

Champagnat, N., 2006. A microscopic interpretation for adaptive dynamics trait substitution sequence models. Stochastic Process. Appl. 116, 1127–1160.

Champagnat, N., Ferrière, R., Méléard, S., 2006. Unifying evolutionary dynamics: from individual stochastic processes to macroscopic models. Theor. Popul. Biol. 69, 297–321.

Crow, J.F., Kimura, M., 1970. An Introduction to Population Genetics Theory. Harper and Row, New York.

Darwin, C., 1859. The Origin of Species. J. Murray, London.

DePristo, M., Weinreich, D., Hartl, D., 2005. Missense meanderings in sequence space: a biophysical view of protein evolution. Nat. Rev. Genet. 6, 678–687.

Desai, M.M., Fisher, D.S., 2007. Beneficial mutation-selection balance and the effect of linkage on positive selection. Genetics 176, 1759–1798.

Dos Vultos, T., Mestre, O., Rauzier, J., Golec, M., Rastogi, N., Rasolofo, V., Tonjum, T., Sola, C., Matic, I., Gicquel, B., 2008. Evolution and diversity of clonal bacteria: the paradigm of mycobacterium tuberculosis. PLoS One 3, e1538EP.

Eigen, M., McCaskill, J., Schuster, P., 1989. The molecular quasi-species. Adv. Chem. Phys. 75, 149–263.

Ewens, W., 1967. The probability of survival of a new mutant in a fluctuating environment. Heredity 22, 438–443.

Ewens, W.J., 1990. Population genetics theory—the past and the future. In: Lessard, S. (Ed.), Mathematical and Statistical Developments of Evolutionary Theory. Kluwer Academic Publishers, Amsterdam, pp. 177–227.

Ewens, W.J., 2004. Mathematical Population Genetics. Springer, New York.

Eyre-Walker, A., Keightley, P., 2007. The distribution of fitness effects of new mutations. Nat. Rev. Genet. 8, 610–618.

Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.

Felsenstein, J., 2011. PHYLIP (Phylogeny inference package) version 3.69.

Fisher, R.A., 1958. The Genetical Theory of Natural Selection. Dover, New York.

Gillespie, J.H., 1974. Natural selection for within-generation variance in offspring number. Genetics 76, 601–606.

Gillespie, J.H., 1975. Natural selection for within-generation variance in offspring number II. Discrete haploid models. Genetics 81, 403–413.

Gillespie, J.H., 1977. Natural selection for variances in offspring numbers: a new evolutionary principle. Am. Nat. 111, 1010–1014.

Holder, K., Bull, J., 2001. Profiles of adaptation in two similar viruses. Genetics 159, 1393–1404.

Iwasa, Y., 1988. Free fitness that always increases in evolution. J. Theoret. Biol. 135, 265–281.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, pp. 21–123.

Kemeny, J.G., Snell, J.L., 1960. Finite Markov Chains. Van Nostrand, New York.

Kimura, M., 1955. Solution of a process of random genetic drift with a continuous model. Proc. Natl. Acad. Sci. USA 41, 144–150.

Kimura, M., 1957. Some problems of stochastic processes in genetics. Ann. Math. Stat. 28, 882–901.

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

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.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Kimura, M., Ohta, T., 1971. Theoretical Aspects of Population Genetics. Princeton University Press, Princeton.

Kingman, J.F.C., 1982. The coalescent. Stoch. Proc. Appl. 13, 235–248.

Kurtz, T.G., 1981. Approximation of Population Processes. In: CBMS-NSF Regional Conference Series in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia.

Lässig, M., 2007. From biophysics to evolutionary genetics: statistical aspects of gene regulation. BMC Bioinformatics 8, S7.

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.

Lenski, R., Elena, S.F., 2003. Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation. Nat. Rev. Genet. 4, 457–469.

Levikson, B., 1977. The age distribution of Markov processes. J. Appl. Probab. 14, 492–506.

Levin, D.A., Peres, Y., Wilmer, E.L., 2009. Markov Chains and Mixing Times. American Mathematical Society, Providence, RI.

Li, W.-H., 1987. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. 24, 337–345.

Maruyama, T., 1970. On the fixation probability of mutant genes in a subdivided population. Genet. Res. Camb. 15, 221–225.

McVean, G.A., Charlesworth, B., 1999. A population genetics model for the evolution of synonymous codon usage: patterns and predictions. Genet. Res. 74, 145–158.

McVean, G.A., 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.

McVean, G.A.T., Vieira, J., 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. Genetics 157, 245–257.

Möhle, M., 2001. Forward and backward diffusion approximations for haploid exchangeable population models. Stoch. Proc. Appl. 95, 133–149.

Möhle, M., Sagitov, S., 2001. A classification of coalescent processes for haploid exchangeable population models. Ann. Probab. 29, 1547–1562.

Moran, P.A.P., 1958. Random processes in genetics. Proc. Cambridge Philos. Soc. 54, 60–71.

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.

Mustonen, V., Kinney, J., Callan, C.G., 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.

Mustonen, V., Lässig, M., 2010. Fitness flux and the ubiquity of adaptive evolution. Proc. Natl. Acad. Sci. USA 107, 4248–4253.

Nielsen, R., DuMont, V.L.B., Hubisz, M.J., Aquadro, C.F., 2007. Maximum likelihood estimation of ancestral codon usage bias parameters in Drosophila. Mol. Biol. Evol. 24, 228–235.

Ochman, H., Selander, R.K., 1984. Evidence for clonal population structure in Escherichia coli. Proc. Natl. Acad. Sci. 81, 198–201.

Ohta, T., 1992. Theoretical study of near neutrality. II. Effect of subdivided population structure with local extinction and recolonization. Genetics 130, 917–923.

Ohta, T., Tachida, H., 1990. Theoretical study of near neutrality. I. Heterozygosity and rate of mutant substitution. Genetics 126, 219–229.

Orr, H., 2001. The genetics of species differences. Trends Ecol. Evol. 16, 343–350.

Orr, H., 2005. The genetic theory of adaptation: a brief history. Nat. Rev. Genet. 6, 119–127.

Otto, S.P., Whitlock, M.C., 1997. The probability of fixation in populations of changing size. Genetics 146, 723–733.

Parsons, T.L., Quince, C., Plotkin, J.B., 2010. Some consequences of demographic stochasticity in population genetics. Genetics 185, 1345–1354.

Poelwijk, F.J., Kiviet, D.J., Weinreich, D.M., Tans, S.J., 2007. Empirical fitness landscapes reveal accessible evolutionary paths. Nature 445, 383–386.

Proulx, S.R., 2000. The ESS under spatial variation with applications to sex allocation. Theor. Popul. Biol. 58, 33–47.

Roberts, F.S., 1979. Measurement Theory with Applications to Decision Making, Utility, and the Social Sciences. Addison-Wesley, Reading.

Rodríguez, F., Oliver, J., Marín, A., Medina, J., 1990. The general stochastic model of nucleotide substitution. J. Theoret. Biol. 142, 485–501.

Rouzine, I.M., Rodrigo, A., Coffin, J.M., 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.

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.

Sella, G., Hirsh, A., 2005. The application of statistical physics to evolutionary biology. Proc. Natl. Acad. Sci. USA 102, 9541–9546.

Shpak, M., 2007. Selection against demographic stochasticity in age-structured populations. Genetics 177, 2181–2194.

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.

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.

Wakeley, J., 2005. The limits of theoretical population genetics. Genetics 169, 1–7.

Watterson, G.A., 1976. Reversibility and the age of an allele. I. Theor. Popul. Biol. 10, 239–253.

Watterson, G.A., 1977. Reversibility and the age of an allele. II. Theor. Popul. Biol. 12, 179–196.

Weinreich, D.M., Delaney, N.F., DePristo, M.A., Hartl, D.L., 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312, 111–114.

Whitlock, M., 2003. Fixation probability and time in subdivided populations. Genetics 164, 767–779.

Wichman, H., Badgett, M., Scott, L., Boulianne, C., Bull, J., 1999. Different trajectories of parallel evolution during viral adaptation. Science 285, 422–424.

Wick, L.M., 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.

Wright, S., 1931. Evolution in Mendelian populations. Genetics 16, 97–159.

Wright, S., 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: Proc. 6th Int. Cong. Genet., vol. 1, 356–366.

Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press, Oxford.

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.

Chapter 4

Yeast TF Evolution

Most traditional studies of molecular evolution rely on simplified models of fitness landscapes, or reconstruct the landscapes empirically based on limited experimental data [5]. However, fitness landscapes are fundamentally shaped by complex molecular 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.

Here we consider evolution of TF binding sites in the yeast *S. cerevisiae*, using the results of the previous chapters. We apply the model of monomorphic evolution according to a substitution process described in chapters 2 and 3 to a collection of 25 *S. cerevisiae* TFs for which models of TF binding affinity and specificity were built using high-throughput *in vitro* measurements of TF-DNA interactions [1]. Our goal is study how energetics of protein-DNA interactions affects the structure of the fitness landscape. That is we infer fitness landscapes as a function of TF binding energy, from observed distributions of TF binding sites in the yeast genome, and rationalize it in terms of a two-state thermodynamic model of TF-DNA binding, as described in Chapter 1. Our analysis sheds light on the genome-wide importance of TF-DNA interactions in regulatory site evolution.

This evolutionary model makes a number of important assumptions, which we shall test for yeast populations: We have assumed the population is monomorphic, in steady state, and that the collection of binding sites in the genome are under similar selection pressures — that is 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 [127]. 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.

4.1 Biophysical model of binding site evolution

Here we shall briefly review the evolutionary model developed in earlier chapters.

As discussed in Chapter 1, the probability of a binding site to be TF-bound is given by the Fermi-Dirac function of the free energy E of TF-DNA interaction [7]:

$$p_{\text{bound}}(E) = \frac{1}{1 + e^{\beta(E-\mu)}}, \tag{4.1}$$

where β is the inverse temperature (≈ 1.7 (kcal/mol)$^{-1}$ at room temperature) and μ is the chemical potential, a function of the TF concentration. Note that $p_{\text{bound}}(E) \approx e^{-\beta(E-\mu)}$ if $E \gg \mu$, resulting in a Boltzmann-like exponential distribution. Under the additive energy model, the energy of a sequence is given by

$$E(\sigma) = \sum_{i=1}^{L} \epsilon_i^{\sigma_i} \tag{4.2}$$

where $\epsilon_i^{\sigma_i}$ are the elements of an 'energy matrix' (EM) representing the energetic contribution of individual nucleotides to the binding energy.

In practice, energy matrices are sometimes expressed as position-specific affinity matrices (PSAMs) or position weight matrices (PWMs), which specify the probability of binding rather than the free energy contribution of nucleotides at each position in a binding site. When the total energy E is greater than the chemical potential μ, $p_{\text{bound}}(E) \approx e^{-\beta(E-\mu)}$, and hence the affinities are related to the energies by an exponential. We use EMs rather than PSAMs/PWMs since EMs are more general: they are applicable to binding on the non-exponential regime of $p_{\text{bound}}(E)$ (when $E - \mu < 0$) and also are readily generalized to more complex models of sequence-dependent energies, such as those with contributions from dinucleotides. While we will use data for the additive model only, our general framework is straightforwardly applicable to more complex cases given appropriate energetic data.

As discussed in chapters 1 to 3, the evolution of a TF binding site may be modeled as a substitution process, assuming the population is monomorphic, that is that $u \ll (LN_e \log N_e)^{-1}$, where u here is the per-nucleotide mutation rate (probability of mutation per base per generation), L is the number of bases in the locus, and N_e is an effective population size [128]. We will assume that the locus of interest is unlinked to the rest of the genome (linkage equilibrium) by frequent recombination with rate ρ, which satisfies $\rho \gg N\mu L$ [52]. Therefore we can consider the evolution of the locus independently from the rest of the genome, so hitchhiking and polymorphism elsewhere can be neglected. We also assume that the locus is short enough that recombination does not occur within the locus itself. In general, we assume lengths of $L \sim \mathcal{O}(10)$ to realistically meet these conditions. Thus the collection of binding sites for a particular TF are independent loci evolving in parallel.

As discussed in Chapter 3, in evolutionary steady state, the probability that the population has binding energy E at the locus is given by

$$\pi(E) = \frac{1}{Z}\pi_0(E)\mathcal{F}(E)^{\nu}. \tag{4.3}$$

where $\mathcal{F}(E)$ is the multiplicative fitness (defined so that the total fitness of a set of independently evolving loci is a product of fitnesses of each one), $\pi_0(E)$ is the neutral distribution of sequences (steady state under no selection), and Z is a normalization constant. As previously discussed, this equation is is applicable to a wide class of population models. As discussed in Chapter 2, the exponent ν is a “scaling” effective population size which is closely related to the standard variance effective population size N_e. For example, $\nu = 2(N_e - 1)$ in the Wright-Fisher model and $\nu = N_e - 1$ in the Moran model of population genetics [55]. Conceptually, both ν with N_e measure the strength of genetic drift [57].

The neutral distribution $\pi(0)$ is approximately a Gaussian distribution of energy in the limit of long sequence lengths, and can be seen as an entropic bias leading the steady state distribution $\pi(E)$ towards the peak of this Gaussian, at high energies. On the other hand, the term $f(E)^{\nu}$ biases the steady state towards low energies where the fitness is highest. Thus, the final position of the steady state is determined by the balance between these driving forces of selection and entropy.

We assume that a site contributes fitness 1 to the organism when it is bound, and fitness $f_0 < 1$ otherwise. Then the fitness contribution, averaged over the bound and unbound states of the TF-DNA complex, is given by

$$\mathcal{F}(E) = \frac{1 + f_0 e^{\beta(E-\mu)}}{1 + e^{\beta(E-\mu)}}. \tag{4.4}$$

as discussed in Chapter 1, equation 1.4.

We may invert Eq. 4.3 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 [19]:

$$\log\left(\frac{\pi(E)}{\pi_0(E)}\right) = \nu \log \mathcal{F}(E) - \log Z. \tag{4.5}$$

Thus given a distribution of evolved binding site sequences π and a neutral distribution π_0, we can use Eq. 4.5 to infer the logarithm of the fitness landscape up to an overall scale and shift. Moreover, given a specific functional form of $\mathcal{F}(E)$, such as the Fermi-Dirac fitness in Eq. 4.4, we can perform a maximum likelihood fit of the observed sequence distribution to infer values of parameters β, μ, ν, and f_0. Indeed, this is our goal.

Degeneracies in the parameter space of this model are an important issue. 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. This is a general property of a model where fitness is an average over two possible phenotypes. Consider a general fitness function

$$\mathcal{F}(\sigma) = p(\sigma) + f_0(1 - p(\sigma)), \tag{4.6}$$

where one phenotype has fitness 1 and occurs with probability $p(\sigma)$, and the other phenotype has fitness f_0 and occurs with probability $1 - p(\sigma)$. In the case of binding sites, the phenotypes are TF-bound and TF-unbound, and $p(\sigma)$ is a Fermi-Dirac function projected from the genotype σ to the energy (Eq. 4.1). The steady-state distribution (Eq. 4.3) depends on the quantity $\mathcal{F}(\sigma)^\nu$, which can be written as:

$$\mathcal{F}(\sigma)^\nu = (1 - \frac{1}{\nu}\gamma(1 - p(\sigma)))^\nu \approx e^{-\gamma(1-p(\sigma))} \tag{4.7}$$

if $\gamma(1 - p(\sigma)) \ll \nu$ or, since $0 \leq 1 - p(\sigma) \leq 1$, 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. 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 depend only on the strength of selection relative to the strength of drift [55, 65], Ns, or in our model, $\nu(1 - f_0) = \gamma$. 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 $Ns = \gamma$ may still be large.

4.1.1 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

$$s(E) = \frac{\mathcal{F}(E + \Delta E)}{\mathcal{F}(E)} - 1 \approx \frac{d \log \mathcal{F}}{dE} \Delta E. \tag{4.8}$$

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

$$\tilde{s}(E) = \left| \frac{d}{dE} \log \mathcal{F}(E) \right| = \frac{\beta(1 - f_0)z}{(1 + z)(1 + f_0 z)}, \quad \text{where} \quad z = e^{\beta(E-\mu)}. \tag{4.9}$$

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. 4.1 we show $\log \mathcal{F}(E)$, $\tilde{s}(E)$, and the derivative

$$\frac{\partial \tilde{s}}{\partial \beta} = \frac{z(1 - f_0)}{(1 + z)^2(1 + f_0 z)^2}[(1 - f_0 z^2) \log z + (1 + z)(1 + f_0 z)]. \tag{4.10}$$

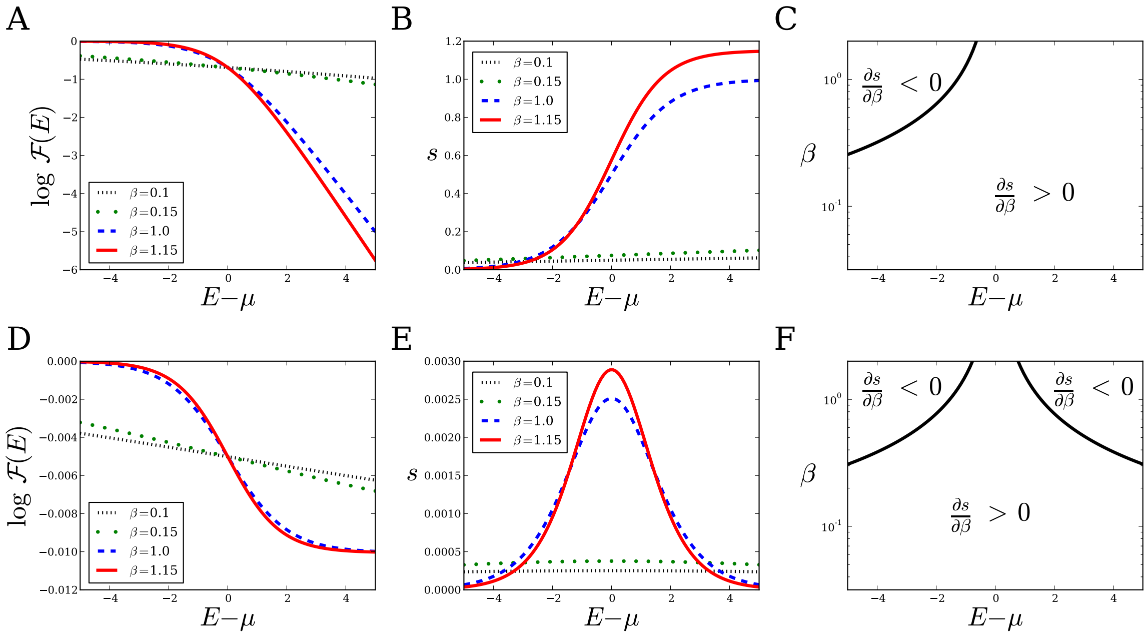


Figure 4.1: Fitness and selection strength plots as functions of energy $E-\mu$ (measured with respect to chemical potential μ) and inverse temperature β. Top row uses $f_0 = 0$; bottom row uses $f_0 = 0.99$. (A,D) Logarithm of Fermi-Dirac fitness versus energy for several values of β; note that the 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. 4.10) with $f_0 = 0.99$.

For $f_0 = 0$ (Fig. 4.1A–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. 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. 4.1C).

This situation changes qualitatively in the regime $E-\mu > 0$ when $f_0 \neq 0$ (Fig. 4.1D-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. 4.10, the boundaries in Fig. 4.1F 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. 4.1F are thus given by the curves $\beta(E - \mu) = \log z_1^*$ for $E - \mu < 0$ and $\beta(E - \mu) = \log z_2^*$ for $E - \mu > 0$.

4.2 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. 1.4).

4.2.1 Methods: A model system to check the assumptions of monomorphism and steady state

We consider a haploid asexual Wright-Fisher process [55]. The population consists of $N_e = 1000$ 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_e. 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. 4.3), and the distribution observed in simulations, π_{obs}, using the total variation distance (TVD):

$$\Delta(\pi_{\text{exp}}, \pi_{\text{obs}}) = \frac{1}{2} \sum_x |\pi_{\text{exp}}(x) - \pi_{\text{obs}}(x)|. \tag{4.11}$$

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. 4.11 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 EM into 100 bins of equal size).

We begin by randomly generating the EM parameters $\epsilon_i^{\sigma_i}$. Each $\epsilon_i^{\sigma_i}$ in the EM 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 EM by a factor χ:

$$\chi^2 = \sum_{i=1}^{L} \sum_{\alpha \in \{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}} \pi_0(\alpha)(\epsilon_i^{\alpha} - \bar{\epsilon}_i)^2 \tag{4.12}$$

where ϵ_i^{α} is the EM element for base α at position i, $L = 10$ is the binding site length, $\bar{\epsilon_i} = \sum_{\alpha \in \{\mathsf{A},\mathsf{C},\mathsf{G},\mathsf{T}\}} \epsilon_i^{\alpha}$ is the average energy contribution at position i, and $\pi_0(\alpha)$ is the background probability of nucleotide α (0.25, $\forall \alpha$ in our simulations). It can be shown that χ is the standard deviation of the random sequence energy distributution, which is approximately Gaussian [15]. We generate the EM 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. 1.4 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 EM 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.

4.2.2 The effect of polymorphism

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. 4.2A). From π_{obs}, we infer the fitness landscape as a function of energy using Eq. 4.5 (Fig. 4.2B).

Additionally, for each u we record the average number of unique sequences present in the population at equilibrium, and compute the total variation distance (TVD; Eq. 4.11) between π_{obs} and the monomorphic prediction (Fig. 4.2C). 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. 4.3 overestimates the fitness of low-affinity sites. The population becomes distinctly 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 $u \ll (LN_e \log N_e)^{-1}$ [128]. Indeed, in the monomorphic limit the expected time between new mutations, $(LN_e u)^{-1}$, must be longer than the expected time over which fixation occurs, which is $\mathcal{O}(N_e)$ generations with probability $1/N_e$ for mutants that fix, and $\mathcal{O}(\log N_e)$ with probability $(N_e - 1)/N_e$ for mutants that go extinct. Thus the total expected time before the mutant either fixes or goes extinct is $\mathcal{O}(\log N_e)$ generations for $N_e \gg 1$ [129]. Thus we must have $(LN_e u)^{-1} \gg \log N_e$ or, equivalently, $u \ll (LN_e \log N_e)^{-1}$. Using $N_e = 1000$ and $L = 10$ as in our simulations yields $u \ll 1.4 \times 10^{-5}$ in the monomorphic limit, consistent with the results in Fig. 4.2C.

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. 4.3A–C). When the

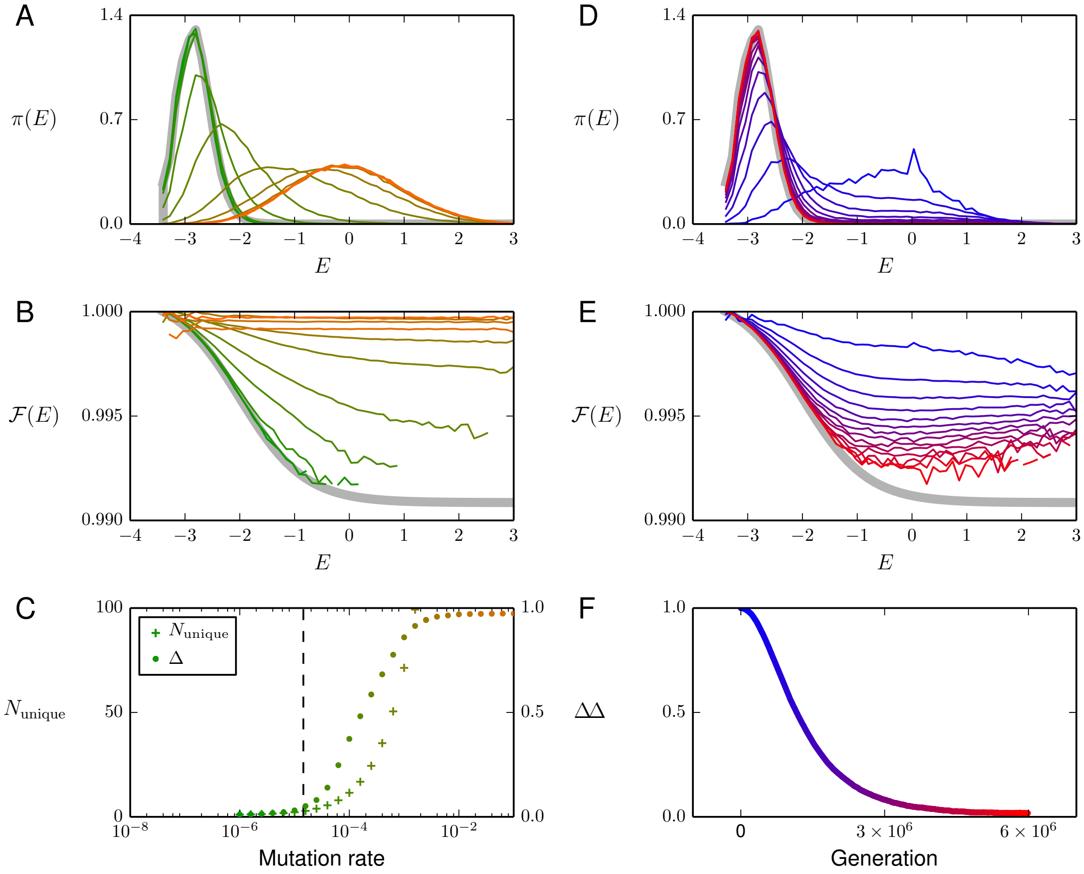


Figure 4.2: The monomorphic limit and steady state of a Wright-Fisher model of population genetics. In (A)–(C) we show results from simulations at various mutation rates, using a fitness function with $f_0 = 0.99$, $\beta = 1.69$ (kcal/mol)$^{-1}$, and $\mu = -2$ kcal/mol. Each mutation rate data point is an average over 10^5 independent runs, as described in Methods. Colors from green to orange correspond to increasing mutation rates. (A) Observed steady-state distributions $\pi_{\text{obs}}(E)$ for various mutation rates. The steady state $\pi(E)$ predicted using Eq. 4.3 is shown in grey. (B) Fitness functions $\mathcal{F}(E)$ predicted using observed distributions $\pi_{\text{obs}}(E)$ in Eq. 4.5. The exact fitness function is shown in gray. Inferred fitness functions are matched to the exact one by using the known population size N_e, and setting the maximum fitness to 1.0 for each curve. (C) For each mutation rate, the total variation distance (TVD) Δ between $\pi_{\text{obs}}(E)$ and $\pi(E)$, and the average number of unique sequences in the population N_{unique} (the degree of polymorphism) are shown. The predicted bound $(N_e L \log N_e)^{-1}$ 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 Methods.

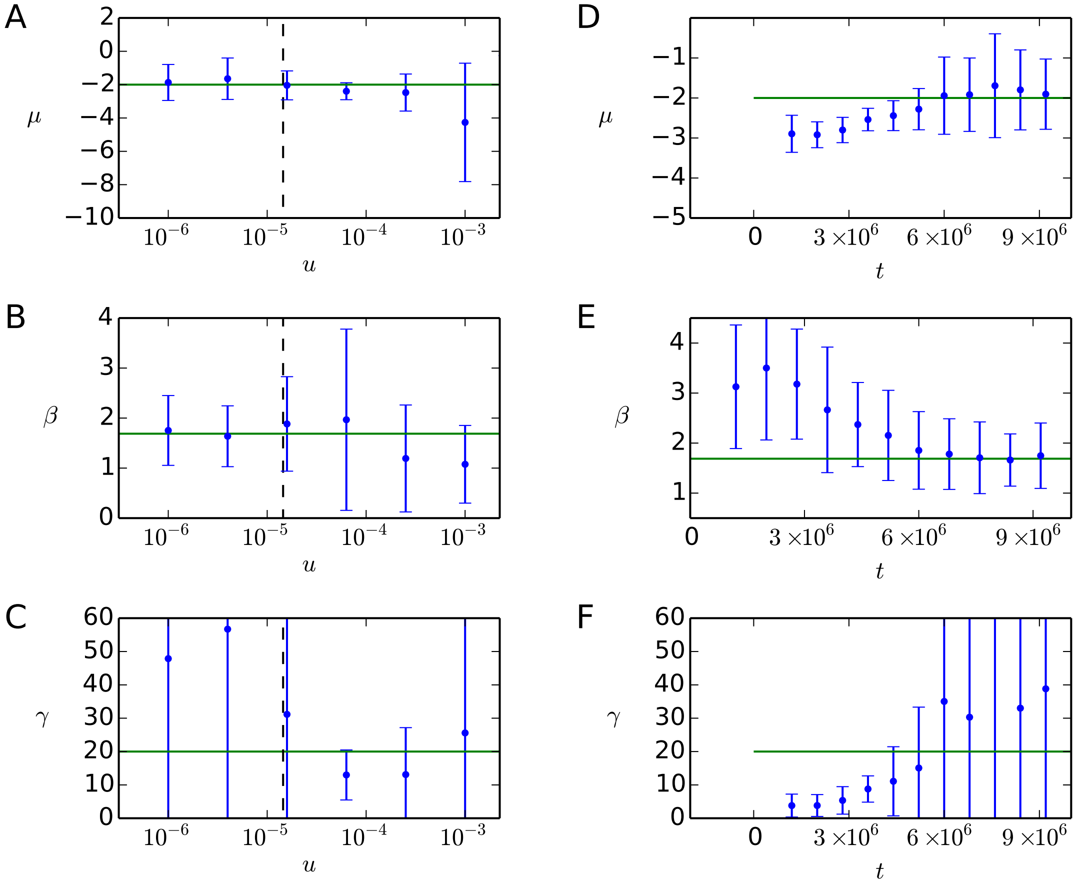


Figure 4.3: Fitted parameters of the Fermi-Dirac function from Wright-Fisher simulations. In (A)–(C) the fitted values of μ, β and $\gamma = \nu(1 - f_0)$ are shown as functions of mutation rate u. For each mutation rate, we generate 200 random samples of 500 sequences from the 10^5 sequences generated in simulations used in Fig. 4.2A–C. We fit the parameters of the fitness function on each sample separately by maximum likelihood (see Methods). 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 $(N_e L \log N_e)^{-1}$ 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 equilibration simulations used in Fig. 4.2D–F. The sampling procedure, the maximum likelihood fit, and the representation of parameter predictions are the same as in (A)–(C).

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. 4.3A,B). For γ, polymorphism has no clear bias on the average inferred value, although it also appears to increase the variation.

4.2.3 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 Methods (Fig. 4.2D), and infer the fitness function (Fig. 4.2E). We also compute the TVD between the observed distribution π_{obs} and the steady-state prediction (Fig. 4.2F). With time, π_{obs} converges to the steady state (Eq. 4.3) and the TVD decays to zero, enabling accurate reconstruction of the fitness function in the threshold region (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. 4.2F places the equilibration timescale at about 4×10^6 generations. As the population equilibrates, accurate inference of the fitness function parameters becomes possible (Fig. 4.3D-F). We see that parameters inferred from a population out of steady tend to underestimate μ and γ and overestimate β.

The application of Eq. 4.5 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. 2.1.1, the monomorphic limit requires $\mu \ll 1/(LN \log N)$ for neutral evolution [91, 92], or $\mu \ll 1/(LN \log(Ns))$ in the presence of beneficial mutations with a typical selection coefficient s, $1 \ll Ns \ll N$ [56]. Assuming that deleterious substitutions are negligible with regard to reaching steady state (due to exponential suppression of their substitution rates), equilibration will be dominated by neutral evolution. Equation 2.1 then implies that the neutral substitution rate is equal to the mutation rate, which is much less than $1/(LN \log N)$ per generation.

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 [130]. 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 average time to make a single neutral substitution step, $LN \log N$, we estimate that evolutionary steady state will be reached in

$$(LN \log N) \times (2L \log(2L)) \text{ generations.} \tag{4.13}$$

For small but nontrivial genomic loci ($L = \mathcal{O}(10)$ bp) 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 exploited this projection to describe evolution of TF binding sites in *S. cerevisiae*, in which the distribution of binding sites has been projected onto free energies of TF-DNA binding [17–20]. The steady state is expected to be reached more quickly in the one-dimensional energy space than in the high-dimensional genotype space [20]. [20] 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 inhibit TF binding [131, 132], 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. **??** to genomic data such as collections of TF binding sites. Current work is in progress to apply these results to evolution of regulatory sites in yeast, exploring the biophysical origins of the underlying fitness landscapes.

4.3 Transcription factor binding sites in yeast

How well does *S. cerevisiae* satisfy the assumptions of our evolutionary model? *S. cerevisiae* is not a purely haploid organism but 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 [133]. Based on the analysis of yeast genomes, wild yeast populations show extremely limited outcrossing and recombination and are geographically distinct [134]. Thus, *S. cerevisiae* may be regarded as haploid to a reasonable approximation, with recombination during the diploid stages unlinking the loci. This is consistent with our model, which assumes a haploid population and independent evolution of binding sites.

Are natural populations of *S. cerevisiae* 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 [133]. Assuming loci of length $L = 10$, this sets a bound on the effective population size N_e 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 [133], 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 Eq. 4.3. Furthermore, in *S. cerevisiae* and *S. paradoxus* the proportion of polymorphic sites in a population has been found to be about 0.001 [133, 135, 136], generally with no more than two alleles segregating at any one site [133]. 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. 4.2C.

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 [137]. This is several times less than the 5–10 million years of divergence time for the most recent speciation event, with *S. paradoxus* [138]. Thus steady state may plausibly be reached for a fast-reproducing organism like *S. cerevisiae* over evolutionary times scales.

4.3.1 Site-specific selection

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. 4.5. 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.

4.3.2 Methods

Binding site and EM data

We obtain curated binding site locations for 125 TFs from Ref. [2], 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. Fro 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 [1]. This study provides PSAMs for each TF determined using the MatrixREDUCE package [46]. We convert the elements of the PSAM $w_{i\alpha}$ to EM 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. [2]. We neglect PHO4 since it does not have any binding sites above the 0.9 threshold of Ref. [2], leaving us with 25 TFs for which both EM and a set of genomic binding sites are available. We align the binding site sequences from Ref. [2] to the corresponding EMs, choosing the alignment that produces the lowest average binding energy for the sites.

Essentiality data

The Yeast Deletion Database classifies genes as essential, tested (nonessential), and unavailable, which number 1156, 6343, and 529 respectively [127, 139]. 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 [165] of *S. cerevisiae* to those of *S. paradoxus* using ClustalW [166]. We measure the rate of nonsynonymous mutations using PAML [167]. 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. [146], 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* [2]. 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.

4.3.3 Results

We classify fitness effects of genes using knockout lethality, which is available in the Yeast Deletion Database [127, 139]. 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. 4.4A 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 Methods, we find no significant difference (at $p = 0.05$ level) between the two groups of sites for any TF except RPN4, for which $p = 0.03$ and the difference in mean energies is 0.24 kcal/mol, and PDR3, for which $p = 0.002$ and the difference in mean energies is 2.3 kcal/mol. The mean p-value of the null model over all TFs is 0.38. In Fig. 4.4B 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 (Table C.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 0.51. We find no significant correlation for any TF at $p = 0.05$ level except MSN2, with $p = 0.046$.

It is possible that regulation of highly-expressed genes may be more tightly controlled. Indeed, gene expression level is weakly, though significantly, correlated with gene essentiality [140]. 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 [140] (Table C.1, column C), and again find no correlation using the Spearman rank correlation except for

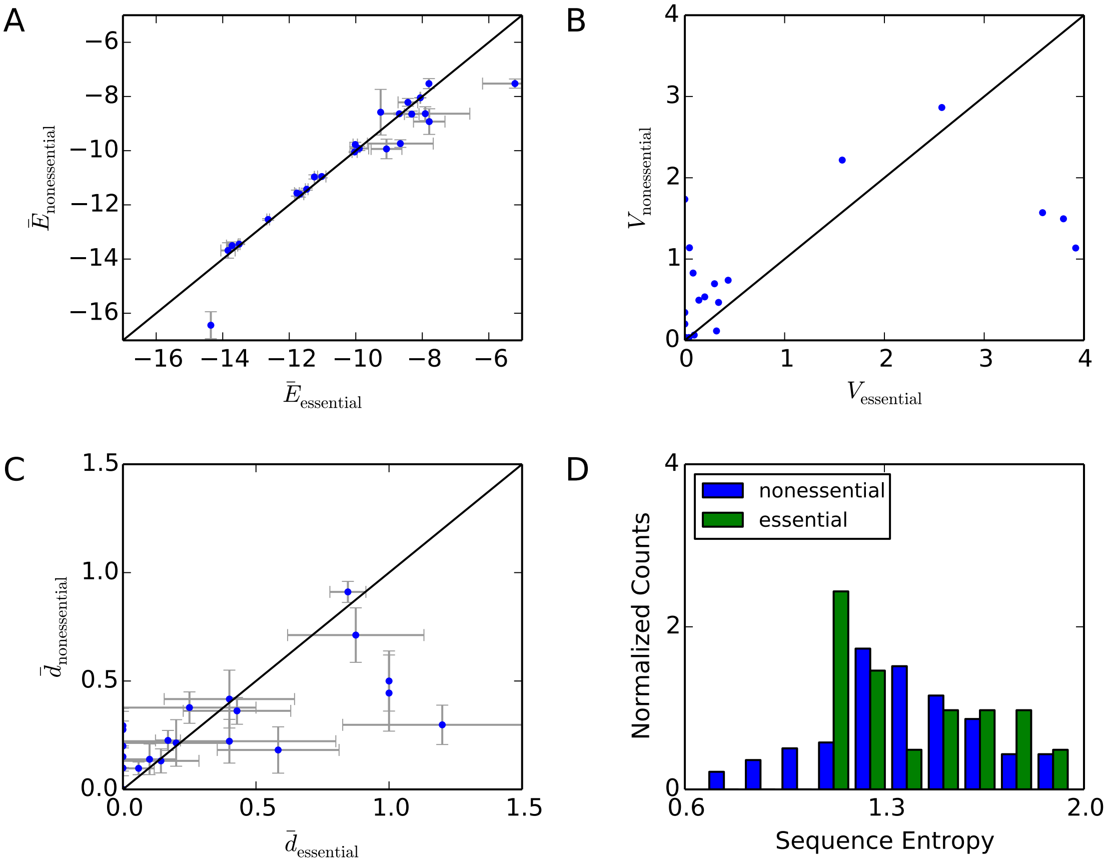


Figure 4.4: Tests of site-specific selection. We divide binding sites for each TF into two groups: those regulating essential and nonessential genes. (A) Comparison of mean binding energies of sites regulating essential ($\bar{E}_{\text{essential}}$) and nonessential genes ($\bar{E}_{\text{nonessential}}$) for each TF in the data set. 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. (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. Vertical and horizontal error bars show the standard error of the mean in each group. In (A)–(C), 25 TFs were used; black diagonal lines have slope one. (D) Normalized histogram of TF binding site sequence entropies, divided into 16 essential and 109 nonessential TFs, for 125 TFs in Ref. [2].

DAL80 ($p = 0.034$), with mean p-value of 0.54.

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 Methods, 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 (Table C.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 0.42.

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* [2] is no different for sites regulating essential genes than for those regulating nonessential genes, as shown in Fig. 4.4C. Using the null model described in Methods, most TFs are above $p = 0.05$ with the exceptions of YAP7 ($p = 0.04$) and PDR3 ($p = 0.003$), with an average p-value of 0.27.

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. [2] 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 [127, 139], and calculate the sequence entropy of binding sites for each TF. The distribution of sequence entropies in Fig. 4.4D 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 0.59 and all p-values above 0.05 (Table C.1, column E). Overall, our findings are in broad agreement with a previous report [20], 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 is consistent with the variance predicted by a model including only drift and site-independent selection.

4.4 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. 4.3 and 4.5. 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. 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 [141]; however, consistent with previous reports [20, 142], 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. 4.5A for REB1 and in table C.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. 4.5 to reconstruct the fitness landscapes as a function of TF binding energy up to an overall scale and shift (Fig. 4.6). Although the fitness functions may be noisy due to imperfect sampling of $\pi(E)$, they nevertheless provide important qualitative insights. In particular, in all landscapes fitness decreases monotonically as binding energy increases, indicating that stronger-binding sites are more fit. Moreover, we observe no fitness penalty for binding too strongly, at least within the range of energies spanned by $\pi(E)$.

Methods: Neutral binding site energy distributions

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

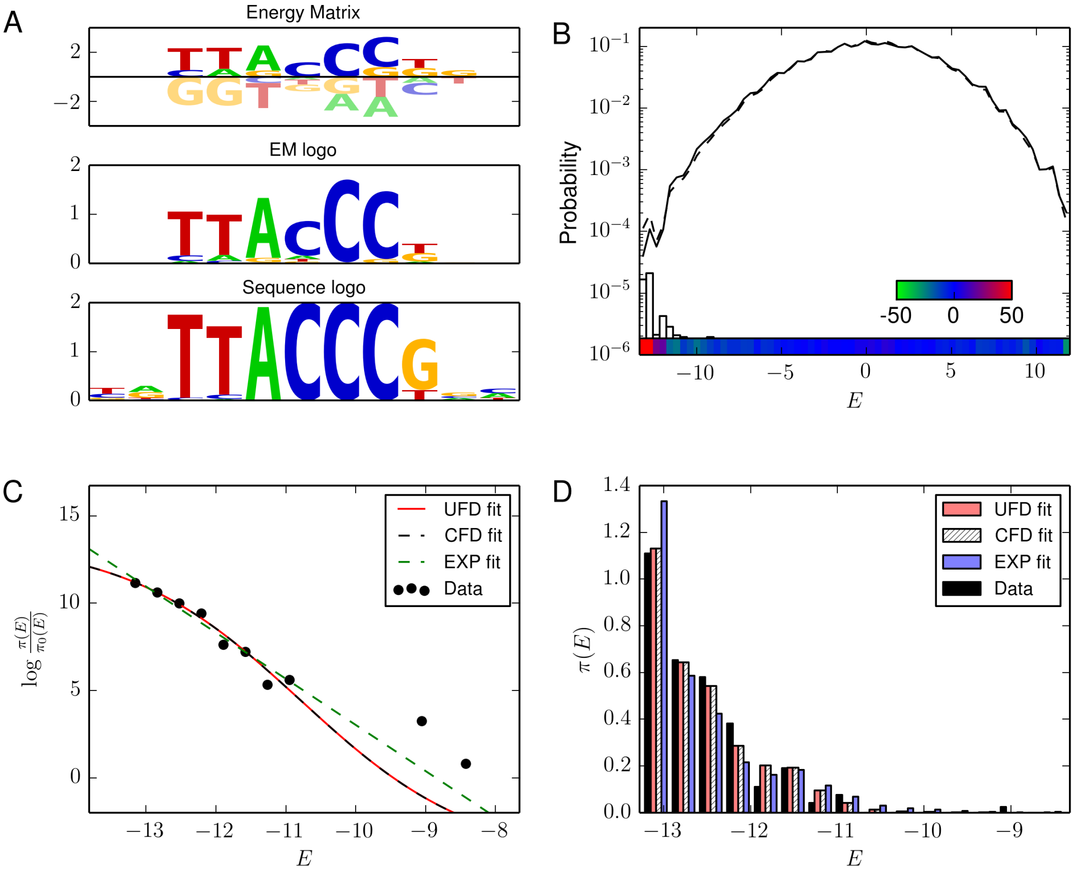


Figure 4.5: Parametric inference of REB1 fitness landscape. (A) From top to bottom: REB1 EM [1], the sequence logo obtained from the EM by assuming a Boltzmann distribution at room temperature at each position in the binding site ($\pi^i(\sigma_i) = \pi_0^i(\sigma_i)e^{-\beta\epsilon_i^{\sigma_i}}/Z_i$), and the sequence logo based on the alignment of REB1 genomic sites. (B) Histogram of energies of intergenic sites calculated using the REB1 EM (dashed line). The neutral distribution of sequence energies expected from the mono- and dinucleotide background model (solid line; see Methods for details). The histogram shows the distribution of functional sites [2]. 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). (C) Fitness function inference. Dots represent data points (as in Fig. 4.6); also shown are the unconstrained fit to the Fermi-Dirac function of Eq. 1.4 ("UFD"; solid red line), constrained fit to the Eq. 1.4 with $f_0 = 0.99$ ("CFD"; dashed black line), and fit to an exponential fitness function ("EXP"; dashed green line). (D) Histogram of binding site energies and its prediction based on the three fits in (C) (Eq. 2.15).

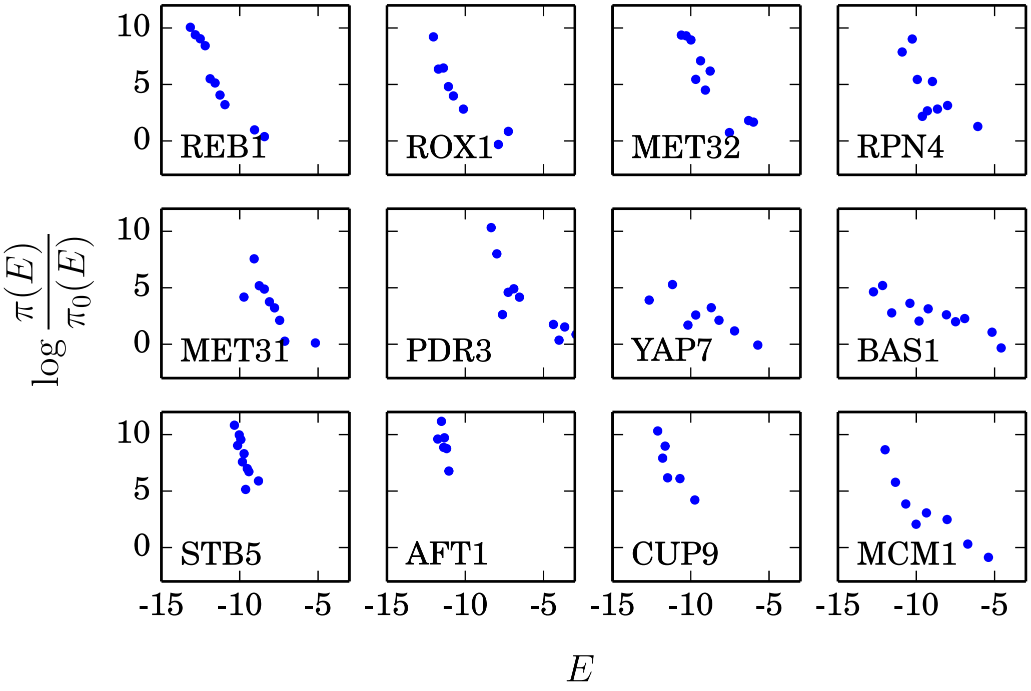


Figure 4.6: Qualitative behavior of fitness landscapes. Shown are plots of $\log(\pi(E)/\pi_0(E))$ for 12 TFs, which, according to Eq. 4.5, equals the logarithm of fitness up to an overall scale and shift. For each TF, sequences are grouped into 15 equal-size energy bins between the minimum and maximum energies allowed by the EM.

$$\pi_0(\sigma) = \pi_0(\sigma_1) \prod_{i=2}^{L} \pi_0(\sigma_{i-1}, \sigma_i), \tag{4.14}$$

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. 4.2 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 C.1, 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 [142].

4.5 Fermi-Dirac landscapes and model selection

For each TF we perform a maximum-likelihood fit of the binding site data to the distribution in Eq. 4.3 with the Fermi-Dirac landscape of Eq. 1.4 (Fig. 4.5, Table S2; see Methods for details). The model of Eq. 1.4 has four fitting parameters: β, μ, ν, and f_0. However, as shown in Sec. 4.1, 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. Although the inverse temperature β and the chemical potential μ have unambiguous physical meanings in the binding probability of Eq. 4.1, we will interpret the fits more broadly to define a class of fitness landscapes with "effective" β and μ, which may not be equal to their physical counterparts. The input to each fit is a collection of genomic TF binding sites $\{\sigma\}$ [2] and the energy matrix from high-throughput *in vitro* TF-DNA binding assays [1], which allows us to assign a binding energy $E(\sigma)$ to each site.

4.5.1 Methods: 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. 1.4). The log-likelihood is given by

$$\log \mathcal{L}(S|\theta) = \sum_{\sigma \in S} \log \pi(\sigma|\theta) = \sum_{\sigma \in S} \log \left(\frac{1}{Z(\theta)} \pi_0(\sigma) (\mathcal{F}(\sigma|\theta))^{\nu} \right), \tag{4.15}$$

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, we carry out its maximization in two steps. We first obtain a global map of the likelihood by calculating the function over a mesh of points in parameter space, over the domain $\beta \in (0.1, 10)$, $\mu \in (-20, 0)$, $\nu \in (10^{-3}, 10^5)$, and $f_0 \in (4.5 \times 10^{-5}, 1 - 4.5 \times 10^{-5})$. We then maximize the likelihood using conjugate-gradient ascent which starts from the approximate global maximum on the mesh.

4.5.2 Results

TF	f_0	$\gamma = \nu(1 - f_0)$	β (kcal/mol)$^{-1}$	$E - \mu$
REB1	0.999	18.3	0.801	≈ 0
ROX1	0.992	403	0.426	> 0
MET32	0.974	132	0.248	> 0
RPN4	4.77×10^{-9}	0.72	3.84	≈ 0
MET31	1.85×10^{-10}	0.547	4.63	≈ 0
PDR3	0.789	4.53×10^3	0.534	> 0
YAP7	6.01×10^{-6}	1.26	1.13	≈ 0
BAS1	2.09×10^{-3}	144	0.246	< 0
STB5	0.167	168	0.301	< 0
AFT1	3.11×10^{-13}	0.617	16.4	≈ 0
CUP9	0.976	243	0.338	> 0
MCM1	0.998	83.8	0.25	> 0

Table 4.1: 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 Table S2 for details).

A summary of maximum-likelihood parameter values for all TFs is shown in Tables 4.1 and C.2, column D. The variation of log-likelihood with fitting parameters is shown in Table

C.2, columns G and H. 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 (Table S2, 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 $s \ll 1$, $N_e s \gg 1$ regime in population genetics).

Five TFs (RPN4, MET31, YAP7, BAS1, and AFT1) have very small values of f_0 (Table 4.1), indicating that on average, removing their binding sites is strongly deleterious to the cell. In these cases, the degeneracy is broken and the global maximum occurs in the vicinity of $f_0 = 0$ (Table S2, column H, insets). Since $1 - f_0 \approx 1$, $\nu \approx \gamma$, a small value in four out of five cases (Table 4.1). 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 $f_0 = 0.167$, which means they are under strong selection but are not necessarily essential.

Since the constrained Fermi-Dirac fits have one less adjustable parameter, it is more consistent to do model selection on the basis of the Akaike information criterion (adjusted for finite-size samples) [143] rather than log-likelihoods:

$$\mathrm{AIC} = 2(k - \log \mathcal{L}) + \frac{2k(k+1)}{n-k-1}, \tag{4.16}$$

where k 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.

Table 4.2 shows the AIC differences between the unconstrained Fermi-Dirac fits (UFD, $k = 4$) and the constrained Fermi-Dirac fits with $f_0 = 0.99$ (CFD, $k = 3$) for each TF. Positive AIC differences indicate that UFD is more favorable. We also calculate the Akaike weights $w \propto e^{-\mathrm{AIC}/2}$, which give the relative likelihood that a given model is the best [143].

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. Indeed, the

TF	$\text{AIC}_{\text{CFD}} - \text{AIC}_{\text{UFD}}$	$\text{AIC}_{\text{EXP}} - \text{AIC}_{\text{UFD}}$	w_{UFD}	w_{CFD}	w_{EXP}
REB1	-2.022	35.832	0.267	0.733	4.42×10^{-9}
ROX1	-2.159	35.051	0.254	0.746	6.21×10^{-9}
MET32	-2.246	10.550	0.245	0.754	0.001
RPN4	17.672	33.683	1.000	1.45×10^{-4}	4.85×10^{-8}
MET31	2.807	11.778	0.801	0.197	0.002
PDR3	-1.750	79.244	0.294	0.706	1.82×10^{-18}
YAP7	-1.988	10.783	0.270	0.729	0.001
BAS1	-2.466	6.007	0.223	0.766	0.011
STB5	-2.737	-7.143	0.025	0.097	0.878
AFT1	-1.104	7.265	0.362	0.628	0.010
CUP9	-2.284	1.689	0.219	0.687	0.094
MCM1	-3.351	-0.167	0.135	0.719	0.146

Table 4.2: 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.

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 k. Out of the five TFs 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 (Table 4.2). Both RPN4 and MET31 are listed as nonessential in the Yeast Deletion Database [127, 139], suggesting an inconsistency in our analysis.

The fits to the Fermi-Dirac fitness landscapes also provide estimates of the effective inverse temperature β and the effective chemical potential μ (Table 4.1). The inferred values of β can be compared to the physical value at room temperature, $\beta_{ph} = 1.69$ (kcal/mol)$^{-1}$. Nine of the TFs (REB1, ROX1, MET32, PDR3, YAP7, BAS1, STB5, CUP9, MCM1) have β's lower than the physical value, while in the other three (RPN4, MET31, AFT1) $\beta > \beta_{ph}$. 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 (Table S2, column G).

The inferred value of μ relative to the distribution of energies E of the binding sites tells us which qualitative regime of the Fermi-Dirac fitness landscape the sites lie in. For five TFs (ROX1, MET32, PDR3, CUP9, MCM1), $E - \mu > 0$, and the sites reside on the exponential tail. Interestingly, $1 - f_0 \ll 1$ for all of these TFs. For another group of five TFs (REB1, RPN4, MET31, YAP7, AFT1), $E - \mu \approx 0$, so that the sites lie on the sharp threshold between fully bound and fully unbound states. In this regime, changing the energy of the site through mutations may lead to a large change in its occupancy by the TF. Finally, for two TFs (BAS1, STB5), $E - \mu < 0$, and the sites lie on the high-fitness plateau. Note that in most of the $E - \mu > 0$ and $E - \mu < 0$ cases, values of μ within a large range fit the data equally well, as long as the binding energies of all sites are well to the right or to the left of the chemical potential (Table S2, column G).

What does $\beta \neq \beta_{ph}$ 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. 4.9). The magnitude of the selection coefficient depends qualitatively on both $E - \mu$ and whether f_0 is zero or nonzero (Fig. 4.1). For five of the TFs (ROX1, MET32, PDR3, CUP9, MCM1), $f_0 \neq 0$, $\beta < \beta_{ph}$, and $E - \mu > 0$. Thus these TFs are in a regime where decreasing β strengthens selection (Fig. 4.1F). In other words, selection is stronger for these binding sites than expected from purely biophysical considerations. For RPN4, MET31, and AFT1, $f_0 \approx 0$, $\beta > \beta_{ph}$, and $E \approx \mu$. Hence $\partial \tilde{s}/\partial \beta > 0$, and selection is again stronger than expected. BAS1 and STB5 exhibit $\beta < \beta_{ph}$ and lie on the high fitness plateau ($E - \mu < 0$), and thus selection is also stronger than expected. In contrast, YAP7 and REB1 exhibit $\beta < \beta_{ph}$ and lie on the threshold $E - \mu \approx 0$, and hence selection is weaker than expected in these two cases.

4.6 Exponential fitness landscape

Next, we 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. 4.6 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.

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

$$\begin{aligned} \pi(\sigma) &= \frac{1}{Z}\pi_0(\sigma)e^{\nu\alpha E(\sigma)} \\ &= \prod_{i=1}^{L}\frac{\pi_0^i(\sigma_i)}{Z_i}e^{\nu\alpha\epsilon_i^{\sigma_i}}, \end{aligned} \tag{4.17}$$

where $E(\sigma)$ is given by Eq. 4.2, $\pi_0(\sigma)$ 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, sites decouple and the distribution of sites $\pi(\sigma)$ completely factorizes. The assumption of factorization underlies the common practice of inferring EMs from log-odds scores of observed genomic binding sites [8]. The log-odds score of a nucleotide σ_i is defined as

$$\mathcal{S}(\sigma_i) = \log\frac{p_i^{\sigma_i}}{\pi_0^i(\sigma_i)} = -\beta\epsilon_i^{\sigma_i} - \log Z_i, \tag{4.18}$$

where $p_i^{\sigma_i}$ is the probability of seeing 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. Eq. 4.18 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. 4.16. The AIC analysis shows that the exponential landscape is significantly poorer than the Fermi-Dirac landscape in all cases except STB5 (Table 4.2). 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).

4.7 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 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 EM model of TF-DNA energetics in which the energy of each position in the site is independent of all the other positions. The EM parameters are inferred from a high-throughput data set in which TF-DNA interactions were studied *in vitro* using a microfluidics device [1]. We consider two types of fitness functions: Fermi-Dirac, which appears naturally from considering TF binding as a two-state process (Eq. 4.1), and exponential, which is motivated by the observation that for many TFs, fitness appears to fall off linearly with energy in log-space.

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. 1.4) 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. 2.15:

$$\pi(E) = \frac{1}{Z}\pi_0(E)\langle \mathcal{F}(E;\mu_i,\beta_i,f_{0i})^{\nu}\rangle_i \equiv \frac{1}{Z}\pi_0(E)\mathcal{F}(E;\bar{\mu},\bar{\beta},\bar{f}_0)^{\bar{\nu}}, \tag{4.19}$$

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. The results of these tests indicate 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 [144–150], 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 is significantly (though weakly) correlated with conservation of those genes [151]. It has therefore been suggested that selection pressures are so strong that only the most nonessential genes experience significant genetic drift [144]. Previous studies have also found that gene expression levels are a more reliable (though still very weak) predictor of selection pressures than essentiality [148], 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.

These results are highly consistent with results showing that for the majority (67%) of binding sites, site occupancy is uncorrelated with gene expression [152], and results showing only a 3% overlap between binding sites determined by ChIP-chip analysis and regulatory interactions determined through mRNA expression microarrays [153]. It has been suggested that this is evidence for the decoupling of TF binding and gene expression in eukaryotes [48].

Available data does not rule out the possibility of time-dependent selection in combination with forms of site-dependent selection we have not accounted for. 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 [154–160]. There is also evidence that TF binding site motifs themselves change over time, through cross species comparison of motifs for certain TFs [161].

4.7.1 Turnover

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 [162–164]. Our evolutionary model can account for this scenario using a promoter-level fitness function. Here we show a toy model demonstrating this.

Consider a promoter with two possible binding sites, as illustrated in figure 4.7. Each binding site may bind to a TF according to the familiar Fermi-Dirac function, but we additionally assume that transcriptional machinery will bind to the promoter as a whole as well, such that the binding energy is proportional to the total number of bound TFs. That is, the fitness of a promoter with two sites with binding energies E_1 and E_2 is

$$f(E_1, E_2) = F(\,\epsilon[F(E_1, \mu) + F(E_2, \mu)]\,,\,\mu')$$

where $F(E, mu) = 1/(1 + e^{-\beta(E-\mu)})$ is the Fermi -Dirac binding probability, μ is the chemical potential of the TF, μ' is the chemical potential of the transciptional machinery, and ϵ is the binding energy of the transcriptional machinery to a single TF.

As illustrated in figure 4.7, only some of the binding sites will be functional. For low population sizes, selection pressure is weak and only one of the two binding sites is typically bound, while both are bound for larger populations.

Thus, although individual binding sites may appear or disappear, the overall binding affinity of the promoter for the transcriptional machinery may stay constant.

4.7.2 Conclusion

Out of 12 TFs with sufficient binding site data, five have $f_0 \approx 0$, indicating a large fitness penalty for deleting such sites. This conclusion is strongly supported by the AIC differences between unconstrained and non-lethal Fermi-Dirac fits for only one TF, RPN4 (Table 4.2).

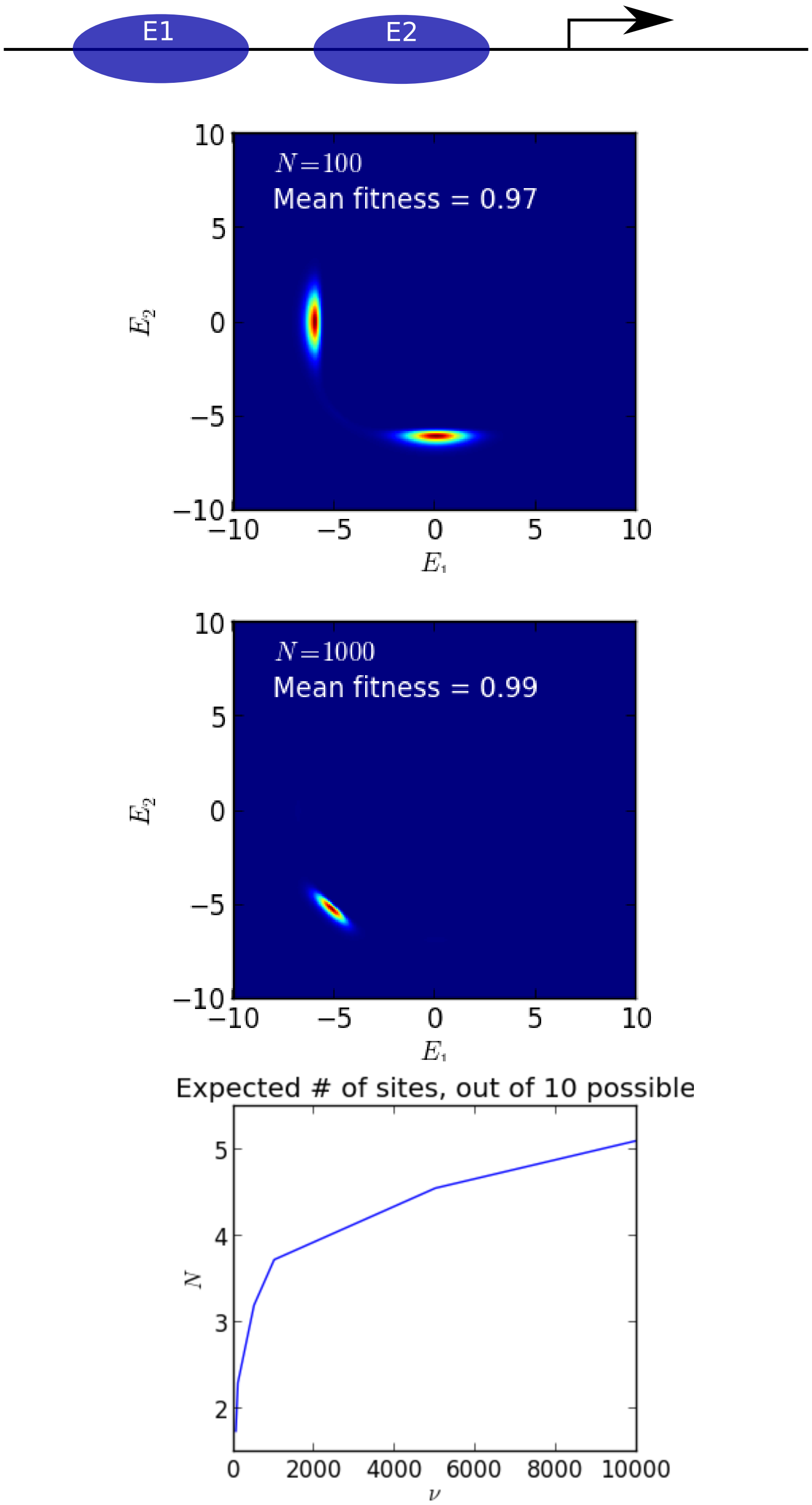


Figure 4.7: (Top) Illustration of a promoter with two binding sites, with binding energies E_1 and E_2. (Middle) For the fitness function defined in equation 4.7.1 , the distribution of phenotypes (E_1, E_2) for $N = 100$ and $N = 1000$, using the monomorphic steady state equation. For $N = 100$ typically only one site is bound at a time. For $N = 1000$ both sites are bound. (Bottom) For promoter with 10 binding sites, the number of bound sites as a function of population size.

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 EM 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 4.2).

We find that in 10 out of 12 cases, fitting an exponential fitness function is less supported by the data than fitting a Fermi-Dirac function (Table 4.2). 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 indicate epistasis among positions and show the limitations of this approximation.

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 4.1). 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 but not dynamic. We also find that the fitted inverse temperature β is typically not close to the value based on room temperature (Table 4.1). This observation suggests selection pressures in addition to those dictated by the energetics of TF binding to its cognate sites.

Appendix A

Diffusion Theory

The Probability of Fixation of a single Mutant

Using the backwards Kolmogorov equation we may obtain the probability of fixation of a single mutant. In the continuum limit, we may do this by setting the final population fraction x to 1 in equation 1.15, and we define the fixation probability after t steps as $u(x_0, t) = \theta(1, x_0, t)$, and ultimate fixation probability $u(x_0) = \lim_{t\to\infty} \theta(1, x_0, t)$. At large times the mutant must either fix or become extinct, and the fixation probability becomes constant in time, allowing us to set $\frac{\partial u(x_0)}{\partial t} = 0$. Substituting this into the Kolmogorov equation, we find the fixation probability must satisfy

$$M(x_0)\frac{\partial u(x_0)}{\partial x_0} + \frac{V(x_0)}{2}\frac{\partial^2 u(x_0)}{\partial^2 x_0} = 0 \tag{A.1}$$

ith boundary conditions $u(0) = 0$ and $u(1) = 1$. The solution is found by elementary methods, rearranging as

$$\frac{d}{dx_0}\left(\log \frac{du(x_0)}{dx_0}\right) = -\frac{2M(x_0)}{V(x_0)}, \tag{A.2}$$

and one then easily obtains

$$u(x_0) = \frac{\int_0^{x_0} G(x)dx}{\int_0^1 G(x)dx} \tag{A.3}$$

using an integrating factor

$$G(x) = e^{-\int \frac{2M(x)}{V(x)}dx}. \tag{A.4}$$

To get the fixation of a probability of a single mutant, one evaluates the result with $x_0 = 1/N$. Substituting the moments for the Wright-Fisher model, one obtains

$$u(p) = \frac{1 - e^{-sNp}}{1 - e^{-sN}} \tag{A.5}$$

and setting $p = 1/N$, for one initial mutant, gives the result for $\theta(s) = u(1/N)$ quoted at the start of this section.

The Time to Fixation of a single Mutant

Although we shall not derive it in full, similar methods give the number of generations until fixation or extinction.

The fixation time is most conveniently derived by writing

$$\bar{t}_{\text{fix}} = \frac{T(x_0)}{u(x_0)} \quad \text{with} \quad T(x_0) = \int_0^\infty t \frac{\partial u(x_0, t)}{\partial t} dt \tag{A.6}$$

where T can be though of as a sum of fixation times weighted by the probability $\frac{\partial u(x_0,t)}{\partial t}$ that the mutant fixes in a short interval dt around time t, and therefore $\bar{t}_{\text{fix}}$ is the normalized sum, giving the average fixation time for the cases where the mutant fixes.

By differentiating then integrating the backwards Kolmogorov equation with respect to time, one can obtain a differential equation for $T(x_0)$, which, with appropriate boundary conditions, leads to the solution

$$\bar{t}_{\text{fix}} = \int_p^1 \psi(x')u(x')(1-u(x'))dx' + \frac{1-u(p)}{u(p)} \int_0^p \psi(x')u^2(x')dx' \tag{A.7}$$

where

$$\psi(x) = \frac{2\int_0^1 G(x')dx'}{G(x)V(x)} \tag{A.8}$$

Through a similar procedure one also obtains the extinction time

$$\bar{t}_{\text{ext}} = \frac{1-u(p)}{u(p)} \int_p^1 \psi(x')(1-u(x'))^2 dx' + \int_0^p \psi(x')u(x')(1-u(x'))dx' \tag{A.9}$$

The Wright-Fisher moments $M(x)$ and $V(x)$ may now be substituted to obtain the WF diffusion limit. The result is too complicated to state here, but it is used in Chapter 2. In the $s = 0$ (neutral) case, however, this result simplifies to

$$\bar{t}_{\text{fix}} = 2(N-1)N \log \frac{N}{N-1} \qquad \approx 2N \tag{A.10}$$

$$\bar{t}_{\text{ext}} = \frac{2N \log N}{N-1} \qquad \approx 2 \log N \tag{A.11}$$

The form of the s-dependent fixation and extinction times are shown in figure 1.4.

Appendix B

Monomorphic-Polymorphic Boundary with Selection

Here we describe the dependence of the fixation/extinction time on the fitness of a single mutant in a monomorphic population, where the mutant's fitness is f_m and the population's is f_p. We shall do so using Kimura's theory, which is valid in the 'diffusion limit', that is, when the change in frequency of an allele per generation is small. This approximation may break down for small populations and strong selection.

It will be convenient to relate their fitnesses through the selection coefficient $s = f_m/f_p - 1$, which ranges from -1 to ∞. The fixation/extinction time is calculated as $\langle t(s,N)\rangle = P_{\text{fix}}t_{\text{fix}} + P_{\text{ext}}t_{\text{ext}}$ using Kimura's theory for haploid populations. The result is shown in Fig S1 for N=1000. As can be seen, it deviates from the actual Wright-Fisher result partly due to the small population and for large s, but is useful as a rough approximation. The full result is

$$\begin{aligned}
\langle t(s,N)\rangle = \frac{e^{-2s}}{(e^{2ns}-1)s}&((e^{2s}+e^{2s(n+1)}-2e^{2ns})\gamma \\
&+(e^{2ns}-1)\,\mathrm{Ei}(2s) \\
&+(e^{2ns}-e^{4ns})\,\mathrm{Ei}(-2s(n-1)) \\
&+(e^{4ns}-e^{2s(n+1)})\,\mathrm{Ei}(-2ns) \\
&+(1-e^{2s})\,\mathrm{Ei}(2ns) \\
&+e^{2s(n+1)}\log(2|s|n(n-1)) \\
&+e^{2s}\log(2|s|n/(n-1)) \\
&-2e^{2ns}\log(2|s|n))
\end{aligned} \tag{B.1}$$

.

Given this function, and given a distribution of mutant selection coefficients $g(s)$, it is then possible to calculate the total expected fixation/extinction time $\int g(s)\langle t(s)\rangle ds$, which

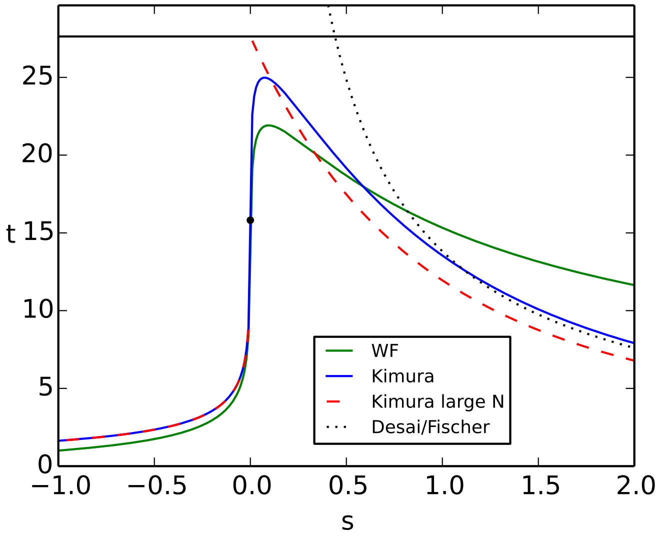


Figure B.1: Expected fixation/extinction time for N=1000. In the neutral limit, this 15.8 generations (black dot). The kimura large-N maximum is the black line. Different computations shown in legend.

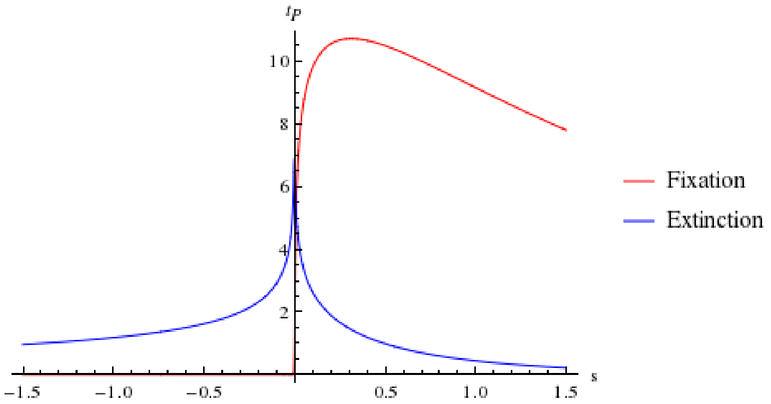


Figure B.2: Contributions to the total fixation time from mutants that go extinct, and these that fix. (ie, $P_{\text{fix}}(s)t_{\text{fix}}(s)$) for N=500

can then be compared to the waiting time in a substitution process to obtain a bound between the polymorphic and monomorphic regimes. For example, for $N = 1000$ and assuming $g(s)$ is a Gaussian with mean 0 and standard deviation of 0.2, the average fixation/extinction time is 14.8 generations.

However, even if we are ignorant of the form of $g(s)$ we can still set some bounds on the expected fixation/extinction time in the diffusion limit. We shall examine the large-N limit of the fixation/extinction time, which is more tractable, and which gives fixation/extinction times of

$$\begin{array}{ll} \frac{2-2e^{-2s}}{s}\log N & \text{for } s > 0 \\ 2\log N & \text{for s} = 0 \\ \frac{1}{s}(e^{-2s}\,\mathrm{Ei}(2s) - \log(-2s) - \gamma) & \text{for } s < 0 \end{array} \tag{B.2}$$

where Ei is the exponential integral function. The large-s limit of the $s > 0$ case reproduces the strong selection result from Desai & Fisher of $\frac{1}{s}\log Ns$. Note that this latter result is only a good approximation to Kimura's diffusion result for large s, however for such large selection the diffusion result deviates significantly from the true value.

In this large N limit, the maximum fixation/extinction time is $4\log N$, which occurs for the small s limit of the $s > 0$ case. This $4\log N$ result provides an upper bound for the fixation time for all non-neutral evolution, for any N.

The lower bound on the fixation time is more complicated. The fixation time is less than the neutral fixation time in two cases: For deleterious mutants, which typically quickly become extinct, and for highly fit mutants which fix very quickly.

In the latter case for very high s, the fixation time becomes small due to the large selective pressure driving the mutants to fixation. It becomes less than the neutral fixation time when $(2 - 2e^{-2s})/s < 1$, which gives $s > 0.8$, or $f_m > 1.8f_p$. Thus, we expect that all beneficial mutants will fix or become extinct between $2\log N$ and $4\log N$ generations, as long as their fitness is no more than 1.8 times the population's fitness. We shall assume such extremely fit mutants cannot occur. This result depends on the diffusion limit which we know breaks down for such strong selection, however judging from the exact WF calculation, it is an underestimate, and ignoring such extremely fit mutants is justified.

Deleterious mutants with $s < 0$, on the other hand, generally go extinct quickly, faster than the neutral fixation time of $2 \log N$, and in the limit of large N they nearly immediately become extinct compared to the neutral fixation time. In the context of the monomorphic-polymorphic regime boundary, this means their effect on dynamics is often negligible and the population will be effectively monomorphic despite their presence. However, if there are a large number of them they may still contribute to the average fixation time.

If the number of deleterious mutants is not too large compared to the number of beneficial mutations ones, they can be ignored completely. This may be the case in evolution out-of-equilibrium. We can account for this scenario by rescaling the mutation rate to $u' = up_b$ where $p_b = \int_0^\infty g(s)ds$ is the fraction of a beneficial mutations. Thus, in this case in the large N limit the monomorphic/polymorphic regime boundary is between the upper bound of $4p_b N \log N = 1$ and the lower bound of $2p_b uN \log N = 1$. This will be a good approximation as long as $\int_{-\infty}^0 g(s)\langle t(s)\rangle ds$ is small relative to $\int_0^\infty g(s)\langle t(s)\rangle ds$, keeping in mind that $t(s) \sim \log N$ for s¿0 and ~ 1 for s¡0.

However, in monomorphic steady state the number of beneficial substitutions must be equal to the number of deleterious substitutions, and since the deleterious mutants are unlikely to fix there must be many more of them. We can make a rough argument that the contribution of deleterious mutants will dominate the fixation/extinction time, based on the known fact that the ratio $P_{\text{fix}}(r)/P_{\text{fix}}(1/r) \approx r^N$, where $r = f_m/f_p$. That is, beneficial mutants are more likely than deleterious ones to fix by a typical factor of r^N, and therefore there must be r^N times as many deleterious mutants in steady state as beneficial ones. This assumes the strength of deleterious mutants is similar to the strength of beneficial ones (as would happen for an EM model). If expected fixation/extinction time for the $s < 0$ case is weighted by this factor of r^N compared to the $s > 0$ case (which only grows as $\log N$) it will dominate the average. Thus we expect the average fixation time to be slightly less than $2 \log N$, since the $s < 0$ fixation/extinction time is always less than the neutral $2 \log N$, and in the steady state we also expect most substitutions to be near neutral (except for unusual configurations of genotype-space).

We demonstrate this for a particular choice of $g(s)$ appropriate for steady state: The condition on $g(s)$ in steady state is that $\int sg(s)P_{\text{fix}}(s)ds = 0$, that the average s is 0. This

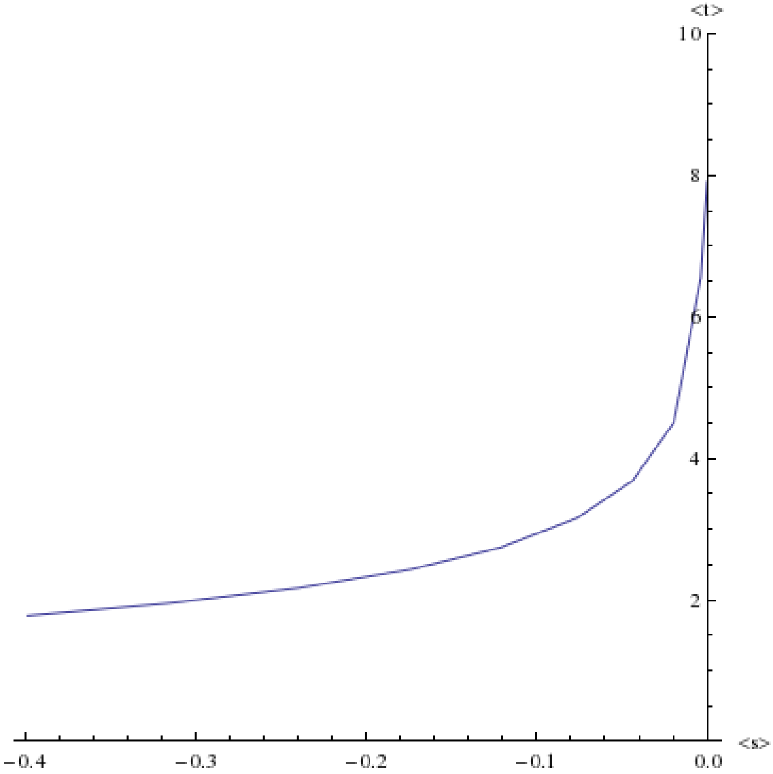


Figure B.3: Plot of the mean s vs the fixation/extinction time, as we vary σ for the choice of $g(s)$ described in the text.

condition is obtained by assuming that the population is in a state with fitness f_0, requiring that the average $\Delta f = f_f - f_0$ due to a substitution to be 0, and rewriting in terms of s. One way to satisfy this constrain is with $g(s) \propto k(s)/P_{\text{fix}}(s)$ for any even function $k(s)$. Here we choose $k(s)$ to be a Gaussian parametrized by its standard deviation σ and we calculate the mean fixation/extinction time and the mean s for varying σ. The result is show in fig xx. As expected, the average fixation/extinction time is less than the neutral time, and follows the $s < 0$ fixation/extinction time closely.

Appendix C

Yeast Selection Tests and Fits

Table C.1: Full summary of tests for correlations between TF-DNA binding energies and growth rates after knockouts of genes regulated by the TF, expression levels of regulated genes, K_A/K_S ratios for regulated genes, and distances between TF sites and the TSS of the regulated gene.

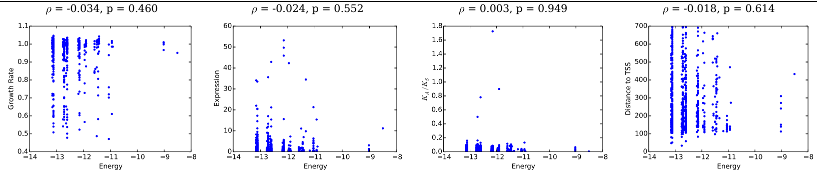
Table C.2: Full summary of parametric fits of fitness landscapes to TF binding site data.

(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

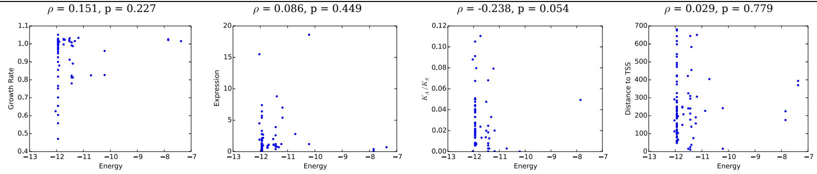
	Essential	**Noness.**
Total Data	211	481
Expr Data	199	419
K_A/K_S Data	192	371
$\bar{E}$	-12.538	-12.634
V	0.338	0.466
$\bar{d}$	0.911	0.845


ρ = -0.034, p = 0.460
ρ = -0.024, p = 0.552
ρ = 0.003, p = 0.949
ρ = -0.018, p = 0.614

ROX1 (nonessential TF)

Total sites: 93
Unique sites: 58

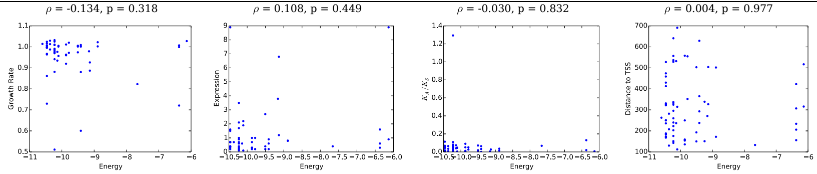
	Essential	**Noness.**
Total Data	17	68
Expr Data	16	63
K_A/K_S Data	17	49
$\bar{E}$	-11.566	-11.769
V	0.083	0.828
$\bar{d}$	0.712	0.875


ρ = 0.151, p = 0.227
ρ = 0.086, p = 0.449
ρ = -0.238, p = 0.054
ρ = 0.029, p = 0.779

MET32 (nonessential TF)

Total sites: 68
Unique sites: 39

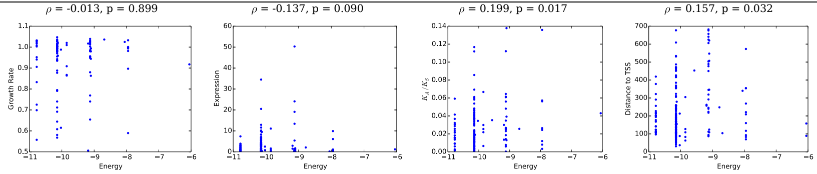
	Essential	**Noness.**
Total Data	5	59
Expr Data	4	47
K_A/K_S Data	4	49
$\bar{E}$	-9.738	-8.66
V	3.917	1.136
$\bar{d}$	0.377	0.25


ρ = -0.134, p = 0.318
ρ = 0.108, p = 0.449
ρ = -0.030, p = 0.832
ρ = 0.004, p = 0.977

RPN4 (nonessential TF)

Total sites: 188
Unique sites: 38

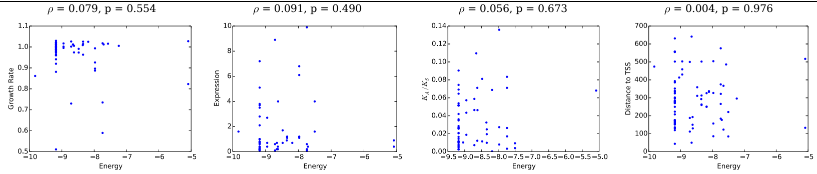
	Essential	**Noness.**
Total Data	70	103
Expr Data	67	87
K_A/K_S Data	65	79
$\bar{E}$	-9.773	-10.016
V	0.297	0.696
$\bar{d}$	0.226	0.169


ρ = -0.013, p = 0.899
ρ = -0.137, p = 0.090
ρ = 0.199, p = 0.017
ρ = 0.157, p = 0.032

MET31 (nonessential TF)

Total sites: 77
Unique sites: 35

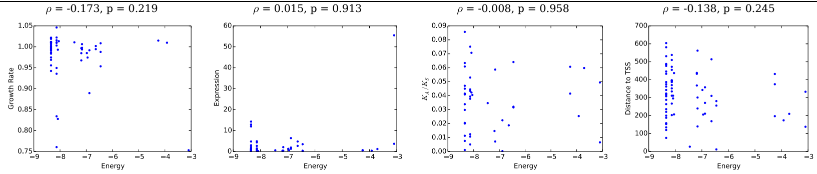
	Essential	**Noness.**
Total Data	10	60
Expr Data	9	51
K_A/K_S Data	10	49
$\bar{E}$	-8.65	-8.318
V	0.434	0.74
$\bar{d}$	0.151	0.0


ρ = 0.079, p = 0.554
ρ = 0.091, p = 0.490
ρ = 0.056, p = 0.673
ρ = 0.004, p = 0.976

PDR3 (nonessential TF)

Total sites: 73
Unique sites: 31

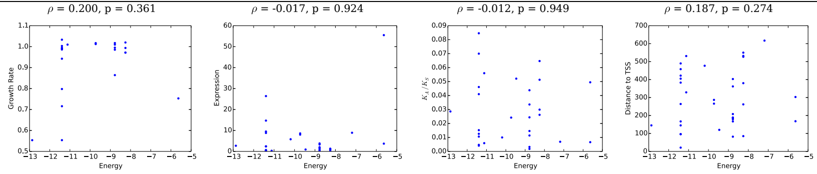
	Essential	**Noness.**
Total Data	5	53
Expr Data	5	48
K_A/K_S Data	5	36
$\bar{E}$	-7.524	-5.214
V	3.794	1.495
$\bar{d}$	0.298	1.2


ρ = -0.173, p = 0.219
ρ = 0.015, p = 0.913
ρ = -0.008, p = 0.958
ρ = -0.138, p = 0.245

YAP7 (nonessential TF)

Total sites: 36
Unique sites: 22

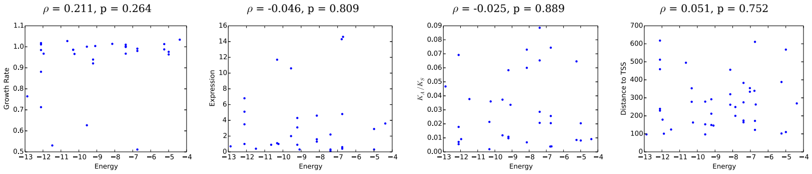
	Essential	**Noness.**
Total Data	13	23
Expr Data	13	22
K_A/K_S Data	11	18
$\bar{E}$	-9.936	-9.076
V	2.575	2.865
$\bar{d}$	0.182	0.583


ρ = 0.200, p = 0.361
ρ = -0.017, p = 0.924
ρ = -0.012, p = 0.949
ρ = 0.187, p = 0.274

BAS1 (nonessential TF)

Total sites: 41
Unique sites: 21

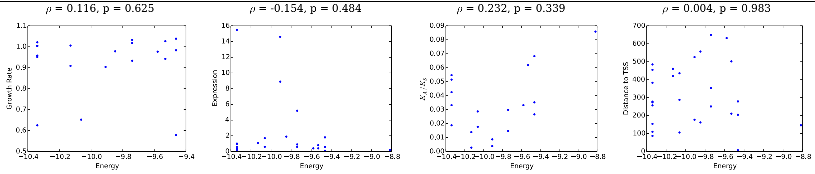
	Essential	Noness.
Total Data	5	30
Expr Data	5	25
K_A/K_S Data	5	28
$\bar{E}$	-8.926	-7.792
V	0.896	6.49
$\bar{d}$	0.214	0.2



STB5 (nonessential TF)

Total sites: 28
Unique sites: 19

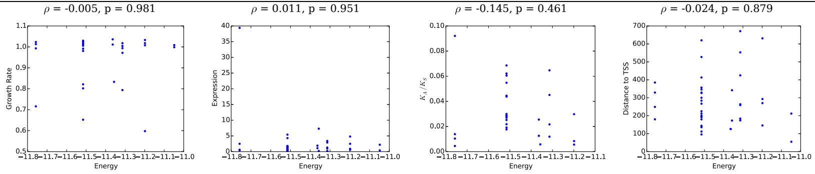
	Essential	Noness.
Total Data	5	20
Expr Data	5	18
K_A/K_S Data	5	14
$\bar{E}$	-9.918	-9.893
V	0.317	0.116
$\bar{d}$	0.222	0.4



AFT1 (nonessential TF)

Total sites: 42
Unique sites: 18

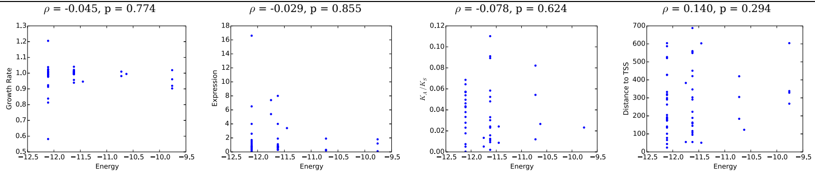
	Essential	Noness.
Total Data	5	31
Expr Data	4	30
K_A/K_S Data	5	23
$\bar{E}$	-11.423	-11.475
V	0.006	0.036
$\bar{d}$	0.417	0.4



CUP9 (nonessential TF)

Total sites: 58
Unique sites: 13

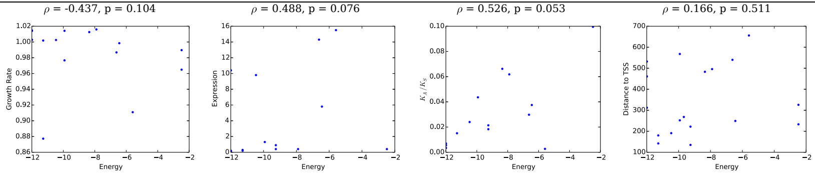
	Essential	Noness.
Total Data	11	43
Expr Data	11	32
K_A/K_S Data	11	31
$\bar{E}$	-11.607	-11.681
V	0.141	0.494
$\bar{d}$	0.139	0.1



MCM1 (essential TF)

Total sites: 18
Unique sites: 13

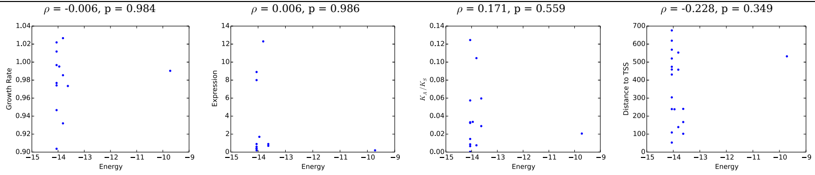
	Essential	Noness.
Total Data	2	15
Expr Data	2	12
K_A/K_S Data	2	12
$\bar{E}$	-8.58	-9.252
V	0.0	9.927
$\bar{d}$	0.8	2.0



CIN5 (nonessential TF)

Total sites: 19
Unique sites: 12

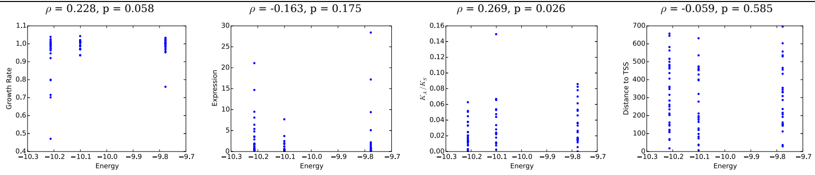
	Essential	Noness.
Total Data	2	15
Expr Data	2	10
K_A/K_S Data	2	12
$\bar{E}$	-13.683	-13.841
V	0.046	1.139
$\bar{d}$	0.5	1.0



GAT1 (nonessential TF)

Total sites: 88
Unique sites: 11

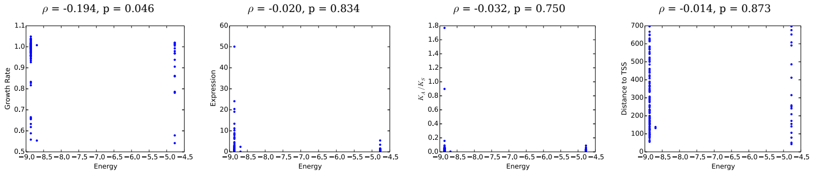
	Essential	Noness.
Total Data	8	71
Expr Data	8	63
K_A/K_S Data	7	61
$\bar{E}$	-10.048	-10.036
V	0.041	0.035
$\bar{d}$	0.362	0.429



MSN2 (nonessential TF)

Total sites: 141
Unique sites: 8

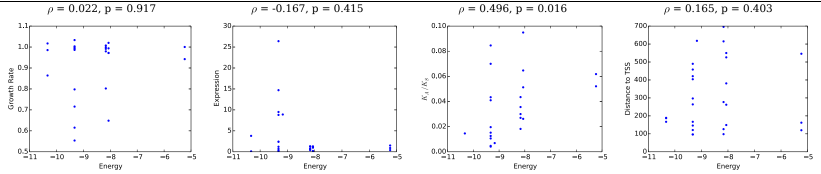
	Essential	Noness.
Total Data	19	108
Expr Data	17	97
K_A/K_S Data	15	87
$\bar{E}$	-8.216	-8.433
V	1.577	2.218
$\bar{d}$	0.097	0.059



CAD1 (nonessential TF)

Total sites: 28
Unique sites: 8

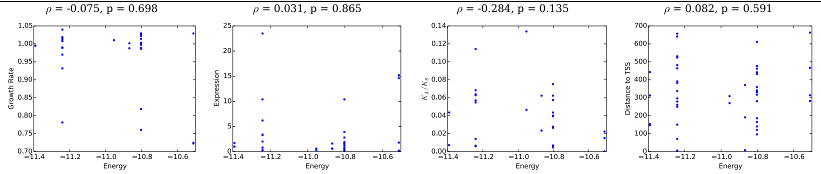
	Essential	Noness.
Total Data	3	25
Expr Data	3	23
K_A/K_S Data	3	20
$\bar{E}$	-8.635	-7.91
V	3.585	1.571
$\bar{d}$	0.2	0.0



ACE2 (nonessential TF)

Total sites: 45
Unique sites: 6

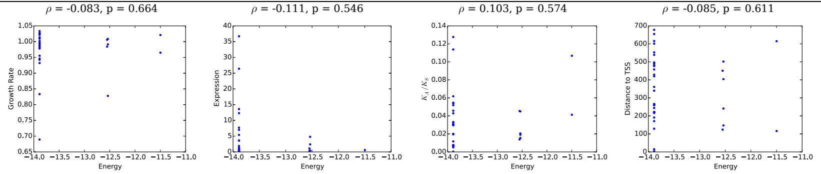
	Essential	Noness.
Total Data	7	29
Expr Data	7	26
K_A/K_S Data	6	23
$\bar{E}$	-10.954	-11.023
V	0.094	0.065
$\bar{d}$	0.0	0.0



YAP3 (nonessential TF)

Total sites: 38
Unique sites: 6

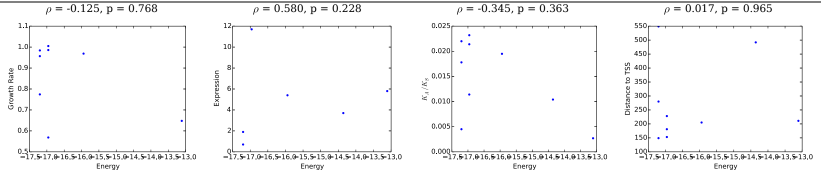
	Essential	Noness.
Total Data	8	30
Expr Data	8	24
K_A/K_S Data	8	24
$\bar{E}$	-13.503	-13.718
V	0.199	0.535
$\bar{d}$	0.276	0.0



GCN4 (nonessential TF)

Total sites: 9
Unique sites: 5

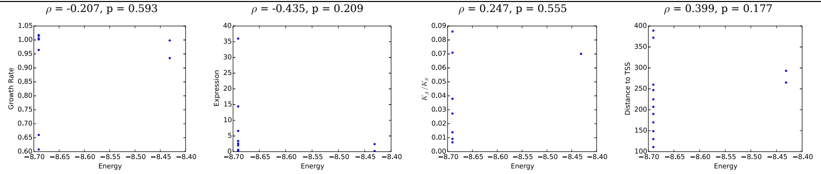
	Essential	Noness.
Total Data	1	8
Expr Data	1	5
K_A/K_S Data	1	8
$\bar{E}$	-16.442	-14.357
V	0.0	1.736
$\bar{d}$	0.429	2.0



MATA2 (nonessential TF)

Total sites: 13
Unique sites: 4

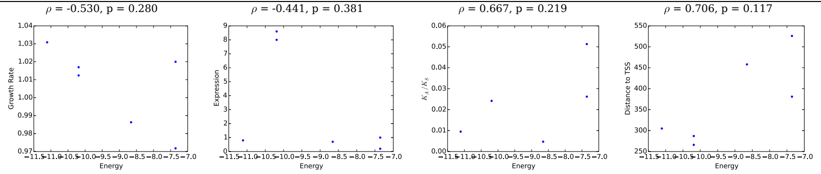
	Essential	Noness.
Total Data	1	10
Expr Data	1	9
K_A/K_S Data	1	7
$\bar{E}$	-8.639	-8.69
V	0.0	0.011
$\bar{d}$	0.444	1.0



YAP1 (nonessential TF)

Total sites: 6
Unique sites: 4

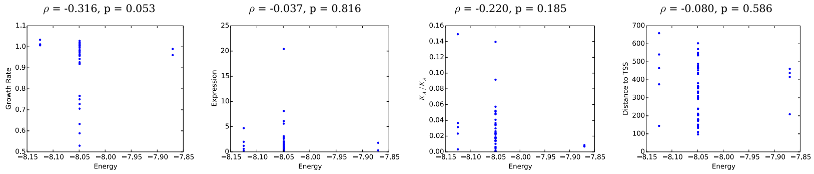
	Essential	Noness.
Total Data	0	6
Expr Data	0	6
K_A/K_S Data	0	5
$\bar{E}$	-9.143	nan
V	nan	2.114
$\bar{d}$	0.6	nan



CBF1 (nonessential TF)

Total sites: 49
Unique sites: 3

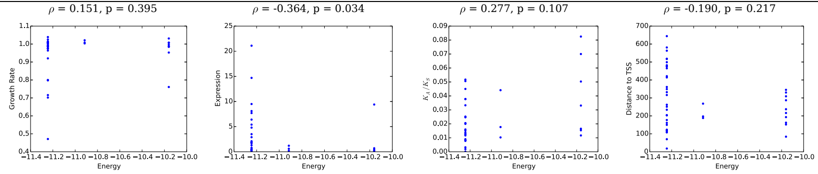
	Essential	**Noness.**
Total Data	7	40
Expr Data	7	36
K_A/K_S Data	6	32
$\bar{E}$	-8.048	-8.06
V	0.001	0.002
$\bar{d}$	0.132	0.143



DAL80 (nonessential TF)

Total sites: 44
Unique sites: 3

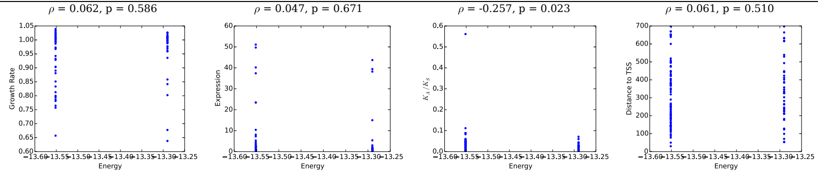
	Essential	**Noness.**
Total Data	4	35
Expr Data	4	30
K_A/K_S Data	4	31
$\bar{E}$	-10.969	-11.245
V	0.0	0.202
$\bar{d}$	0.294	0.0



AFT2 (nonessential TF)

Total sites: 118
Unique sites: 2

	Essential	**Noness.**
Total Data	17	82
Expr Data	15	70
K_A/K_S Data	15	63
$\bar{E}$	-13.45	-13.505
V	0.01	0.016
$\bar{d}$	0.099	0.0



SKO1 (essential TF)

Total sites: 12
Unique sites: 2

	Essential	**Noness.**
Total Data	1	11
Expr Data	1	10
K_A/K_S Data	1	9
$\bar{E}$	-7.525	-7.801
V	0.0	0.343
$\bar{d}$	0.0	0.0

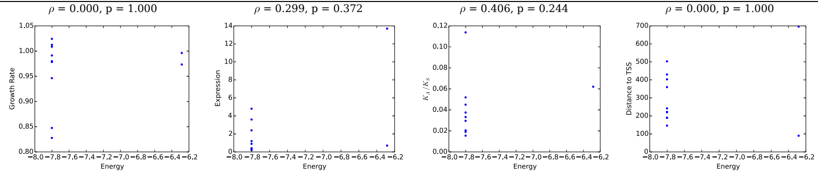
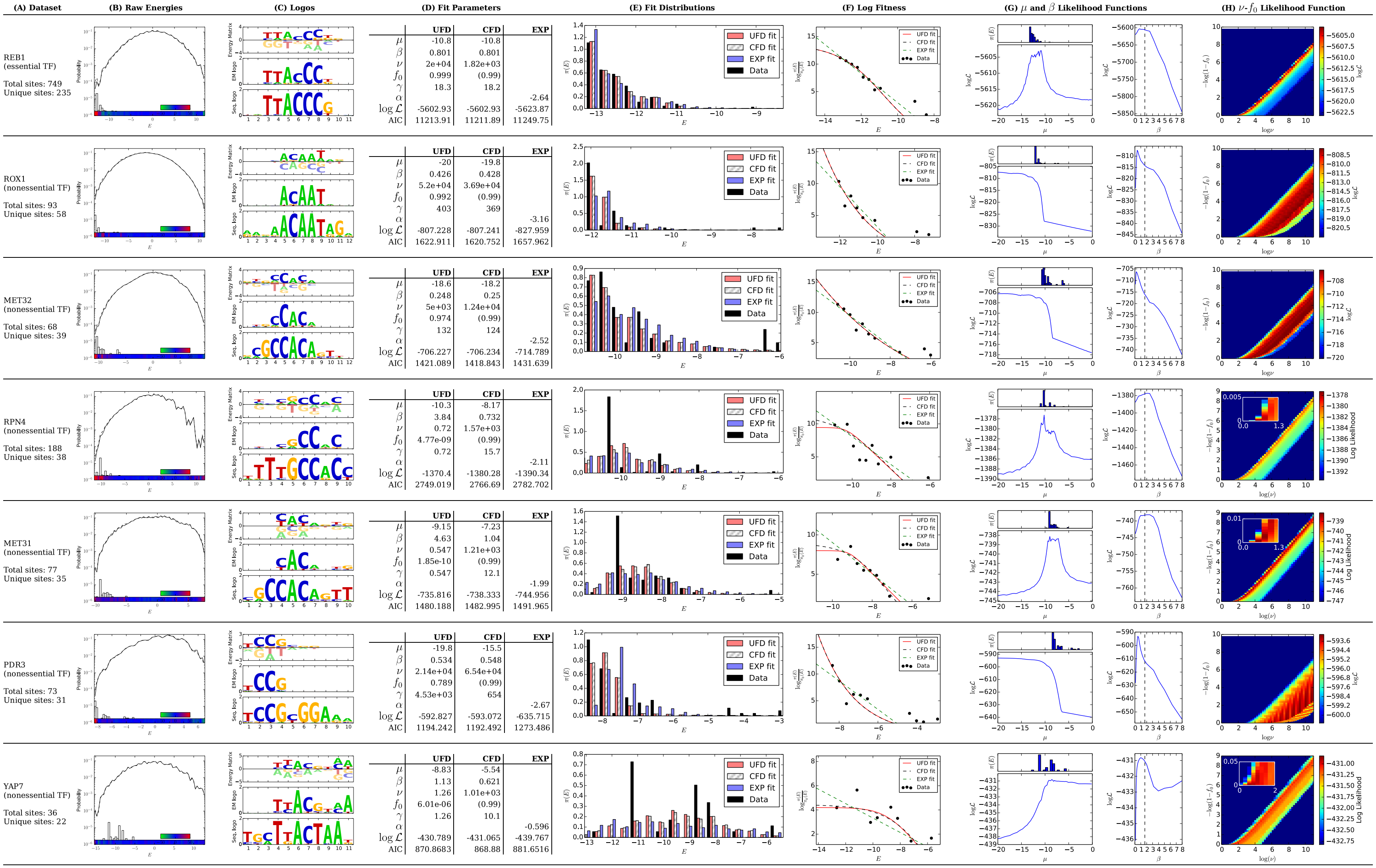


Table C.1: **Full summary of tests for site-specific selection.** For 25 TFs we compute TF-DNA interaction energies (in kcal/mol) for each site. Columns from left to right: (A) Essentiality of the TF according to the Yeast Deletion Database; total number of binding sites for each TF; total number of 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 $\bar{d}$ between *S. cerevisiae* and *S. paradoxus* sites regulating essential and nonessential genes. (B) Growth rate in strains with 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.



REB1
(essential TF)

Total sites: 749
Unique sites: 235

	UFD	CFD	EXP
μ	-10.8	-10.8	
β	0.801	0.801	
ν	2e+04	1.82e+03	
f_0	0.999	(0.99)	
γ	18.3	18.2	
α			-2.64
$\log\mathcal{L}$	-5602.93	-5602.93	-5623.87
AIC	11213.91	11211.89	11249.75

ROX1
(nonessential TF)

Total sites: 93
Unique sites: 58

	UFD	CFD	EXP
μ	-20	-19.8	
β	0.426	0.428	
ν	5.2e+04	3.69e+04	
f_0	0.992	(0.99)	
γ	403	369	
α			-3.16
$\log\mathcal{L}$	-807.228	-807.241	-827.959
AIC	1622.911	1620.752	1657.962

MET32
(nonessential TF)

Total sites: 68
Unique sites: 39

	UFD	CFD	EXP
μ	-18.6	-18.2	
β	0.248	0.25	
ν	5e+03	1.24e+04	
f_0	0.974	(0.99)	
γ	132	124	
α			-2.52
$\log\mathcal{L}$	-706.227	-706.234	-714.789
AIC	1421.089	1418.843	1431.639

RPN4
(nonessential TF)

Total sites: 188
Unique sites: 38

	UFD	CFD	EXP
μ	-10.3	-8.17	
β	3.84	0.732	
ν	0.72	1.57e+03	
f_0	4.77e-09	(0.99)	
γ	0.72	15.7	
α			-2.11
$\log\mathcal{L}$	-1370.4	-1380.28	-1390.34
AIC	2749.019	2766.69	2782.702

MET31
(nonessential TF)

Total sites: 77
Unique sites: 35

	UFD	CFD	EXP
μ	-9.15	-7.23	
β	4.63	1.04	
ν	0.547	1.21e+03	
f_0	1.85e-10	(0.99)	
γ	0.547	12.1	
α			-1.99
$\log\mathcal{L}$	-735.816	-738.333	-744.956
AIC	1480.188	1482.995	1491.965

PDR3
(nonessential TF)

Total sites: 73
Unique sites: 31

	UFD	CFD	EXP
μ	-19.8	-15.5	
β	0.534	0.548	
ν	2.14e+04	6.54e+04	
f_0	0.789	(0.99)	
γ	4.53e+03	654	
α			-2.67
$\log\mathcal{L}$	-592.827	-593.072	-635.715
AIC	1194.242	1192.492	1273.486

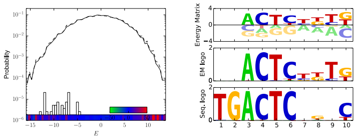
YAP7
(nonessential TF)

Total sites: 36
Unique sites: 22

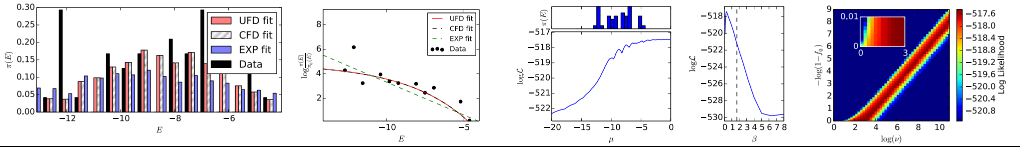
	UFD	CFD	EXP
μ	-8.83	-5.54	
β	1.13	0.621	
ν	1.26	1.01e+03	
f_0	6.01e-06	(0.99)	
γ	1.26	10.1	
α			-0.596
$\log\mathcal{L}$	-430.789	-431.065	-439.767
AIC	870.8683	868.88	881.6516

BAS1
(nonessential TF)

Total sites: 41
Unique sites: 21

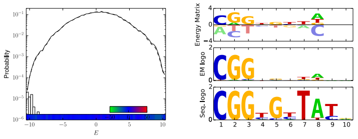


	UFD	CFD	EXP
μ	9.14	13.4	
β	0.246	0.245	
ν	144	4e+04	
f_0	0.00209	(0.99)	
γ	144	400	
α			-0.53
$\log\mathcal{L}$	-517.33	-517.328	-523.838
AIC	1043.771	1041.305	1049.779

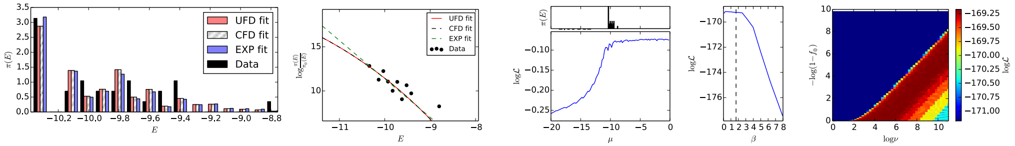


STB5
(nonessential TF)

Total sites: 28
Unique sites: 19

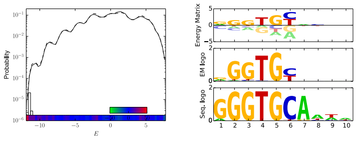


	UFD	CFD	EXP
μ	-1.92	-1.8	
β	0.301	0.318	
ν	202	1.98e+04	
f_0	0.167	(0.99)	
γ	168	198	
α			-4.36
$\log\mathcal{L}$	-169.172	-169.173	-169.393
AIC	348.0831	345.346	340.9398

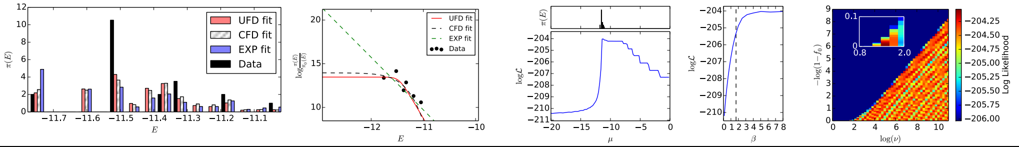


AFT1
(nonessential TF)

Total sites: 42
Unique sites: 18

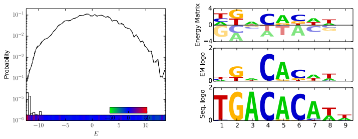


	UFD	CFD	EXP
μ	-11.5	-10.8	
β	16.4	4.13	
ν	0.617	1.98e+03	
f_0	3.11e-13	(0.99)	
γ	0.617	19.8	
α			-6.02
$\log\mathcal{L}$	-203.447	-204.12	-210.57
AIC	415.9751	414.8716	423.24

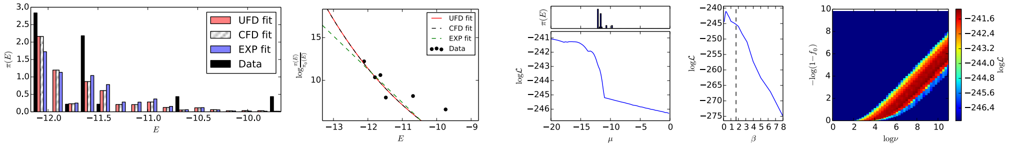


CUP9
(nonessential TF)

Total sites: 58
Unique sites: 13

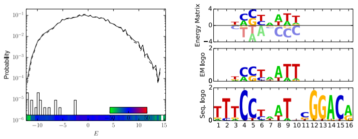


	UFD	CFD	EXP
μ	-20	-19.6	
β	0.338	0.341	
ν	1.02e+04	2.2e+04	
f_0	0.976	(0.99)	
γ	243	220	
α			-3.89
$\log\mathcal{L}$	-241.004	-241.017	-245.19
AIC	490.7627	488.4784	492.4514



MCM1
(essential TF)

Total sites: 18
Unique sites: 13



	UFD	CFD	EXP
μ	-20.1	-20	
β	0.25	0.252	
ν	5.05e+04	8.18e+03	
f_0	0.998	(0.99)	
γ	83.8	81.8	
α			-0.861
$\log\mathcal{L}$	-348.811	-348.817	-353.141
AIC	708.6989	705.3483	708.532

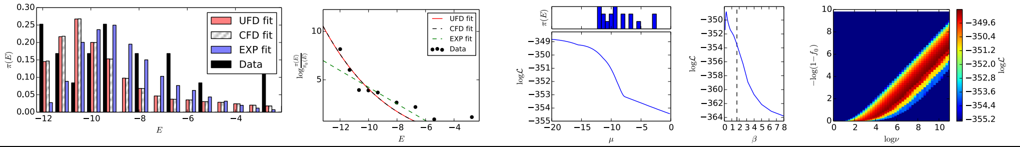


Table C.2: **Summary of fitness landscape fits to TF binding site data.** We consider 12 TFs which have more than 12 unique binding site sequences. Each row corresponds to a TF, ranked in the decreasing order of the number of unique binding site sequences. Columns, from left to right: (A) Summary of TF binding site data. (B) Same as Fig. 5B. (C) Same as Fig. 5A. (D) Fitted values of fitness landscape parameters and maximized log-likelihoods for the unconstrained fit to the Fermi-Dirac function of Eq. 7 ("UFD"), constrained fit to the Eq. 7 with $f_0 = 0.99$ ("CFD"), and fit to an exponential fitness function ("EXP"). (E) Same as Fig. 5D. (F) Same as Fig. 5C. (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 create the log-likelihood plots, either μ or β were held fixed while all the other parameters were re-optimized. (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 create the log-likelihood heatmaps, both ν and f_0 were held fixed while all the other parameters were re- optimized. Note that in F and H, the maximum values do not always match those listed in D because we employ an additional round of conjugate-gradient ascent after locating the approximate maximum on the grid.

Bibliography

[1] P. M. Fordyce, D. Gerber, D. Tran, J. Zheng, H. Li, J. L. DeRisi, and S. R. Quake, “De novo identification and biophysical characterization of transcription-factor binding sites with microfluidic affinity analysis,” *Nature Biotech*, vol. 28, pp. 970–975, 2010.

[2] K. Chen, E. van Nimwegen, N. Rajewsky, and M. L. Siegal, “Correlating gene expression variation with cis-regulatory polymorphism in Saccharomyces cerevisiae,” *Genome Biol Evol*, vol. 2, pp. 697–707, 2010.

[3] J. H. Gillespie, “Natural selection for within-generation variance in offspring number ii. discrete haploid models,” *Genetics*, vol. 81, pp. 403 –413, 1975.

[4] M. Lynch, “The limits to knowledge in quantitative genetics,” in *Evolutionary Biology* (M. Clegg, M. Hecht, and R. Macintyre, eds.), vol. 32 of *Evolutionary Biology*, pp. 225–237, Springer US, 2000.

[5] I. G. Szendro, M. F. Schenk, J. Franke, J. Krug, and J. A. de Visser, “Quantitative analyses of empirical fitness landscapes,” *J Stat Mech*, vol. P01005, 2013.

[6] R. Der, C. L. Epstein, and J. B. Plotkin, “Generalized population models and the nature of genetic drift,” *Theoretical Population Biology*, vol. 80, pp. 80–99, 2011.

[7] O. G. Berg and P. H. von Hippel, “Selection of DNA binding sites by regulatory proteins: Statistical-mechanical theory and application to operators and promoters,” *J Mol Biol*, vol. 193, pp. 723–743, 1987.

[8] G. D. Stormo and D. S. Fields, “Specificity, free energy and information content in protein-DNA interactions,” *TIBS*, vol. 23, pp. 109–113, 1998.

[9] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, “Identification of

direct residue contacts in proteinprotein interaction by message passing," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 67–72, 2009.

[10] C. Jckel, P. Kast, and D. Hilvert, "Protein design by directed evolution," *Annu. Rev. Biophys.*, vol. 37, pp. 153–173, 2008.

[11] M. Raviscioni, P. Gu, M. Sattar, A. J. Cooney, and O. Lichtarge, "Correlated evolutionary pressure at interacting transcription factors and dna response elements can guide the rational engineering of dna binding specificity," *Journal of Molecular Biology*, vol. 350, pp. 402–415, 2005.

[12] A. D. Ault and J. R. Broach, "Creation of gpcr-based chemical sensors by directed evolution in yeast," *Protein Engineering Design and Selection*, vol. 19, pp. 1–8, 2006.

[13] J. J. Agresti, E. Antipov, A. R. Abate, K. Ahn, A. C. Rowat, J.-C. Baret, M. Marquez, A. M. Klibanov, A. D. Griffiths, and D. A. Weitz, "Ultrahigh-throughput screening in drop-based microfluidics for directed evolution," *PNAS*, vol. 107, pp. 4004–4009, 2010.

[14] R. Fasan, Y. T. Meharenna, C. D. Snow, T. L. Poulos, and F. H. Arnold, "Evolutionary history of a specialized p450 propane monooxygenase," *Journal of Molecular Biology*, vol. 383, pp. 1069–1080, 2008.

[15] A. M. Sengupta, M. Djordjevic, and B. I. Shraiman, "Specificity and robustness in transcription control networks," *Proc Natl Acad Sci USA*, vol. 99, pp. 2072–2077, 2002.

[16] U. Gerland and T. Hwa, "On the selection and evolution of regulatory DNA motifs," *J Mol Evol*, vol. 55, pp. 386–400, 2002.

[17] J. Berg and M. Lässig, "Stochastic evolution of transcription factor binding sites," *Biophysics (Moscow)*, vol. 48, pp. S36–S44, 2003.

[18] J. Berg, S. Willmann, and M. Lässig, "Adaptive evolution of transcription factor binding sites," *BMC Evol Biol*, vol. 4, p. 42, 2004.

[19] M. Lässig, "From biophysics to evolutionary genetics: statistical aspects of gene regulation," *BMC Bioinformatics*, vol. 8, p. S7, 2007.

[20] V. Mustonen, J. Kinney, C. G. Callan, and M. Lässig, "Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites," *Proc Natl Acad Sci USA*, vol. 105, pp. 12376–12381, 2008.

[21] J. D. Bloom, J. J. Silberg, C. O. Wilke, D. A. Drummond, C. Adami, and F. H. Arnold, "Thermodynamic prediction of protein neutrality," *Proc Natl Acad Sci USA*, vol. 102, pp. 606–611, 2005.

[22] M. A. DePristo, D. M. Weinreich, and D. L. Hartl, "Missense meanderings in sequence space: a biophysical view of protein evolution," *Nat Rev Genet*, vol. 6, pp. 678–687, 2005.

[23] J. D. Bloom, S. T. Labthavikul, C. R. Otey, and F. H. Arnold, "Protein stability promotes evolvability," *Proc Natl Acad Sci USA*, vol. 103, pp. 5869–5874, 2006.

[24] K. B. Zeldovich, P. Chen, and E. I. Shakhnovich, "Protein stability imposes limits on organism complexity and speed of molecular evolution," *Proc Natl Acad Sci USA*, vol. 104, pp. 16152–16157, 2007.

[25] J. D. Bloom, A. Raval, and C. O. Wilke, "Thermodynamics of neutral protein evolution," *Genetics*, vol. 175, pp. 255–266, 2007.

[26] S. Bershtein, K. Goldin, and D. S. Tawfik, "Intense neutral drifts yield robust and evolvable consensus proteins," *J Mol Biol*, vol. 379, pp. 1029–1044, 2008.

[27] J. D. Bloom and M. J. Glassman, "Inferring stabilizing mutations from protein phylogenies: Application to influenza hemagglutinin," *PLoS Comput Biol*, vol. 5, p. e1000349, 2009.

[28] M. Manhart and A. V. Morozov, "Path-based approach to random walks on networks characterizes how proteins evolve new functions," *Phys Rev Lett*, vol. 111, p. 088102, 2013.

[29] M. Manhart and A. V. Morozov, "Statistical physics of evolutionary trajectories on fitness landscapes," in *First-Passage Phenomena and Their Applications* (R. Metzler, G. Oshanin, and S. Redner, eds.), Singapore: World Scientific, 2013.

[30] M. Manhart, A. Haldane, and A. V. Morozov, "A universal scaling law determines time reversibility and steady state of substitutions under selection," *Theor Popul Biol*, vol. 82, pp. 66–76, 2012.

[31] A. haldane, M. Manhart, and A. V. Morozov, "Biophysical fitness landscapes for transcription factor binding sites," *Submitted*, 2014.

[32] A. J. Stewart, S. Hannenhalli, and J. B. Plotkin, "Why transcription factor binding sites are ten nucleotides long," *Genetics*, vol. 192, pp. 973–985, 2012.

[33] M. Ptashne and A. Gann, *Genes and Signals.* Cold Spring Harbor: Cold Spring Harbor Laboratory Press, 2002.

[34] I. Erb and E. V. Nimwegen, "Nimwegen e: Statistical features of yeast's transcriptional regulatory code," in *IEE Proceedings Systems Biology ICCSB*, pp. 111–118, 2006.

[35] J. C. Perez and E. A. Groisman, "Evolution of transcriptional regulatory circuits in bacteria," *Cell*, vol. 138, pp. 233–244, 2009.

[36] Y. Zhao, D. Granas, and G. D. Stormo, "Inferring binding energies from selected binding sites," *PLoS Computational Biology*, vol. 5, pp. e1000590EP –, 2009.

[37] Y. Zhao, S. Ruan, M. Pandey, and G. D. Stormo, "Improved models for transcription factor binding site identification using nonindependent interactions," *Genetics*, vol. 191, pp. 781–790, 2012.

[38] R. Siddharthan, "Phylogibbs-mp: Module prediction and discriminative motif-finding by gibbs sampling," *PLoS Comput Biol*, vol. 4, pp. e1000156EP –, 2008.

[39] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-F. Wang, D. Coburn,

D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk, "Diversity and complexity in dna recognition by transcription factors," *Science*, vol. 324, pp. 1720–1723, 2009.

[40] R. Stoltenburg, C. Reinemann, and B. Strehlitz, "Selexa (r)evolutionary method to generate high-affinity nucleic acid ligands," *Biomolecular Engineering*, vol. 24, pp. 381–403, 2007.

[41] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young, "Transcriptional regulatory networks in Saccharomyces cerevisiae," *Science*, vol. 298, pp. 799–804, 2002.

[42] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J.-B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young, "Transcriptional regulatory code of a eukaryotic genome," *Nature*, vol. 431, pp. 99–104, 2004.

[43] K. MacIsaac, T. Wang, D. B. Gordon, D. Gifford, G. Stormo, and E. Fraenkel, "An improved map of conserved regulatory sites for saccharomyces cerevisiae," *BMC Bioinformatics*, vol. 7, p. 113, 2006.

[44] S. Mukherjee, M. F. Berger, G. Jona, X. S. Wang, D. Muzzey, M. Snyder, R. A. Young, and M. L. Bulyk, "Rapid analysis of the dna-binding specificities of transcription factors with dna microarrays," *Nat Genet*, vol. 36, pp. 1331–1339, 2004.

[45] M. F. Berger, A. A. Philippakis, A. M. Qureshi, F. S. He, P. W. E. III, and M. L. Bulyk, "Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities," *Nat Biotech*, vol. 24, pp. 1429–1435, 2006.

[46] B. C. Foat, A. V. Morozov, and H. J. Bussemaker, "Statistical mechanical modeling

of genome-wide transcription factor occupancy data by MatrixREDUCE," *Bioinformatics*, vol. 22, pp. e141–e149, 2006.

[47] M. Geertz and S. J. Maerkl, "Experimental strategies for studying transcription factordna binding specificities," *Briefings in Functional Genomics*, vol. 9, no. 5-6, pp. 362–373, 2010.

[48] Z. Wunderlich and L. A. Mirny, "Different gene regulation strategies revealed by analysis of binding motifs," *TIG*, vol. 25, pp. 434–440, 2009.

[49] S. Wright, "Evolution in Mendelian populations," *Genetics*, vol. 16, pp. 97–159, 1931.

[50] S. Wright, "The roles of mutation, inbreeding, crossbreeding and selection in evolution," *Proc. 6th Int. Cong. Genet.*, vol. 1, pp. 356–366, 1932.

[51] R. A. Fisher, *The Genetical Theory of Natural Selection.* New York: Dover, 1958.

[52] V. Mustonen and M. Lässig, "Fitness flux and the ubiquity of adaptive evolution," *Proc. Natl. Acad. Sci. USA*, vol. 107, pp. 4248–4253, 2010.

[53] M. Eigen, J. McCaskill, and P. Schuster, "The molecular quasi-species," *Adv. Chem. Phys.*, vol. 75, pp. 149–263, 1989.

[54] P. A. P. Moran, "Random processes in genetics," *Proc. Camb. Philos. Soc.*, vol. 54, pp. 60–71, 1958.

[55] W. Ewens, *Mathematical Population Genetics.* New York: Springer, 2004.

[56] M. M. Desai and D. S. Fisher, "Beneficial mutationselection balance and the effect of linkage on positive selection," *Genetics*, vol. 176, pp. 1759–1798, 2007.

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

[58] H. Ochman and R. K. Selander, "Evidence for clonal population structure in escherichia coli," *Proceedings of the National Academy of Sciences*, vol. 81, pp. 198–201, 1984.

[59] L. M. Wick, H. Weilenmann, and T. Egli, “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*, vol. 148, pp. 2889–2902, 2002.

[60] T. Dos Vultos, O. Mestre, J. Rauzier, M. Golec, N. Rastogi, V. Rasolofo, T. Tonjum, C. Sola, I. Matic, and B. Gicquel, “Evolution and diversity of clonal bacteria: The paradigm of mycobacterium tuberculosis,” *PLoS ONE*, vol. 3, p. e1538EP, 2008.

[61] M. Achtman, “Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens,” *Annu. Rev. Microbiol.*, vol. 62, pp. 53–70, 2008.

[62] G. A. T. McVean and J. Vieira, “Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in drosophila,” *Genetics*, vol. 157, pp. 245 –257, 2001.

[63] Z. Yang and R. Nielsen, “Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage,” *Mol. Biol. Evol.*, vol. 25, pp. 568–579, 2008.

[64] A. S. Lauring and R. Andino, “Quasispecies theory and the behavior of rna viruses,” *PLoS Pathog*, vol. 6, pp. e1001005EP –, 2010.

[65] J. Wakeley, “The limits of theoretical population genetics,” *Genetics*, vol. 169, pp. 1–7, 2005.

[66] T. L. Parsons, C. Quince, and J. B. Plotkin, “Some consequences of demographic stochasticity in population genetics,” *Genetics*, vol. 185, pp. 1345 –1354, 2010.

[67] W. Ewens, “The probability of survival of a new mutant in a fluctuating environment,” *Heredity*, vol. 22, pp. 438–443, 1967.

[68] T. Maruyama, “On the fixation probability of mutant genes in a subdivided population,” *Genet. Res. Camb.*, vol. 15, pp. 221–225, 1970.

[69] S. P. Otto and M. C. Whitlock, “The probability of fixation in populations of changing size,” *Genetics*, vol. 146, pp. 723 –733, 1997.

[70] M. Möhle, "Forward and backward diffusion approximations for haploid exchangeable population models," *Stoch. Proc. Appl.*, vol. 95, pp. 133–149, 2001.

[71] M. Möhle and S. Sagitov, "A classification of coalescent processes for haploid exchangeable population models," *Ann. Prob.*, vol. 29, pp. 1547–1562, 2001.

[72] M. Whitlock, "Fixation probability and time in subdivided populations," *Genetics*, vol. 164, pp. 767–779, 2003.

[73] M. Kimura, "On the probability of fixation of mutant genes in a population," *Genetics*, vol. 47, pp. 713–719, 1962.

[74] M. Kimura, "Evolutionary rate at the molecular level," *Nature*, vol. 217, pp. 624–626, 1968.

[75] J. Summers and S. Litwin, "Examining the theory of error catastrophe," *Journal of Virology*, vol. 80, pp. 20–26, 2006.

[76] M. Kimura and T. Ohta, *Theoretical Aspects of Population Genetics.* Princeton: Princeton University Press, 1971.

[77] L. J. S. Allen, *An Introduction to Stochastic Processes with Applications to Biology.* Boca Raton: Chapman and Hall, CRC, second ed., 2011.

[78] Z. Yang, *Computational Molecular Evolution.* Oxford: Oxford University Press, 2006.

[79] T. H. Jukes and C. R. Cantor, "Evolution of protein molecules," in *Mammalian protein metabolism* (H. N. Munro, ed.), pp. 21–123, New York: Academic Press, 1969.

[80] M. Kimura, "A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences," *J. Mol. Evol.*, vol. 16, pp. 111–120, 1980.

[81] K. Tamura and M. Nei, "Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees," *Mol. Biol. Evol.*, vol. 10, pp. 512–526, 1993.

[82] J. Felsenstein, "Evolutionary trees from DNA sequences: A maximum likelihood approach," *J. Mol. Evol.*, vol. 17, pp. 368–376, 1981.

[83] J. Felsenstein, "PHYLIP (Phylogeny Inference Package) version 3.69," 2011.

[84] F. Rodríguez, J. Oliver, A. Marín, and J. Medina, "The general stochastic model of nucleotide substitution," *J. Theor. Biol.*, vol. 142, pp. 485–501, 1990.

[85] D. Barry and J. A. Hartigan, "Asynchronous distance between homologous DNA sequences," *Biometrics*, vol. 43, pp. 261–276, 1987.

[86] F. S. Roberts, *Measurement Theory with Applications to Decisionmaking, Utility, and the Social Sciences.* Reading: Addison-Wesley, 1979.

[87] G. Sella and A. E. Hirsh, "The application of statistical physics to evolutionary biology," *Proc Natl Acad Sci USA*, vol. 102, pp. 9541–9546, 2005.

[88] R. Lenski and S. F. Elena, "Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation," *Nat. Rev. Genet.*, vol. 4, pp. 457–469, 2003.

[89] C. Cannings, "The latent roots of certain markov chains arising in genetics: A new approach, i. haploid models," *Adv. Appl. Prob.*, vol. 6, pp. 260–290, 1974.

[90] Y. Iwasa, "Free fitness that always increases in evolution," *J. Theor. Biol.*, vol. 135, pp. 265–281, 1988.

[91] N. Champagnat, "A microscopic interpretation for adaptive dynamics trait substitution sequence models," *Stoch. Process. Appl.*, vol. 116, pp. 1127–1160, 2006.

[92] N. Champagnat, R. Ferrière, and S. Méléard, "Unifying evolutionary dynamics: From individual stochastic processes to macroscopic models," *Theor. Pop. Biol.*, vol. 69, pp. 297–321, 2006.

[93] G. W. Harrison, "Numerical solution of the fokker planck equation using moving finite elements," *Numerical Methods for Partial Differential Equations*, vol. 4, pp. 219–232, 1988.

[94] M. Kimura, “Solution of a process of random genetic drift with a continuous model,” *Proc. Natl. Acad. Sci. USA*, vol. 41, pp. 144–150, 1955.

[95] H. Wichman, M. Badgett, L. Scott, C. Boulianne, and J. Bull, “Different trajectories of parallel evolution during viral adaptation,” *Science*, vol. 285, pp. 422–424, 1999.

[96] J. Bull, M. Badgett, and H. Wichman, “Big-benefit mutations in a bacteriophage inhibited with heat,” *Mol. Biol. Evol.*, vol. 17, pp. 942–950, 2000.

[97] K. Holder and J. Bull, “Profiles of adaptation in two similar viruses,” *Genetics*, vol. 159, pp. 1393–1404, 2001.

[98] R. Barrett, L. M’Gonigle, and S. Otto, “The distribution of beneficial mutant effects under strong selection,” *Genetics*, vol. 174, pp. 2071–2079, 2006.

[99] H. Orr, “The genetics of species differences,” *Trends Ecol. Evol.*, vol. 16, pp. 343–350, 2001.

[100] H. Orr, “The genetic theory of adaptation: A brief history,” *Nat. Rev. Genet.*, vol. 6, pp. 119–127, 2005.

[101] A. Eyre-Walker and P. Keightley, “The distribution of fitness effects of new mutations,” *Nat. Rev. Genet.*, vol. 8, pp. 610–618, 2007.

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

[103] R. Barrett, R. MacLean, and G. Bell, “Mutations of intermediate effect are responsible for adaptation in evolving *Pseudomonas fluorescens* populations,” *Biol. Lett.*, vol. 2, pp. 236–238, 2006.

[104] R. Bürger, *The Mathematical Theory of Selection, Recombination, and Mutation.* New York: Wiley, 2000.

[105] S. R. Proulx, “The ESS under spatial variation with applications to sex allocation,” *Theor. Pop. Biol.*, vol. 58, pp. 33–47, 2000.

[106] M. Shpak, “Selection against demographic stochasticity in Age-Structured populations,” *Genetics*, vol. 177, pp. 2181 –2194, 2007.

[107] J. H. Gillespie, “Natural selection for within-generation variance in offspring number,” *Genetics*, vol. 76, pp. 601 –606, 1974.

[108] W.-H. Li, “Models of nearly neutral mutations with particular implications for non-random usage of synonymous codons,” *J. Mol. Evol*, vol. 24, pp. 337–345, 1987.

[109] M. Kimura, “Some problems of stochastic processes in genetics,” *Ann. Math. Stat.*, vol. 28, pp. 882–901, 1957.

[110] J. G. Kemeny and J. L. Snell, *Finite Markov Chains.* New York: Van Nostrand, 1960.

[111] P. Sjödin, I. Kaj, S. Krone, M. Lascoux, and M. Nordborg, “On the meaning and existence of an effective population size,” *Genetics*, vol. 169, pp. 1061–1070, 2005.

[112] J. H. Gillespie, “Natural selection for variances in offspring numbers: A new evolutionary principle,” *Am. Nat.*, vol. 111, pp. 1010–1014, 1977.

[113] G. A. Watterson, “Reversibility and the age of an allele. ii.,” *Theor. Pop. Biol.*, vol. 12, pp. 179–196, 1977.

[114] B. Levikson, “The age distribution of markov processes,” *J. Appl. Prob.*, vol. 14, pp. 492–506, 1977.

[115] T. G. Kurtz, *Approximation of Population Processes.* CBMS-NSF Regional Conference Series in Applied Mathematics, Philadelphia: Society for Industrial and Applied Mathematics, 1981.

[116] W. J. Ewens, “Population genetics theory – the past and the future,” in *Mathematical and Statistical Developments of Evolutionary Theory* (S. Lessard, ed.), pp. 177–227, Amsterdam: Kluwer Academic Publishers, 1990.

[117] J. F. C. Kingman, “The coalescent,” *Stoch. Proc. Appl.*, vol. 13, pp. 235–248, 1982.

[118] G. A. Watterson, “Reversibility and the age of an allele. i.,” *Theor. Pop. Biol.*, vol. 10, pp. 239–253, 1976.

[119] M. G. Bulmer, “The selection-mutation-drift theory of synonymous codon usage,” *Genetics*, vol. 129, pp. 897–907, 1991.

[120] G. A. McVean and B. Charlesworth, “A population genetics model for the evolution of synonymous codon usage: patterns and predictions,” *Genet. Res.*, vol. 74, pp. 145–158, 1999.

[121] G. A. McVean and J. Vieira, “The evolution of codon preferences in drosophila: a maximum-likelihood approach to parameter estimation and hypothesis testing,” *J. Mol. Evol.*, vol. 49, pp. 63–75, 1999.

[122] R. Nielsen, V. L. B. DuMont, M. J. Hubisz, and C. F. Aquadro, “Maximum likelihood estimation of ancestral codon usage bias parameters in drosophila,” *Mol. Biol. Evol.*, vol. 24, pp. 228–235, 2007.

[123] S. Bershtein, M. Segal, R. Bekerman, N. Tokuriki, and D. S. Tawfik, “Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein,” *Nature*, vol. 444, pp. 929–932, 2006.

[124] D. M. Weinreich, N. F. Delaney, M. A. DePristo, and D. L. Hartl, “Darwinian evolution can follow only very few mutational paths to fitter proteins,” *Science*, vol. 312, pp. 111–114, 2006.

[125] F. J. Poelwijk, D. J. Kiviet, D. M. Weinreich, and S. J. Tans, “Empirical fitness landscapes reveal accessible evolutionary paths,” *Nature*, vol. 445, pp. 383–386, 2007.

[126] H. A. Orr, “Fitness and its role in evolutionary genetics,” *Nat Rev Genet*, vol. 10, pp. 531–539, 2009.

[127] E. A. Winzeler, D. D. Shoemaker, A. Astromoff, H. Liang, K. Anderson, B. Andre, R. Bangham, R. Benito, J. D. Boeke, H. Bussey, A. M. Chu, C. Connelly, K. Davis, F. Dietrich, S. W. Dow, M. El Bakkoury, F. Foury, S. H. Friend, E. Gentalen, G. Giaever, J. H. Hegemann, T. Jones, M. Laub, H. Liao, N. Liebundguth, D. J. Lockhart, A. Lucau-Danila, M. Lussier, N. M’Rabet, P. Menard, M. Mittmann, C. Pai, C. Rebischung, J. L. Revuelta, L. Riles, C. J. Roberts, P. Ross-MacDonald, B. Scherens,

M. Snyder, S. Sookhai-Mahadeo, R. K. Storms, S. Véronneau, M. Voet, G. Volckaert, T. R. Ward, R. Wysocki, G. S. Yen, K. Yu, K. Zimmermann, P. Philippsen, M. Johnston, and R. W. Davis, "Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis," *Science*, vol. 285, pp. 901–906, 1999.

[128] N. Champagnat, "A microscopic interpretation for adaptive dynamics trait substitution sequence models," *Stoch Proc Appl*, vol. 116, pp. 1127–1160, 2006.

[129] M. Kimura and T. Ohta, "The average number of generations until fixation of a mutant gene in a finite population," *Genetics*, vol. 61, pp. 763–771, 1969.

[130] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov Chains and Mixing Times.* Providence, RI: American Mathematical Society, 2009.

[131] A. Sarai and Y. Takeda, "Lambda repressor recognizes the approximately 2-fold symmetric half-operator sequences asymmetrically," *Proc. Natl. Acad. Sci. USA*, vol. 86, pp. 6513–6517, 1989.

[132] N. Lehming, J. Sartorius, B. Kisters-Woike, B. von Wilcken-Bergmann, and B. Muller-Hill, "Mutant lac repressors with new specificities hint at rules for protein-DNA recognition," *EMBO J.*, vol. 9, pp. 615–621, 1990.

[133] I. J. Tsai, D. Bensasson, A. Burt, and V. Koufopanou, "Population genomics of the wild yeast Saccharomyces paradoxus: Quantifying the life cycle," *Proc Natl Acad Sci USA*, vol. 105, pp. 4957–4962, 2008.

[134] B. Dujon, "Yeast evolutionary genomics," *Nat Rev Genet*, vol. 11, pp. 512–524, 2010.

[135] G. Liti, D. M. Carter, A. M. Moses, J. Warringer, L. Parts, S. A. James, R. P. Davey, I. N. Roberts, A. Burt, V. Koufopanou, I. J. Tsai, C. M. Bergman, D. Bensasson, M. J. T. O'Kelly, A. van Oudenaarden, D. B. H. Barton, E. Bailes, A. N. Nguyen, M. Jones, M. A. Quail, I. Goodhead, S. Sims, F. Smith, A. Blomberg, R. Durbin, and E. J. Louis, "Population genomics of domestic and wild yeasts," *Nature*, vol. 458, pp. 337–341, 2009.

[136] S. W. Doniger, H. S. Kim, D. Swain, D. Corcuera, M. Williams, S.-P. Yang, and J. C. Fay, “A catalog of neutral and deleterious polymorphism in yeast,” *PLoS Genet*, vol. 4, p. e1000183, 2008.

[137] J. C. Fay and J. A. Benavides, “Evidence for domesticated and wild populations of ¡italic¿saccharomyces cerevisiae¡/italic¿,” *PLoS Genet*, vol. 1, pp. e5EP –, 2005.

[138] T. Replansky, V. Koufopanou, D. Greig, and G. Bell, “Saccharomyces sensu stricto as a model system for evolution and ecology,” *Trends in Ecology & Evolution*, vol. 23, pp. 494–501, 2008.

[139] G. Giaever, A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, A. P. Arkin, A. Astromoff, M. El Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K.-D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Guldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kotter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C. Wang, T. R. Ward, J. Wilhelmy, E. A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis, and M. Johnston, “Functional profiling of the saccharomyces cerevisiae genome,” *Nature*, vol. 418, pp. 387–391, 2002.

[140] F. C. P. Holstege, E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander, and R. A. Young, “Dissecting the regulatory circuitry of a eukaryotic genome,” *Cell*, vol. 95, pp. 717–728, 1998.

[141] M. W. Hahn, J. E. Stajich, and G. A. Wray, “The effects of selection against spurious transcription factor binding sites,” *Mol Biol Evol*, vol. 20, pp. 901–906, 2003.

[142] M. Djordjevic, A. M. Sengupta, and B. I. Shraiman, “A biophysical approach to

transcription factor binding site discovery," *Genome Res*, vol. 13, pp. 2381–2390, 2003.

[143] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodal Inference: A Practical Information-Theoretic Approach.* New York: Springer-Verlag, second ed., 2002.

[144] I. K. Jordan, I. B. Rogozin, Y. I. Wolf, and E. V. Koonin, "Essential genes are more evolutionarily conserved than are nonessential genes in bacteria," *Genome Res*, vol. 12, pp. 962–968, 2002.

[145] C. Pal, B. Papp, and L. D. Hurst, "Genomic function (communication arising): Rate of evolution and gene dispensability," *Nature*, vol. 421, pp. 496–497, 2003.

[146] J. Zhang and X. He, "Significant impact of protein dispensability on the instantaneous rate of protein evolution," *Mol Biol Evol*, vol. 22, pp. 1147–1155, 2005.

[147] J. K. Choi, S. C. Kim, J. Seo, S. Kim, and J. Bhak, "Impact of transcriptional properties on essentiality and evolutionary rate," *Genetics*, vol. 175, pp. 199–206, 2007.

[148] D. M. Krylov, Y. I. Wolf, I. B. Rogozin, and E. V. Koonin, "Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution," *Genome Res*, vol. 13, pp. 2229–2235, 2003.

[149] Z. Wang and J. Zhang, "Why is the correlation between gene importance and gene evolutionary rate so weak?," *PLoS Genet*, vol. 5, p. e1000329, 2009.

[150] G. Fang, E. Rocha, and A. Danchin, "How essential are nonessential genes?," *Mol Biol Evol*, vol. 22, pp. 2147–2156, 2005.

[151] A. E. Hirsh and H. B. Fraser, "Protein dispensability and rate of evolution," *Nature*, vol. 411, pp. 1046–1049, 2001.

[152] F. Gao, B. Foat, and H. Bussemaker, "Defining transcriptional networks through integrative modeling of mrna expression and transcription factor binding data," *BMC Bioinformatics*, vol. 5, p. 31, 2004.

[153] Z. Hu, P. J. Killion, and V. R. Iyer, "Genetic reconstruction of a functional transcriptional regulatory network," *Nat Genet*, vol. 39, pp. 683–687, 2007.

[154] S. W. Doniger and J. C. Fay, "Frequent gain and loss of functional transcription factor binding sites," *PLoS Comput Biol*, vol. 3, p. e99, 2007.

[155] D. Raijman, R. Shamir, and A. Tanay, "Evolution and selection in yeast promoters: Analyzing the combined effect of diverse transcription factor binding sites," *PLoS Comput Biol*, vol. 4, p. e7, 2008.

[156] I. Tirosh, A. Weinberger, D. Bezalel, M. Kaganovich, and N. Barkai, "On the relation between promoter divergence and gene expression evolution," *Mol Syst Biol*, vol. 4, 2008.

[157] B. B. Tuch, D. J. Galgoczy, A. D. Hernday, H. Li, and A. D. Johnson, "The evolution of combinatorial gene regulation in fungi," *PLoS Biol*, vol. 6, p. e38, 2008.

[158] R. Jovelin and P. Phillips, "Evolutionary rates and centrality in the yeast gene regulatory network," *Genome Biol*, vol. 10, p. R35, 2009.

[159] S. Wuchty, Z. N. Oltvai, and A.-L. Barabasi, "Evolutionary conservation of motif constituents in the yeast protein interaction network," *Nat Genet*, vol. 35, pp. 176–179, 2003.

[160] A. D. J. van Dijk, S. van Mourik, and R. C. H. J. van Ham, "Mutational robustness of gene regulatory networks," *PLoS ONE*, vol. 7, p. e30591, 2012.

[161] C. R. Baker, B. B. Tuch, and A. D. Johnson, "Extensive DNA-binding specificity divergence of a conserved transcription regulator," *Proc Natl Acad Sci USA*, vol. 108, pp. 7493–7498, 2011.

[162] X. He, T. S. P. C. Duque, and S. Sinha, "Evolutionary origins of transcription factor binding site clusters," *Mol Biol Evol*, vol. 29, pp. 1059–1070, 2012.

[163] B. Z. He, A. K. Holloway, S. J. Maerkl, and M. Kreitman, "Does positive selection drive transcription factor binding site turnover? a test with drosophila cis-regulatory modules," *PLoS Genet*, vol. 7, p. e1002053, 2011.

[164] N. Habib, I. Wapinski, H. Margalit, A. Regev, and N. Friedman, “A functional selection model explains evolutionary robustness despite plasticity in regulatory networks,” *Mol Syst Biol*, vol. 8, 2012.

[165] J. M. Cherry, E. L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E. T. Chan, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. R. Engel, D. G. Fisk, J. E. Hirschman, B. C. Hitz, K. Karra, C. J. Krieger, S. R. Miyasato, R. S. Nash, J. Park, M. S. Skrzypek, M. Simison, S. Weng, and E. D. Wong, “Saccharomyces genome database: the genomics resource of budding yeast,” *Nucl Acids Res*, vol. 40, pp. D700–D705, 2012.

[166] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins, “Clustal W and Clustal X version 2.0,” *Bioinformatics*, vol. 23, pp. 2947–2948, 2007.

[167] Z. Yang, “PAML 4: Phylogenetic Analysis by Maximum Likelihood,” *Mol Biol Evol*, vol. 24, pp. 1586–1591, 2007.