



Long-term evolution on complex fitness landscapes when mutation is weak

David M. McCandlish¹

Received: 31 December 2017 / Revised: 4 August 2018 / Accepted: 6 August 2018 / Published online: 19 September 2018
© The Genetics Society 2018

Abstract

Understanding evolution on complex fitness landscapes is difficult both because of the large dimensionality of sequence space and the stochasticity inherent to population-genetic processes. Here, I present an integrated suite of mathematical tools for understanding evolution on time-invariant fitness landscapes when mutations occur sufficiently rarely that the population is typically monomorphic and evolution can be modeled as a sequence of well-separated fixation events. The basic intuition behind this suite of tools is that surrounding any particular genotype lies a region of the fitness landscape that is easy to evolve to, while other pieces of the fitness landscape are difficult to evolve to (due to distance, being across a fitness valley, etc.). I propose a rigorous definition for this “dynamical neighborhood” of a genotype which captures several aspects of the distribution of waiting times to evolve from one genotype to another. The neighborhood structure of the landscape as a whole can be summarized as a matrix, and I show how this matrix can be used to approximate the expected waiting time for certain evolutionary events to occur and to provide an intuitive interpretation to existing formal results on the index of dispersion of the molecular clock.

Introduction

Understanding evolutionary dynamics on general fitness landscapes remains an important project for theoretical population genetics (Wright 1932; Kauffman and Levin 1987; Kauffman 1993; Gavrillets 2004). This project consists of two complementary sets of goals. First, there is a quantitative component: we must be able to mathematically characterize the dynamics that a fitness landscape induces for an evolving population. For instance, we might want to know the number of mutations fixed before a population reaches a local fitness maximum (Gillespie 1983, 1984, 1991; Macken and Perelson 1989; Orr 1998; Jain and Seetharaman 2011; Seetharaman and Jain 2014) or the expected time for a population to cross a fitness valley (Iwasa et al. 2004; Weinreich and Chao 2005; Weissman et al. 2009). Second, there is a qualitative component whose goal is to simplify the structure of the fitness landscape and the evolutionary dynamics it induces in order to aid

understanding and allow useful generalizations. Common approaches in this vein include methods to summarize the ruggedness of fitness landscapes through correlation distances and the Walsh-Fourier transform (Weinberger 1990, 1991, Fontana et al. 1993; Stadler 2003; Weinreich et al. 2013; Neidhart et al. 2013; Poelwijk et al. 2016) and studying epistasis in various other frameworks (Wolf et al. 2000; Kondrashov and Kondrashov 2001; Phillips 2008; Beerenwinkel et al. 2007; Carneiro and Hartl 2010; Østman et al. 2011; Szendro et al. 2013; Tenaillon 2014; Ferretti et al. 2016; Bank et al. 2016; Crona et al. 2017). It also includes methods meant to summarize the repeatability of evolutionary trajectories by considering the rank ordering of genotypes by fitness (Weinreich 2005; Weinreich et al. 2005, 2006), or by looking at the variation between trajectories (Lobkovsky et al. 2011; de Visser and Krug 2014). The need for this second component arises due to the size of fitness landscapes available to be analyzed. The fitness landscapes available today, whether computational (Kauffman 1993; Gruner et al. 1996; Wagner 1996, 2005, 2011; Bornberg-Bauer and Chan 1999; Fontana 2002; Cowperthwaite and Meyers 2007) or empirical (Weinreich et al. 2006; Lozovsky et al. 2009; de Visser et al. 2009; Pitt and Ferré-D’Amaré 2010; Hinkley et al. 2011; Sarkisyan et al. 2016; Wu et al. 2016; Aguilar-Rodríguez et al. 2017), contain anywhere from 16 or 32 genotypes up to hundreds

✉ David M. McCandlish
mccandlish@cshl.edu

¹ Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

of thousands of genotypes; a full description of the evolutionary dynamics at this scale would be unwieldy at best and unintelligible at worst. More subtly, these high-dimensional fitness landscapes may have geometric features that are very different from those suggested by (Wright 1932)'s original pictorial metaphor (Gavrilets and Gravner 1997; Gavrilets 1997, 2003, 2004, Gravner et al. 2007), and thus developing a new, more accurate set of intuitions for evolutionary dynamics in high-dimensional spaces remains an important challenge.

These two goals, calculation and comprehension, need not remain separate. Indeed, one of the signs of a mature theory is that the same set of quantities can be used for both. Here, I present a method for both calculating and understanding various quantities of evolutionary interest for a population of constant size evolving on an arbitrary, time-invariant fitness landscape. However, to develop this theory we must make two key assumptions. First, we assume that mutation is weak relative to drift, so that the population is typically monomorphic and each new mutation that enters the population is fixed or lost before the next mutation occurs. In this regime we can model evolution using an “origin-fixation” model, where evolution takes the form of a Markov chain on the set of genotypes and the population is modeled as a single particle that jumps from genotype to genotype with each substitution (see McCandlish and Stoltzfus 2014, for a review). Second, we assume a time-reversible mutation model and allow deleterious substitutions. Under these assumptions the evolutionary process can be modeled as a reversible Markov chain (Iwasa 1988; Berg et al. 2004; Sella and Hirsh 2005), a class of mathematical models for which a simple and elegant mathematical theory is available (e.g., Kelly 1979; Kielson 1979; Aldous and Fill 2002). Unlike Gillespie's classical strong-selection weak-mutation formalism (Gillespie 1983, 1984) where evolution ends when a local fitness maximum is reached, these models also capture the transition from the adaptive transient into a period of long-term nearly neutral evolution (Hartl et al. 1985; Hartl and Taubes 1996; McCandlish et al. 2015a) where, over much longer time scales, the population first explores the structure of the fitness landscape as a whole and then finally transitions into a stationary stochastic process where all information about the starting genotype has been lost.

Previously, I have used this framework to visualize the structure of fitness landscapes (McCandlish 2011), to characterize the influence of the fitness landscape on the evolutionary accessibility of genotypes (McCandlish 2013), and to understand the impact of epistasis on the process of adaptation (McCandlish et al. 2015b) and the dynamics of reversion (McCandlish et al. 2016). Here, I will provide a characterization of what it means for one genotype to be “nearby” or evolutionarily accessible from another. While

on a fully random landscape one can think about evolutionary nearness using mutational distance (Jain and Krug 2007), in complex fitness landscapes the evolutionary process is highly anisotropic, as populations may more frequently snake around on high-dimensional fitness ridges rather than cross even relatively narrow valleys (van Nimwegen and Crutchfield 2000). Nonetheless, in our weak-mutation/reversible Markov chain framework it turns out that we can define a correspondingly anisotropic “dynamical neighborhood” around a genotype that reflects how populations tend to move through the fitness landscape and which exactly characterizes the set of genotypes that are “near” to the focal genotype in several well-defined senses.

The key formal object underlying our definition of the dynamical neighborhood is a matrix known in the Markov chain literature as the fundamental matrix (Meyer 1975; Hunter 1983; Aldous and Fill 2002). It captures how being at a particular genotype changes the expected amount of time spent at any other genotype. I use the fundamental matrix to analyze the problem of waiting times for certain classes of evolutionary events to occur (e.g., arriving at a genotype with a particular phenotype or experiencing a certain type of substitution), and apply similar ideas to analyze the index of dispersion of the counting process associated with these events. Both analyses come down to computing the expected number of extra events that will occur in the future, given that an event just occurred. The main intuition is that if, e.g., the genotypes that code for a phenotype are all in each other's dynamical neighborhoods, then the times that the population displays that phenotype will occur in large clumps. Since the index of dispersion is low for clock-like, regular processes, and high for clumpy, sporadic processes, such a pattern will tend to result in an elevated index of dispersion for the associated counting process. Moreover, the expected waiting time for a population to evolve such a phenotype will tend to be elevated relative to a phenotype that is displayed at more even intervals. This is because the clumps of time when the population is fixed for that phenotype are separated by large temporal gaps where the phenotype is not observed.

In what follows I will first explain the general theory and then demonstrate the application of these techniques by analyzing a one-codon fitness landscape (for other analyses of this landscape, see McCandlish 2011, 2013). See Table 1 for a summary of mathematical notation.

Methods

Population-genetic model

Assume a haploid, asexual population of size N , where mutations arise sufficiently infrequently that each new

mutation is typically lost or fixed before the next new mutation enters the population. We can capture the dynamics of such a population by defining a transition matrix P such that $P(i, j)$ is the per-generation probability that a population currently fixed for genotype i will produce a mutant of genotype j that is destined to become fixed in the population (for a review, see McCandlish and Stoltzfus 2014). In particular, if we model evolution as a Wright–Fisher process, we have:

$$P(i, j) = \begin{cases} NM(i, j) \frac{1 - e^{-2(f(j) - f(i))}}{1 - e^{-2N(f(j) - f(i))}} & \text{if } i \neq j \\ 1 - \sum_{k \neq i} P(i, k) & \text{if } i = j \end{cases} \quad (1)$$

where $M(i, j)$ is the mutation rate from genotype i to genotype j , and $f(i)$ is the Malthusian fitness of genotype i . We assume that the genotypes are path-connected by mutation and that the mutational dynamics are reversible, i.e., that for all i, j they satisfy $\pi_M(i)M(i, j) = \pi_M(j)M(j, i)$ for some probability distribution given by the vector π_M . Then the stationary distribution π such that $\pi^T P = \pi^T$ is given by:

$$\pi(i) = \frac{\pi_M(i) e^{2(N-1)f(i)}}{\sum_k \pi_M(k) e^{2(N-1)f(k)}}, \quad (2)$$

and $\pi(i)P(i, j) = \pi(j)P(j, i)$ for all genotype i and j , so that the Markov chain defined by P is also reversible (Sella and Hirsh 2005). Although for clarity we have presented this discussion in terms of a haploid, Wright–Fisher population and have used Kimura (1962)’s formula for the fixation probability, the evolutionary dynamics can also be modeled as a reversible Markov chain for diploid populations assuming that the fitness effects of alleles are additive. One can also use the classical probability of fixation from Fisher (1930) or Wright (1931) (see, e.g., Berg et al. 2004, McCandlish et al. 2015a) to produce a model involving only the compound parameter $S = 4Ns$, or use the exact fixation probability for a Moran process (Sella and Hirsh 2005) so that the Markov chain itself becomes exact in the limit as the mutation rate approaches zero. The above theory applies in all these cases with straight-forward modifications to the formulas for P and π .

Results

The dynamical neighborhood of a genotype

Given that the population starts at some genotype i , in expectation it spends more time at some genotypes and less at others than it would have had its initial genotype been drawn from the stationary distribution π . In particular, in

expectation such a population spends:

$$Z(i, j) \equiv \sum_{t=0}^{\infty} (P^t(i, j) - \pi(j)) \quad (3)$$

extra generations at genotype j . This matrix Z plays a central role in the theory of finite state Markov chains and can be used to calculate many important quantities. For instance, let $E_{\pi}T_i$ be the expected waiting time for the population to first become fixed for genotype i given that its initial genotype is drawn from the stationary distribution, and E_jT_i be the expectation of the same waiting time starting from initial genotype j . Then

$$E_{\pi}T_i = \frac{Z(i, i)}{\pi(i)}, \quad (4)$$

and

$$E_jT_i = E_{\pi}T_i - \frac{Z(j, i)}{\pi(i)} \quad (5)$$

(see, e.g., Aldous and Fill 2002; Hunter 1983).

In previous work (McCandlish 2013), I argued that $E_{\pi}T_i$, the expected waiting time to first arrive at genotype i when the initial state is drawn from the stationary distribution, provides a natural index for the “findability” of a genotype in the sense that the rank ordering of the $E_{\pi}T_i$ implies a rank ordering on several other aspects of the distribution of waiting times for the population to arrive at specific genotypes. In particular, for finite state reversible Markov chains we have the following relations (Chapter 3, Section 4.1, Aldous and Fill 2002; McCandlish 2013):

$$E_iT_j - E_jT_i = E_{\pi}T_j - E_{\pi}T_i \quad (6)$$

and

$$\frac{P_{\pi}(T_i < T_j)}{P_{\pi}(T_j < T_i)} = \frac{E_iT_j}{E_jT_i}, \quad (7)$$

where the notation $P_{\pi}(T_i < T_j)$ denotes the probability that the population reaches i before j given that its initial state is drawn from π . Together, the two above results imply that:

$$E_{\pi}T_i \leq E_{\pi}T_j \Leftrightarrow E_jT_i \leq E_iT_j \Leftrightarrow P_{\pi}(T_i < T_j) \geq P_{\pi}(T_j < T_i). \quad (8)$$

Thus, the rank ordering of the expected waiting time to evolve to a genotype given a stationary start also determines which of a pair of genotypes is most likely to evolve first from a stationary start as well as the asymmetry in expected waiting times from a nonstationary start.

Further analysis of the matrix Z allows us to extend these ideas to define the “dynamical neighborhood” of a genotype

Table 1 List of mathematical symbols

Symbol	Interpretation
N	Population size
$f(i)$	Malthusian fitness of genotype i
$M(i,j)$	Mutation rate from genotype i to genotype j
X_t	Population's genotype at time t
$P(i,j)$	Probability that $X_t = j$ given that $X_{t-1} = i$
$\pi_M(i)$	Stationary probability of genotype i under neutrality
$\pi(i)$	Stationary probability of genotype i
$Z(i,j)$	Expected extra generations spent at genotype j for a population that starts at genotype i
$E_\pi T_i$	Expected number of generations to evolve to genotype i when initial state is drawn from stationary distribution
$E_j T_i$	Expected number of generations to evolve to genotype i for a population that starts at genotype j
$P_\pi(T_i < T_j)$	Probability of evolving to genotype i before genotype j when initial state is drawn from stationary distribution
$\mathcal{N}(i)$	Dynamical neighborhood of genotype i
μ, l, \mathcal{A}	Mutation rate, sequence length, and alphabet, when assuming complete neutrality and a uniform rate of point mutations
$C(i,j)$	Probability that an event occurs at time t given $X_{t-1} = i$ and $X_t = j$
$V(i,j)$	Probability that an event occurs at time t and $X_t = j$ given $X_{t-1} = i$
$\mathbf{1}$	Vector whose entries are all 1
$g(i)$	Probability that an event occurs at time t given that $X_{t-1} = i$
$c(i)$	Probability at stationarity that a population is at genotype i at time t given that an event has occurred at time t
$E_\pi g$	Rate that an event defined by V occurs at stationarity
$E_\pi T_V$	Expected waiting time for an event defined by V to occur when the initial state is drawn from the stationary distribution
\mathcal{P}	Set of genotypes that encode a specific phenotype
R_∞	Asymptotic index of dispersion of the counting process for the event defined by V
D	Diagonal matrix with $D(i, i) = \pi(i)$
$x_{i,j}(t)$	Probability that population is at genotype j at time t given that it was at genotype i at time 0
λ_k, l_k, r_k	The k th largest eigenvalue of P and its associated left and right eigenvectors
τ_k	Time-scale associated λ_k
u_k	Embedding axis from (McCandlish 2011) associated with r_k

i as a set of genotypes that are unusually “near” to i with respect to the dynamics of an evolving population. In particular, let us define the dynamical neighborhood of i , $\mathcal{N}(i)$, to be the set of genotypes j such that starting at i increases the expected amount of time a population will spend at j in excess of the time it would have spent there starting at stationarity, i.e., with $Z(i, j) > 0$. In symbols, let

$$\mathcal{N}(i) \equiv \{j | Z(i, j) > 0\}. \quad (9)$$

In Appendix A we show that $\mathcal{N}(i)$ is a mutationally contiguous set of genotypes that includes i . Furthermore, $\mathcal{N}(i)$ consists exactly of the alternative genotypes j with the following properties:

1. The probability of being at genotype j is enriched when starting at i (i.e., $Z(i, j) > 0$).
2. The probability of being at genotype i is enriched

when starting at j (i.e., $Z(j, i) > 0$).

3. The expected waiting time to evolve from i to j is less than the expected waiting time to evolve j from an equilibril start (i.e., $j \in \mathcal{N}(i) \Leftrightarrow E_j T_j < E_\pi T_j$).
4. The expected waiting time to evolve from j to i is less than the expected waiting time to evolve i from an equilibril start (i.e., $j \in \mathcal{N}(i) \Leftrightarrow E_j T_i < E_\pi T_i$).

where the first two statements follow from the definition of $\mathcal{N}(i)$ and the relation $\pi(i)Z(i, j) = \pi(j)Z(j, i)$ (which follows from $\pi(i)P^t(i, j) = \pi(j)P^t(j, i)$ and the definition of Z), and the latter two combine these facts with Eq. (5). Notice that these relations are symmetric between i and j , so that in addition we have $j \in \mathcal{N}(i) \Leftrightarrow i \in \mathcal{N}(j)$.

To gain some better intuition about $\mathcal{N}(i)$, it is helpful to consider two limiting cases. First, consider the neutral case where the space of genotypes are sequences of length l drawn from an alphabet \mathcal{A} and all possible point mutations

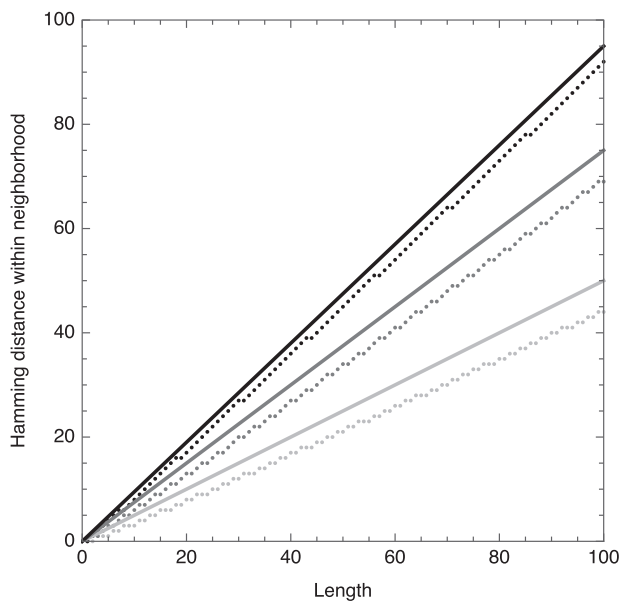


Fig. 1 Size of the dynamical neighborhood for a neutral fitness landscapes with $|\mathcal{A}|$ alleles per site as a function of sequence length. Light gray, gray, and black dots show the maximal Hamming distance from i within $\mathcal{N}(i)$ for $|\mathcal{A}| = 2, 4,$ and $20,$ respectively. Solid lines are again colored light gray, gray, and black for $|\mathcal{A}| = 2, 4,$ and $20,$ respectively and show expected distance from i to a random sequence as a function of sequence length, demonstrating that this distance can serve as a crude approximation of the size of $\mathcal{N}(i)$

occur at a uniform rate μ . In this case, there is an analytical formula for the entries of Z where for fixed l and $|\mathcal{A}|$ the sign of $Z(i, j)$ depends only on the Hamming distance between sequences i and j (see Appendix B). Using this analytical formula, we can plot the maximum Hamming distance from i still included in $\mathcal{N}(i)$ (Fig. 1). Intuitively, one might guess that the sequences in the dynamical neighborhood of i would be those sequences closer to i than would be expected for a random sequence, i.e., sequences with distance less than $l(|\mathcal{A}| - 1)/|\mathcal{A}|$. The solid lines in Fig. 1 show that this is reasonable guess but that typically the neutral dynamical neighborhoods are in fact slightly smaller. Second, consider the case where at short times the strong-selection weak-mutation dynamics provide a good approximation, so that a population starting at i after a short period is likely to end up at one of n fitness peaks p_1, \dots, p_n , where the probability of first arriving at peak p_k is given by w_k . Assuming this adaptation is sufficiently rapid, we then have $Z(i, j) \approx \sum_{k=1}^n w_k Z(p_k, j)$, so that the dynamical neighborhood of i consists of the genotypes that are enriched when starting at the peaks that populations at i tend to adapt to. In particular, we expect $\mathcal{N}(i) \subseteq \bigcup_{k=1}^n \mathcal{N}(p_k)$.

It is helpful to compare the dynamical neighborhood of a genotype i with its mutational neighborhood, that is, the set of genotypes that i can directly mutate to. These neighborhoods are not necessarily the same. First, the dynamical neighborhood of a genotype might include other genotypes

many substitutions away. Second, it is easy to construct examples where the mutational neighborhood includes genotypes that are not in the dynamical neighborhood. For instance, this circumstance often arises in multi-peaked fitness landscapes where most deleterious substitutions from a fitness peak result in a return of the population to the same peak via a reversion or compensatory substitution, but a few special mutations into the fitness valley have a tendency to resolve to a different peak (we will see some examples of this later). Just as the path of a particle is continuous when its position at each moment remains in the neighborhood of its last position and discontinuous if it jumps to a position that is outside the neighborhood of its last position, we can distinguish between continuous substitutions in genotype space where the population fixes a mutation that is in the dynamical neighborhood of the focal genotype and discontinuous substitutions where the population fixes a mutation that is outside its dynamical neighborhood (Fontana and Schuster 1998a, 1998b; Stadler et al. 2001). The intuition here is that most substitutions are continuous in that the population remains in the same “region” of the fitness landscape, such as in the basin of attraction of a particular fitness peak, but that certain special mutations move the population to a new region of the landscape, such as mutations into a fitness valley that tend to resolve via peak shift rather than reversion. While relatively rare, such discontinuous substitutions have several unusual properties which I will highlight in what follows.

The findability of phenotypes and waiting times for specific subsets of substitutions

It is natural to ask whether the theory for the findability of genotypes (McCandlish 2013) can be extended to a theory for the findability of phenotypes. In the context of modeling the evolution of a population under weak mutation as a reversible Markov chain, we can identify a phenotype of interest with the subset of genotypes that produce that phenotype, for instance a phenotype might correspond to the set of nucleotide sequences that code for a particular amino acid sequence. Thus, to extend the theory of findability to phenotypes we must calculate the expected time for a population whose initial state is drawn from the stationary distribution to first become fixed for any member of the set of genotypes that produce our phenotype of interest. Unfortunately, there is no general, simple, exact formula analogous with Eq. (4) for the expected waiting time to arrive at a subset of states. While it is possible to find the waiting time for a subset of states using the theory of absorbing Markov chains (Kemeny and Snell 1976), here I suggest a simple approximation that provides insight into the features of a phenotype that make it easy to evolve.

The idea is that a phenotype will typically be found in a “region” of the fitness landscape. Thus, the times when a population is fixed for a phenotype will come in clumps. Furthermore, if the population spends a set fraction of its time fixed for a particular phenotype, the larger these clumps are, the larger the gaps between them must be. Thus, if we can estimate the size of these clumps, we can estimate the waiting time to evolve a phenotype as the ratio between the clump size and the stationary frequency of the phenotype (Aldous 1989). Furthermore, it seems that the matrix Z should provide a way to estimate this clump size. In particular, given that the population is currently at some genotype i that codes for our phenotype of interest, we know the expected excess amount of time that the population will spend at any other genotype j that also encodes that phenotype. Thus, we should be able to develop an estimate of the clump size by appropriately summing over the entries of Z .

To develop this idea further, we will consider waiting times for a somewhat broader category of evolutionary events that encompasses not only arriving at a genotype that produces a phenotype of interest, but can include other evolutionary events of interest such as the waiting time for a specific substitution (or set of substitutions) to occur or the waiting time for a mutation to enter the population that would produce a particular genotype or phenotype. The key unifying feature of all these events is that they occur with some fixed probability given the most recent transition of the evolutionary Markov chain. More formally, if we use X_t to denote the genotype of the population at time t , for each pair i, j , let our event occur with probability $C(i, j)$ at time t given that $X_{t-1} = i$ and $X_t = j$. We can define such an event using a matrix V , where $V(i, j) = C(i, j)P(i, j)$ for some set of constants $0 \leq C(i, j) \leq 1$. Thus, for example, if we are interested in the findability of phenotypes, we can consider the event of arriving or remaining at our phenotype of interest, in which case $C(i, j)$ is 1 if j encodes our phenotype of interest and is 0 otherwise so that the event matrix V is simply the transition matrix P with the columns corresponding to all phenotypes besides our phenotype of interest set to zero. But we can also consider other events such as the event of experiencing a substitution from some specific genotype i to some other specific genotype j , in which case C is the matrix of all 0s except for its i, j th entry which is 1, and V is the matrix of all 0s except for its i, j th entry which is $P(i, j)$.

Now, given a specific choice for this matrix V , we can derive two other useful vectors. First, let g be the vector such that $g(i)$ is the probability that an event occurs in generation t given that the population is at genotype i in generation $t - 1$. Then $g = V\mathbf{1}$, where $\mathbf{1}$ is the vector of all ones, or in other words $g(i) = \sum_j V(i, j)$. Second, let c be the probability distribution describing the genotype of a

population at stationarity at time t conditional on an event having occurred at time t . Then, if we write the frequency of events at stationarity as $E_\pi g = \pi^T g$, we can also write $c^T = \pi^T V / E_\pi g$, so that $c(j) = \sum_i \pi(i) V(i, j) / \sum_i \sum_k \pi(i) V(i, k)$.

Assuming stationarity, conditioning on an event occurring means that the location of the population is distributed as c rather than π . As a result, the population is expected to spend more time at some genotypes and less time at others in the short term, relative to if we had let the population start at stationarity without having conditioned on an event occurring. These times are given by the vector $c^T Z$, whose j th entry is given by $\sum_i c(i) Z(i, j)$. Now, $g(j)$ gives the probability that an event occurs given that the population is at genotype j . Thus, the total expected number of extra events that occur given that a population is at stationarity and an event has just occurred is given by the sum over all states of the product of the extra time spent at each state $c^T Z$ and the probability of an event occurring at that state g , i.e.:

$$\sum_j \left(g(j) \sum_i c(i) Z(i, j) \right) = c^T Z g. \quad (10)$$

Adding 1 to this estimate to account for the initial event that we assume has just occurred gives us an expected clump size of $1 + c^T Z g$. Now, if the expected clump size were one, and the long-term rate of events was $E_\pi g$ the expected waiting time would be approximately $1/E_\pi g$ for rare events $E_\pi g \ll 1$. But our estimated expected clump size is $1 + c^T Z g$ times as large, and so the gaps between clumps must also be $1 + c^T Z g$ times as large, suggesting our final approximation:

$$E_\pi T_V \approx \frac{1 + c^T Z g}{E_\pi g}. \quad (11)$$

Moreover, this result is exact for any event whose corresponding matrix V is entirely composed of zeros except for a single row or column (see Appendix C).

This analysis has surprising implications for dynamically discontinuous substitutions, i.e., substitutions from a genotype to one of its mutational neighbors that is not contained in its dynamical neighborhood. As noted before, such substitutions can often be identified in multi-peaked fitness landscapes as substitutions into fitness valleys that tend to resolve away from the ancestral peak, resulting in a peak shift. More precisely, consider the occurrence of substitutions from a particular genotype i to another particular genotype j that is not in the dynamical neighborhood of i . The corresponding event matrix V has only one nonzero element, in particular $V(i, j) = P(i, j)$, (i.e., $V(m, n) = P(i, j)$ for $m = i, n = j$ and 0 otherwise). At the time of such an event, the population is assuredly at genotype j . Thus, $c^T Z g = P(i, j) Z(j, i)$. Because $j \notin \mathcal{N}(i)$, $Z(j, i) \leq 0$ and so $c^T Z g \leq 0$, i.e., given that such a substitution has just occurred we actually expect to have a deficit of additional events over

the time-scale until the chain again approaches stationarity. This results in $E_{\pi}T_V \leq \frac{1}{E_{\pi}g}$. In other words, even though discontinuous substitutions run counter to the overall dynamical tendencies of a population evolving on a fitness landscape, and are thus intuitively unlikely, they also occur in very small clumps, which results in the waiting time to observe such an event from a stationary start being shorter than would be expected based on their overall frequency of occurrence.

Another surprising implication concerns the relationship between the expected waiting time to evolve to a phenotype and the mutational robustness of that phenotype, i.e., the fraction of mutational neighbors of the constituent genotypes that code for the same phenotype (Wagner 2005, 2008). Intuitively, robust phenotypes should have large $c^T Zg$ since once a population arrives at such a phenotype many mutations produce the same phenotype, resulting in a longer clump of times during which the population displays that phenotype, and hence a longer waiting time to observe that phenotype. Conversely, phenotypes encoded by genotypes spread broadly around the genotypic space will tend to be faster to evolve. We can see this by noting that under weak mutation the clumpiness of arriving or staying at a genotype that encodes a given phenotype will be dominated by the event of staying at that phenotype (because we assume that mutations and hence substitutions occur rarely) so that for a phenotype encoded by a set of genotypes \mathcal{P} we have $g(i) \approx 0$ for $i \notin \mathcal{P}$ and $g(i) \approx 1$ for $i \in \mathcal{P}$. Thus, in this case $c^T Zg \approx (1/|\mathcal{P}|) \sum_{k,k' \in \mathcal{P}} Z(k,k')$, and this sum is reduced if some of the $Z(k,k')$ are negative, as would typically occur if the elements of \mathcal{P} are distant from each other in genotypic space.

The index of dispersion of the molecular clock for general events

We have just seen that the term $c^T Zg$ can be viewed as a measure of the extent to which events defined by the matrix V occur in a clumpy manner. Traditionally, the clumpiness of events in models of molecular evolution has been measured by the index of dispersion of the associated counting process. And indeed the formula for the asymptotic index of dispersion of an event defined by the matrix V can also be written in terms of $c^T Zg$ (Appendix D):

$$R_{\infty} = 1 - E_{\pi}g + 2c^T Zg \tag{12}$$

where R_{∞} is the limit, as t goes to infinity, of the ratio between (1) the variance in the number of events that have occurred by time t and (2) the mean number of events that have occurred by time t . Because a Poisson process has an index of dispersion of one, it is common to call a continuous-time stochastic process overdispersed if its index of dispersion is greater than one and underdispersed

if its index of dispersion is less than one. In discrete time, the analog of a Poisson process has geometric waiting times between events; because $R_{\infty} = 1 - E_{\pi}g$ for such an event, we will say that an event is overdispersed if $R_{\infty} > 1 - E_{\pi}g$ and underdispersed if $R_{\infty} < 1 - E_{\pi}g$. Thus, the sign of $c^T Zg$ determines whether the event defined by V is overdispersed or underdispersed.

While formally equivalent expressions for R_{∞} have appeared elsewhere in the literature (e.g., Iwasa 1993; Cutler 2000; Bloom et al. 2007; Raval 2007; Houchmandzadeh and Vallade 2016), they have typically been phenomenological (e.g., Cutler 2000) or difficult to interpret (e.g., Houchmandzadeh and Vallade 2016), so that the main advantage of Eq. (12) is in interpretability and intuitive understanding. In particular, the term $c^T Zg$ has an intuitive interpretation in terms of the expected number of extra events that occur given that an event has just occurred, and furthermore each of the component terms has a clear biological meaning, since c is the distribution describing the state of the population just after an event, g describes the rate at which events occur, and Z describes the neighborhood structure of the fitness landscape. That is, Eq. (12) makes it clear that the index of dispersion is essentially a measure of whether having an event occur (c) tends to put populations in regions of the fitness landscape (Z) where additional events are more or less likely to occur (g), and that the magnitude and direction of this effect is captured by the single compound term $c^T Zg$.

Much interest in the literature on the index of dispersion concerns deriving conditions where the event process is overdispersed. A sufficient condition for $R_{\infty} \geq 1 - E_{\pi}g$ met for many useful choices of V is that $DV\mathbf{1} = (DV)^T\mathbf{1}$, where D is the diagonal matrix with π down its main diagonal (i.e., $\sum_j \pi(i)V(i,j) = \sum_j \pi(j)V(j,i)$ for all i , see Appendix E for a proof). In words, this means that at stationarity, the probability of arriving at a state i when an event occurs is always equal to the probability of leaving state i just before an event occurs. Clearly this holds for the substitution process since by reversibility $\pi(i)P(i,j) = \pi(j)P(j,i)$ (see also Houchmandzadeh and Vallade 2016), but it also holds for any subset of the set of substitutions that is symmetric, for instance in codon models it holds for both the set of synonymous substitutions and the set of nonsynonymous substitutions.

On the other hand it need not hold for other asymmetric classes of substitutions such as the class of all advantageous substitutions. Indeed, our analysis so far has identified one class of substitutions that are guaranteed to be underdispersed on an individual basis, namely dynamically discontinuous substitutions to genotypes outside the dynamical neighborhood of their parent genotype. Such individual substitutions are obviously underdispersed since they have negative $c^T Zg$.

Application to a one-codon fitness landscape

To illustrate the theoretical tools we have developed so far, we will use them to analyze a one-codon fitness landscape with selection for serine and against stop codons and mutation rates based on mutation accumulation results in yeast from Lynch et al. (2008) (see McCandlish 2011, 2013, for details and previous results on this fitness landscape). It will be helpful to incorporate the visualization techniques that I have previously developed (McCandlish 2011) because the resulting visualizations integrate well with the theory described here.

We start by briefly reviewing the visualization technique. The goal is to produce a low-dimensional representation of the fitness landscape that best captures the dynamics of the evolutionary Markov chain. In particular, if we let $x_{i,t}(j)$ be the probability that a population with initial genotype i is at genotype j at generation t then we can write

$$x_{i,t}^T = x_{i,0}^T P^t = \pi^T + \sum_{k=2}^n \lambda_k^t l_k^T r_k(i), \quad (13)$$

where $\lambda_1 = 1 > \lambda_2 \geq \dots \geq \lambda_n > 0$ are the eigenvalues of P , and the l_k and r_k are, respectively, the associated left and right eigenvectors, normalized so that $(r_k)^T D r_k = 1$ and $(l_k)^T D^{-1} l_k = 1$.

The way to understand the visualization method is that for $k \geq 2$, the l_k each sum to zero and are best viewed as a set of deviations from the stationary distribution π each of which decays independently in time as λ_k^t , i.e., over time-scale $\tau_k \equiv 1/(1 - \lambda_k)$. The choice of the initial genotype i determines the weight on each of these deviations from the stationary distribution, and the values of these weights are given by the corresponding $r_k(i)$. Over time, the fastest decaying deviations corresponding to large k have essentially decayed to zero, leaving only the deviations corresponding to small k . Thus, if we plot each genotype i at, e.g., $(r_2(i), r_3(i))$, we get a low-dimensional visualization of the fitness landscape where clusters of genotypes correspond to sets of genotypes that produce similar distributions for the location of the population once the quickly decaying deviations have decayed away. The positions of the genotypes in the resulting plot thus reflect the most slowly decaying from deviations from stationarity, which capture the major barriers to diffusion in the fitness landscape. In practice we rescale the axes of the visualizations by the square-root of the associated relaxation time to plot the genotypes at $(u_2(i), u_3(i))$, where $u_k \equiv \sqrt{\tau_k} r_k$, so that the axes have units of square-root generations and squared Euclidean distance optimally approximates $E_i T_j + E_j T_i$ (see McCandlish 2011 for details and other properties of the visualizations).

Figure 2 shows the resulting visualization. Because selection is for serine, the landscape is dominated by the

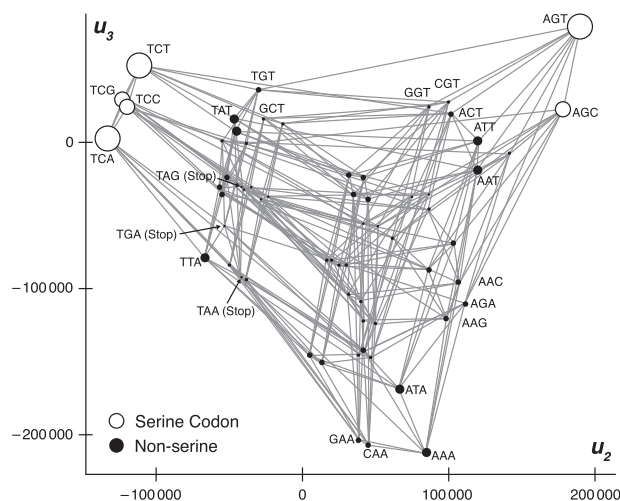


Fig. 2 Visualization of the one-codon fitness landscape using the method of (McCandlish 2011). White circles show serine codons, black circles show all other codons and circle area is proportional to stationary frequency. Edges denote point mutations. Population size $N = 3000$; fitness is 1.001 for serine codons, 0.999 for stop codons and 1.000 for all other codons, and natural selection is modeled using the exact probability of fixation for a Moran process. Mutation rates are based on Lynch et al. 2008 (see McCandlish 2011 for details)

two mutationally disconnected sets of serine codons, TCN and AGY shown as white circles. These two sets are broadly separated along the u_2 -axis because populations must cross a fitness valley to get from a TCN serine to an AGY serine or vice versa, indicating that the most slowly decaying deviation from the stationary distribution corresponds to an excess probability of being found in one group of serines or the other. The u_3 -axis also has an intuitive interpretation and separates the fit serines from genotypes several mutations away from either serine cluster.

To get a more detailed view on the evolutionary consequences of starting at any particular genotype, we can use these visualizations to plot particular rows of Z and hence display the dynamical neighborhoods of the corresponding genotypes. These visualizations are shown in Fig. 3 for three choices of initial genotype, TCA, AAA, and AGT, and emphasize the regional structure of the fitness landscape. We see that TCA and AGT are not in each other's dynamical neighborhood, while AAA and AGT are. Moreover, we can see that starting at AAA leads to an expectation of spending extra time in the neutral portion of the landscape and at the AGY serines, but tends to lead to spending less time at the TCN serines. We can also see that all of AAA's mutational neighbors are in its dynamical neighborhood, but this is not the case for either TCA or AGT, which each have one mutant neighbor outside their dynamical neighborhood. This emphasizes the fact that the probability of a peak shift after a deleterious fixation depends on the precise position of the deleterious genotype in the fitness landscape. Many deleterious substitutions will

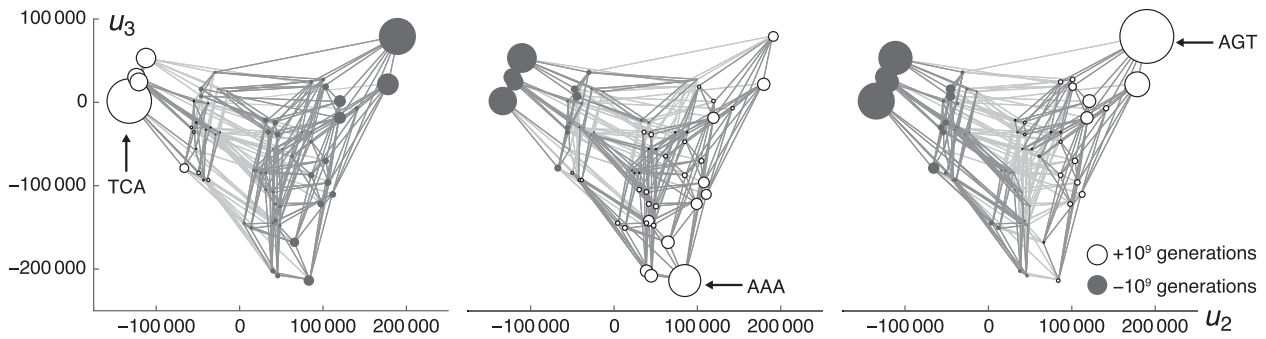


Fig. 3 The expected excess or deficit of time a population spends at each genotype, for three choices of initial genotypes indicated by the arrows. The dynamical neighborhood of the initial genotype is shown with white circles corresponding to an expected excess, while other

genotypes experience an expected deficit shown with gray circles, with the area of each circle proportional to the corresponding time; darker gray edges connect genotypes where the sign of the effect is the same

typically result in a return to the same peak in the short term, but the particular deleterious substitutions that leave the dynamical neighborhood are those that are least likely to return to the original peak.

Besides using Z to define a natural notion of neighborhood for genotypes, we also used it to estimate waiting times for particular events to occur. To investigate the accuracy of approximating $E_{\pi}T_V$ as $(1 + c^T Zg)/E_{\pi}g$, we can compare $E_{\pi}T_V$, as calculated exactly using the theory of absorbing Markov chains, with our approximation. First we will choose V to be the event of arriving or remaining at each particular amino acid or stop for the single codon fitness landscape for $N = 3000$. The approximation is always an overestimate of the true waiting time (except for the amino acids with only one codon, for which it is exact). The maximum error is 7.7% for serine which is the amino acid with the shortest expected waiting time (Fig. 4).

In (McCandlish 2013) I discussed an interesting phenomenon for the findability of genotypes in which the stationary frequency of a genotype is positively correlated with the clumpiness of the times that a population is fixed for that genotype. Because the expected waiting time to evolve a genotype from a stationary start is the ratio of these factors (Eq. (4)), this results in a decreased range of waiting times relative to the range of stationary frequencies. Here we observe the same phenomenon for the findability of amino acids. In particular, while there is a 349-fold range of equilibrium frequencies for amino acids and stop, the corresponding $E_{\pi}T_V$ vary by only 124-fold.

Besides addressing the problem of findability, we also developed formulas for the asymptotic index of dispersion R_{∞} for general events. In the case of the serine landscape we have been discussing, for the molecular clock ($V(i, j) = P(i, j)$ for $i \neq j$ and 0 otherwise) we have $R_{\infty} = 2.42$. What accounts for this overdispersion? To better understand the sources of overdispersion in this fitness landscape, it is

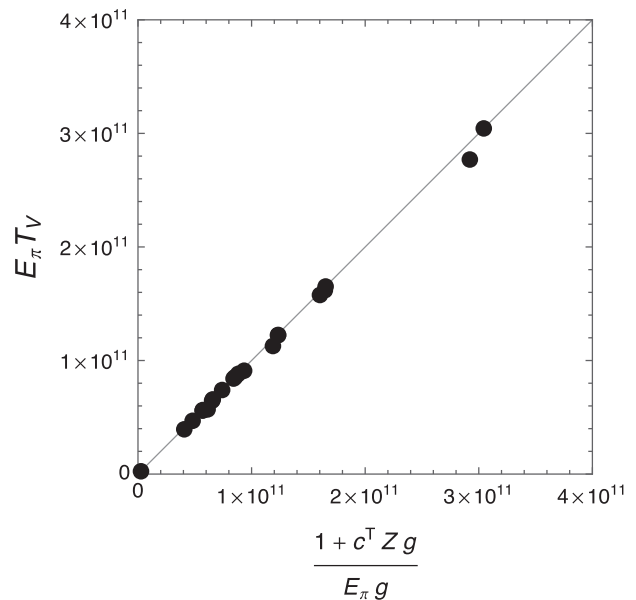


Fig. 4 $E_{\pi}T_V$, as determined numerically using the theory of absorbing Markov chains, is plotted against the approximation $(1 + c^T Zg)/E_{\pi}g$ for the single codon fitness landscape with selection for serine and $N = 3000$. Here, V corresponds to the event of arriving or remaining at a codon that codes for a particular amino acid or stop. The line $y = x$ is shown in gray; $R^2 = 0.998$

helpful to re-write Z in terms of the l_k and r_k

$$Z = \sum_{k \geq 2} \tau_k r_k l_k^T. \tag{14}$$

Using this decomposition, we can see that each deviation from equilibrium makes an independent contribution to the index of dispersion:

$$c^T Zg = \sum_{k \geq 2} c^T (\tau_k r_k l_k^T) g. \tag{15}$$

Thus, to the extent that we can identify particular axes in the visualizations with particular qualitative features of the

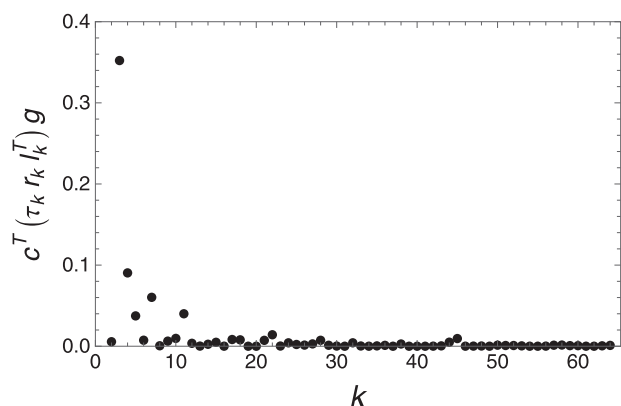


Fig. 5 The contribution to the clumpiness of the molecular clock corresponding to the deviation from equilibrium given by l_k , $c^T(\tau_k r_k l_k^T)g$, is shown for $k \geq 2$ for the one-codon fitness landscape with selection for serine and against stop codons. The sum of these contributions is $c^T Zg$. Notice the very large contribution from l_3

fitness landscape, we can quantify the manner in which these features affect the asymptotic index of dispersion.

For the case at hand, the contributions to $c^T Zg$ are shown in Fig. 5. We notice immediately that a single deviation from equilibrium, l_3 , makes the dominant contribution to the overdispersion ($c^T(\tau_3 r_3 l_3^T)g / (c^T Zg) = 0.49$). Looking back at Fig. 2, we see that the corresponding axis of the visualization, u_3 , separates the two clusters of serine codons from a large neutral plateau consisting of codons that code for other amino acids. Thus, the elevated index of dispersion largely reflects the fact that the substitution rate is low when the population is fixed at one of the two serine peaks, but experiences a higher substitution rate during the epochs that it wanders the neutral plane below. To put this another way, given that a substitution has just occurred, the population is more likely to be on the neutral plateau than on one of the peaks, and hence is likely to experience further substitutions in the short term.

Although $c^T Zg$ is indeed nonnegative when we consider all substitutions together, we have emphasized that $c^T Zg$ need not be positive for certain exceptional classes of events. To give an example where $c^T Zg$ is negative involving more than a single substitution, consider the occurrence of substitutions from codons that code for threonine to the TCN cluster of serines. Because the codons that code for threonine are of the form ACN, each of these codons is mutationally adjacent to a serine of the form TCN (the corresponding substitutions are shown in Fig. 6). For this class of substitutions, we have $R_\infty = 0.95$, so that they occur in an underdispersed manner. In this case it turns out that by far the dominant contribution is from $c^T(\tau_2 r_2 l_2^T)g$, corresponding to the axis between the two serine clusters in the visualization. This makes perfect sense, since after an ACN to TCN substitution, the population is at the TCN fitness peak, and thus has an extremely negative value for $c^T r_2$. But, in order to have another ACN to TCN substitution, the

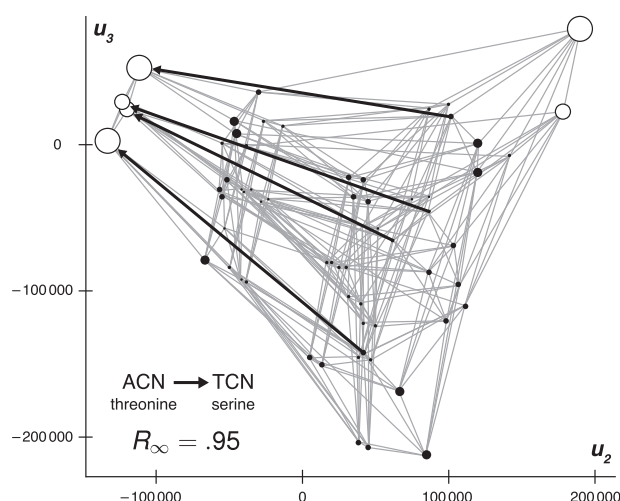


Fig. 6 A visualization of the one-codon fitness landscape with selection for serine and against stop codons. Substitutions from ACN threonines to TCN serines are shown with black arrows

population must first return to a codon that codes for threonine. This takes a very long time, (time-scale τ_2) because it requires a deleterious fixation for the population to leave the fitness peak and because the threonine codons are in a sense closer to the AGY serines, as shown by their positive values along the r_2 -axis (positive coordinates along u_2) and the dynamical neighborhood structure of the landscape (e.g., none of the threonines are in the dynamical neighborhood of TCA, while three out of four of the threonines are in the dynamical neighborhood of AGT, see Fig. 3). Thus, the fact that a substitution from ACN to TCN corresponds to a shift from the “basin of attraction” of one fitness peak to another explains why having had such a substitution decreases the expected number of these substitutions that will occur in the future.

Finally, it is interesting to consider the relationship between mutational robustness and the waiting time for an event to occur. For instance, the codons that code for serine come in two clusters, TCN and AGY, and thus exhibit substantial mutational robustness in that 26% of mutations from a serine codon lead to another serine codon. How much easier would serine be to evolve if it were coded for by six random codons, and how much harder would it be to evolve if it exhibited the even more robust arrangement of codons shown by arginine or leucine?

To answer these questions, we modify our example slightly so that each possible point mutation occurs at rate 10^{-10} per generation, and consider an ensemble of alternative scrambled genetic codes where each codon is randomly reassigned to a different amino acid while keeping the multiplicity of each amino acid constant. We then calculate the expected waiting time to evolve serine (the uniform mutation rate removes the compositional biases from the running example, so that the equilibrium frequency of

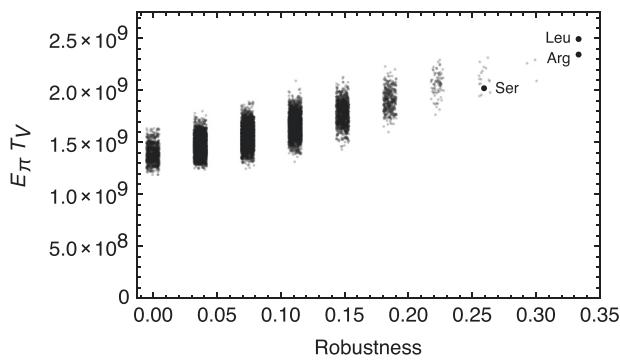


Fig. 7 Relationship between findability and robustness for phenotypes. Small dots show the expected waiting time to evolve serine versus the fraction of mutations from serine codons that produce other serine codons for 10,000 scrambled amino acid codes (horizontal jitter is added to the robustness for clarity). Pearson correlation = 0.78. Large dots show the values for serine, leucine, and arginine under the standard genetic code

serine is constant independent of which six codons are assigned to serine). The results are shown in Fig. 7. We see that the expected waiting time to evolve serine is longer for more robust arrangements of serine codons (Pearson correlation = 0.78), and that the arrangements corresponding to arginine, leucine, and serine itself have extremely high robustness compared to random arrangements (only 19 of the 10,000 random arrangements have robustness as great or greater than serine under the standard code and no arrangements were observed with robustness as high as arginine or leucine). The average waiting time to evolve a serine was 1.6×10^9 generations for random codes, and for random codes where no two serine codons are adjacent the expected waiting time was 1.4×10^9 . In contrast, the expected waiting time was 2.0×10^9 generations for the serine arrangement, 2.3×10^9 generations for the arginine arrangement and 2.5×10^9 generations for the leucine arrangement. This shows that the high degree of robustness conferred by the block structure of the standard genetic code results in substantially longer waiting times to evolve a selected phenotype than would be observed under random genotype–phenotype maps (for more on the role of robustness and evolvability in the evolution of the genetic code, see Koonin and Novozhilov 2009, 2017).

Discussion

Whereas the transition matrix of the weak-mutation Markov chain expresses the local, short-term rules governing the evolutionary dynamics, the fundamental matrix Z integrates over time to capture the expected consequences of these rules. Although as a technical matter many matrices can play the role of Z in the theory of finite state Markov chains (see Sections 7.2 and 7.3, Hunter 1983), our choice is based

on a preference for clear and intuitive interpretations: $Z(i,j)$ is the expected excess or deficit of time that a population spends at genotype j given that the population starts at genotype i relative to how much time a population would have spent at i at stationarity. Even though our presentation has been in discrete time, almost all of these results extend to continuous time with some minor modifications (Aldous and Fill 2002). Furthermore, while we have assumed that the evolutionary process takes the form of a reversible-time Markov chain, we expect the methods introduced here to work well even when strict reversibility is violated. More specifically, the results for the dynamical neighborhood depend primarily on the sign pattern of Z (and particularly that $Z(i,j)$ and $Z(j,i)$ have the same sign) rather than the exact condition $\pi(i)Z(i,j) = \pi(j)Z(j,i)$ guaranteed by reversibility, and Z itself is continuous in the entries of P so long as a mutational path exists with nonzero probability between any pair of genotypes. Moreover, Eq. (12) on the asymptotic index of dispersion applies exactly to nonreversible chains as well, demonstrating the general theoretical importance of our quantifier of clumpiness $c^T Z g$.

Based on the matrix Z , we have introduced the notion of the dynamical neighborhood of a genotype. Investigating the structure of these neighborhoods provides a powerful tool for understanding the structure of fitness landscapes as it allows us to see what the landscape looks like from the perspective of a particular genotype. This is especially useful in conjunction with visualizations based on the eigenvectors of P , as shown in Fig. 3, because the visualization of the structure of the landscape and the entries of the rows of Z contain complementary information. The visualizations are low dimensional and approximately right for all genotypes but discard information from fast time scales (i.e., from the eigenvectors not used to construct the visualizations). On the other hand Z itself is high-dimensional, but for any given row of Z we can plot it on the visualizations to get detailed information on the expected effect of starting at genotype i that is exact and integrates over all time scales.

Given an evolutionary neighborhood structure, a natural question concerns which transitions in observed evolutionary trajectories are continuous (that is remain within the neighborhood) versus those that are discontinuous (substitutions of mutational neighbors that are not dynamical neighbors). Fontana and Schuster (1998a, 1998b) and Stadler et al. (2001) previously proposed a notion of a *phenotypic* neighborhood, where the phenotypic neighbors of a phenotype are those alternative phenotypes that are produced by mutation at high rates so that they generically appear whenever the focal phenotype is present. Discontinuous phenotypic changes then corresponded to the fixation of rare new phenotypes whose discovery requires a long waiting time from the focal phenotype. (Fontana and

Schuster 1998a, 1998b) found that a single discontinuous change would often open up new directions of continuous change, resulting in a burst of substitutions. In our context, discontinuous substitutions are ones that vastly alter the fate of an evolving population, transporting it to a new region of the fitness landscape that it was previously unlikely to reach in the short term.

Expressing a related view to Fontana and Schuster (1998a, 1998b) on the role of the biased production of variation in giving direction to evolution, Schaper and Louis (2014) presented several approximations for the waiting time to produce a given phenotype based on a mean field approximation where the waiting time is inversely proportional to the average rate at which a phenotype is produced together with a multiplicative term accounting for various population-genetic factors. Translating our current results into the case of a population evolving on a network of high-fitness genotypes, the results here based on $c^T Zg$ provide a more nuanced approximation by taking into account the structure of the network and in a way that automatically adjusts for population-genetic parameters such as sequence length and mutation rate.

Previous theoretical treatments of the index of dispersion of the molecular clock for stationary evolution under weak mutation have typically emphasized either the structure of high fitness genotype networks with varying neutral mutation rates (Bloom et al. 2007; Raval 2007; Bastolla et al. 2007), or treated the problem for uncorrelated fitness landscapes, where clumpiness in the substitution processes arises due to a pattern of stasis when the population is at high fitness punctuated by rapid substitutions during the period when the population is at low fitness (Iwasa 1993; McCandlish et al. 2014). Our analysis here can accommodate both these factors, and indeed we see both at play in our serine example: overdispersion is driven by having different substitution rates in different regions of the fitness landscape, but this difference in substitution rate is driven by the difference between the lower substitution rate at the fitness peaks and the higher rate on the neutral plain below. More generally, because the most slowly decaying deviations from stationarity make the largest contributions to Z via their large τ_k , we expect $c^T Zg$ to be largest when g covaries with the r_k for small r_k , which interestingly implies that $c^T Zg$ will be maximized when g appears to change monotonically across our visualizations.

The theory presented here is focused on the expected long-term dynamics of the weak-mutation Markov chain. It therefore assumes that the selection coefficients, mutation rates and population size all remain constant over the relevant time scales. Moreover, the methods employed here are based on modeling the evolutionary process and therefore assume that all relevant sites are included the model. This contrasts with some techniques

such as the examination of rank orders of genotypes (Weinreich et al. 2005) that can usefully be applied to sub-landscapes or in the presence of incomplete data. As a result, empirical applications of the techniques described here must largely be limited to computational models and genomic segments whose action is conserved and where the size of the genotypic space is small enough to allow exhaustive measurements of the fitness for each possible genotype. Promising applications that fit these criteria include codon models for sites with conserved amino acid preferences (Bazykin 2015), conserved transcription factor binding sites (Aguilar-Rodríguez et al. 2017), and splice sites. Another reason that the number of sites involved in the model must be limited is that we assume that the total mutation rate is small. This is necessary in order to model the evolution of the population as a Markov chain on the set of genotypes, and we have leveraged the simplicity and tractability of finite state Markov chains to produce a relatively simple and intelligible mathematical theory. Although some analytical progress has been made in modeling the evolution of finite polymorphic populations on highly simplified fitness landscapes (e.g., Weissman et al. 2009; Goyal et al. 2012; Jain and Krug 2007), we still lack a general theory that would apply to finite polymorphic populations on complex fitness landscapes, making it difficult to address the types of questions we have considered here in that more realistic context.

Data archiving

A mathematica file containing the code used to produce the figures and conduct the numerical calculations has been deposited in Dryad, <https://doi.org/10.5061/dryad.9b5q2pv>.

Acknowledgments Many thanks to Arjan de Visser, Santiago Elena, Inês Fragata, and Sebastian Matuszewski for organizing this theme issue on fitness landscapes, and to two anonymous reviewers for many helpful comments on the manuscript.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Appendix

(A) The dynamical neighborhood of i is connected and includes i

We want to show that for a fixed i , the set of genotypes j such that $Z(i,j) > 0$ is connected, in the sense that there is a path l_1, \dots, l_k such that $Z(i, l_1), \dots, Z(i, l_k) > 0$ and $P(j, l_1)P(l_1, l_2) \cdots P(l_k, i) > 0$. In fact, we will show that for a

fixed choice of genotype i and constant a , the set of genotypes j with $E_j T_i < a$ is connected. Because $E_j T_j = (Z(j, j) - Z(i, j))/\pi(j)$ and $E_\pi T_j = Z(j, j)/\pi(j)$, this reduces to the previous condition for $a = E_\pi T_i$

Suppose to the contrary that the set A of genotypes j such that $E_j T_i < a$ is not connected. Then let A' be the component of A that contains i (i.e., let A' be the maximal connected subset of A containing i), and let A'' be some other component of A . Now, let k be a genotype in A'' and consider $E_k T_i$. We can write $E_k T_i = ET^{(1)} + ET^{(2)}$ where $ET^{(1)}$ is the expected time for a population that begins at k to first arrive at a genotype that is not in A and then $ET^{(2)}$ is the expected remaining waiting time for the population to first arrive at i . Of course, $ET^{(1)} > 0$ since the population begins at genotype $k \in A$. Furthermore, $ET^{(2)} \geq a$ since at $T^{(1)}$ the population is no longer in A . Thus we have:

$$E_k T_i = ET^{(1)} + ET^{(2)} > ET^{(2)} > a \tag{A1}$$

which contradicts our original assumption that $k \in A$. Thus, the set of genotypes j such that $E_j T_i < a$ is connected, as is the set of genotype j such that $Z(i, j) > 0$.

We also need to show that i is in its own dynamical neighborhood, i.e., $Z(i, i) > 0$. Let $Z_t(i, j) \equiv \sum_{m=0}^{t-1} (P^m(i, j) - \pi(j))$. From Eq. (13), we can write $Z_t(i, i)$ as:

$$Z_t(i, i) = \sum_{k \geq 2} \frac{1 - \lambda_k^t}{1 - \lambda_k} r_k(i) l_k(i) = \sum_{k \geq 2} \frac{1 - \lambda_k^t}{1 - \lambda_k} q_k(i)^2 \tag{A2}$$

which is clearly nonnegative and increasing with t . Now, $Z_1(i, i) = 1 - \pi(i)$, which is positive except in the trivial case, which we choose to exclude, of a fitness landscape with only one genotype. Because Z_t is increasing with t , it follows that $Z(i, i) = Z_\infty(i, i) \geq 1 - \pi > 0$, so that i is in its own dynamical neighborhood.

(B) Analytical formula for Z in the neutral case

In the neutral case, we have $P = I - \mu L$, where L is the Laplacian matrix of the l -dimensional Hamming graph on $|\mathcal{A}|$ letters. Because the eigendecomposition of L is known and can be expressed in terms of the Krawtchouk polynomials (Stadler and Happel 1999), we can plug this eigendecomposition into Eq. (14) to get:

$$Z(i, j) = \sum_{k=1}^l \frac{1}{\mu k |\mathcal{A}|} \sum_{q=0}^{\min\{k, d\}} \frac{(-1)^q (|\mathcal{A}| - 1)^{k-q}}{|\mathcal{A}|^q} \binom{d}{q} \binom{l-d}{k-q}, \tag{A3}$$

where d is the Hamming distance between sequence i and sequence j .

(C) Proof that the approximation in Eq. (11) is exact for any event that either always ends or always begins in a single genotype

We want to show that $E_\pi T_V = (1 + c^T Z g)/E_\pi g$ for any V that is all zeros except that either (A) a single row of V contains nonzero entries (the event always begins in a particular genotype i) or (B) a single column of V contains nonzero entries (an event always ends in a single genotype i).

First we will consider the expected waiting time until an event that always begins in genotype i . The expected time until such an event can be viewed as the sum of two waiting times: the waiting time to first reach i , and then the waiting time for the event to occur when starting at i . Thus, we have

$$E_\pi T_V = E_\pi T_i + E_i T_V. \tag{A4}$$

Now $E_\pi T_i$ is simply $Z(i, i)/\pi(i)$, and so what we want to know is $E_i T_V$.

Given that the population is at i , the total number of times the population will be at i before T_V is geometrically distributed, where the success probability is $g(i)$. Thus, on average there are $(1 - g(i))/g(i)$ failures before the first success. After a failure, the population is distributed as:

$$\frac{(P(i, 1) \dots P(i, n)) - (V(i, 1) \dots V(i, n))}{1 - g(i)} \tag{A5}$$

and one step has occurred. So after a failure, the expected time to return to i is:

$$1 + \sum_j \frac{P(i, j) - V(i, j)}{1 - g(i)} E_j T_i = 1 + \frac{Z(i, i)}{\pi(i)} - \frac{\sum_j P(i, j) Z(j, i)}{(1 - g(i)) \pi(i)} + \frac{\sum_j V(i, j) Z(j, i)}{(1 - g(i)) \pi(i)} = 1 + \frac{Z(i, i)}{\pi(i)} - \frac{Z(i, i) + \pi(i) - 1}{(1 - g(i)) \pi(i)} + \frac{c^T Z g}{(1 - g(i)) \pi(i)} = \frac{1 - \pi(i) g(i) - Z(i, i) g(i) + c^T Z g}{(1 - g(i)) \pi(i)} \tag{A6}$$

In addition, there is exactly one success, which takes exactly one time step. Thus, multiplying the duration for each failure by the expected number of failures and adding one for the final success, we have:

$$E_i T_V = \frac{1 - \pi(i) g(i) - Z(i, i) g(i) + c^T Z g}{\pi(i) g(i)} + 1 = \frac{1 - Z(i, i) g(i) + c^T Z g}{\pi(i) g(i)} \tag{A7}$$

Plugging Eq. (A7) into Eq. (A4) gives us:

$$E_{\pi}T_V = \frac{Z(i,i)}{\pi(i)} + \frac{1-Z(i,i)g(i)+c^T Zg}{\pi(i)g(i)} \tag{A8}$$

$$= \frac{1+c^T Zg}{E_{\pi}g}$$

as required.

Next, let us consider an event V that always ends in genotype i . We first treat the special case that an event occurs each time the population transitions into some specific state i , (that is, we let $V(k,j) = P(k,j)$ for $j = i$ and 0 otherwise). Since the population is always at genotype i when an event occurs, c is the vector with a 1 in position i and 0 elsewhere. Furthermore, since an event happens with each transition into i , g is simply the i th column of P .

Now, we turn to calculating $E_{\pi}T_V$. Because an event can only occur on a transition into state i , no event can occur at $t = 0$, and so the first possible time for an event to occur is $t = 1$. Moreover, because the initial genotype is drawn from the stationary distribution π at $t = 0$ the location of the population is still described by the stationary distribution at $t = 1$ and so we have $E_{\pi}T_V = 1 + E_{\pi}T_i$, i.e.:

$$E_{\pi}T_V = \frac{Z(i, i)}{\pi(i)} + 1. \tag{A9}$$

To show that this expression is equal to $(1 + c^T Zg)/E_{\pi}g$, we will need the following identity for $Z(i,i)$:

$$Z(i, i) = \sum_{t=0}^{\infty} (P^t(i, i) - \pi(i)) \tag{A10}$$

$$= 1 - \pi(i) + \sum_{t=1}^{\infty} (P^t(i, i) - \pi(i)) \tag{A11}$$

$$= 1 - \pi(i) + \sum_{t=1}^{\infty} \left(\sum_k P^{t-1}(i, k)P(k, i) - \sum_k \pi(k)P(k, i) \right) \tag{A12}$$

$$= 1 - \pi(i) + \sum_k \left(\left(\sum_{t=1}^{\infty} P^{t-1}(i, k) - \pi(k) \right) P(k, i) \right) \tag{A13}$$

$$= 1 - \pi(i) + \sum_k Z(i, k)P(k, i), \tag{A14}$$

where we have used the fact $\pi(i) = \sum_k \pi(k)P(k, i)$ and the definition of Z . Noting that for our choice of V we have $c^T Zg = \sum_k Z(i, k)P(k, i)$ and $E_{\pi}g = \pi(i)$, we can substitute the above expression for $Z(i,i)$ into Eq. (A9) to obtain

$$E_{\pi}T_V = \frac{1 - \pi(i) + \sum_k Z(i, k)P(k, i)}{\pi(i)} + 1 \tag{A15}$$

$$= \frac{1 - E_{\pi}g + c^T Zg}{E_{\pi}g} + 1 \tag{A16}$$

$$= \frac{1 + c^T Zg}{E_{\pi}g} \tag{A17}$$

as required.

To extend this result to events V where events only occur upon transition to state i but do not occur on every transition to i , we consider an augmented Markov chain with transition matrix P^* and event V^* chosen so that the distribution of waiting times for V^* is identical to the distribution for V , but we add an auxiliary state $n + 1$ so that events only occur on transitions to state $n + 1$ and occur on each such transition.

In particular, we define P^* as

$$P^*(j, k) = \begin{cases} P(j, k) & \text{for } j \neq i, j \neq n + 1, k \neq i, k \neq n + 1 \\ P(i, k) & \text{for } j = i \text{ or } j = n + 1, k \neq i, k \neq n + 1 \\ P(j, i) - V(j, i) & \text{for } k = i \\ V(j, i) & \text{for } k = n + 1, \end{cases} \tag{A18}$$

and V^* as

$$V^*(j, k) = \begin{cases} V(j, i) & \text{for } k = n + 1, j \neq n + 1 \\ V(i, i) & \text{for } k = n + 1, j = n + 1 \\ 0 & \text{otherwise.} \end{cases} \tag{A19}$$

Then P^* with event V^* has dynamics identical to P with event V except that we have split the state i into two states so that the population goes to $n + 1$ instead of i when an event occurs. Defining π^*, c^*, Z^* and g^* based on P^* and V^* , we then have $(c^*)^T Z^* g^* = c^T Zg$, $E_{\pi^*} g^* = E_{\pi}g$, and $E_{\pi^*} T_{V^*} = E_{\pi} T_V$. Thus,

$$E_{\pi} T_V = E_{\pi^*} T_{V^*} \tag{A20}$$

$$= \frac{1 + (c^*)^T Z^* g^*}{E_{\pi^*} g^*} \tag{A21}$$

$$= \frac{1 + c^T Zg}{E_{\pi}g}, \tag{A22}$$

as required.

(D) The index of dispersion for generalized events

We will proceed in somewhat more generality than the main text and consider a general ergodic transition matrix P with events defined by a matrix V as before. To write down a formula for R_{∞} , it will be helpful to write down formulas for $E_{\pi}I_t$, which is the probability that an event occurs at time

t given that the initial genotype is a draw from the equilibrium distribution, and $E_\pi(I_{t_1}I_{t_2})$, which is the probability that an event occurs at both time t_1 and time t_2 given that the initial genotype is a draw from the equilibrium distribution. Of course,

$$E_\pi I_t = \pi^T P^{t-1} g = \pi^T g = E_\pi g \tag{A23}$$

and for $t_1 < t_2$

$$\begin{aligned} E_\pi(I_{t_1}I_{t_2}) &= \pi^T P^{t_1-1} V P^{t_2-t_1-1} g \\ &= \pi^T V P^{t_2-t_1-1} g \\ &= (E_\pi g)(c P^{t_2-t_1-1} g) \end{aligned} \tag{A24}$$

since the probability that an event occurs at time t_1 is $E_\pi g$, and given that an event occurs at t_1 the distribution of the location of the population is given by c .

Writing S_t for the number of events that have occurred through time t , using Eq. (A23) we have:

$$\begin{aligned} E_\pi S_t &= E_\pi \left(\sum_{m=1}^t I_m \right) \\ &= \sum_{m=1}^t E_\pi I_m \\ &= t E_\pi g. \end{aligned} \tag{A25}$$

Furthermore using the general formula for the variance of a sum of random variables and noting that $E_\pi(I_{m_1}I_{m_2}) = E_\pi(I_{m_2}I_{m_1})$ depends only on $|m_2 - m_1|$ and $E_\pi(I_m)^2 = E_\pi I_m$, we have:

$$\begin{aligned} \text{Var}_\pi S_t &= \text{Var}_\pi \left(\sum_{m=1}^t I_m \right) \\ &= \sum_{m=1}^t \text{Var}_\pi I_m + 2 \sum_{m_1=1}^{t-1} \sum_{m_2=m_1+1}^t \text{Cov}_\pi(I_{m_1}, I_{m_2}) \\ &= \sum_{m=1}^t (E_\pi I_m - (E_\pi I_m)^2) \\ &\quad + 2 \sum_{m_1=1}^{t-1} \sum_{m_2=m_1+1}^t (E_\pi(I_{m_1}I_{m_2}) - E_\pi I_{m_1} E_\pi I_{m_2}) \\ &= t E_\pi g (1 - E_\pi g) + 2(E_\pi g) \sum_{m=1}^{t-1} (t-m)(c^T P^{m-1} g - E_\pi g) \\ &= E_\pi S_t \left(1 - E_\pi g + 2 \sum_{m=1}^{t-1} \left(1 - \frac{m}{t} \right) (c^T P^{m-1} g - E_\pi g) \right). \end{aligned} \tag{A26}$$

Concentrating on the last term, we have $E_\pi g = \pi^T g = c^T \Pi g$, where $\Pi \equiv \mathbf{1}\pi^T$ is a projection matrix corresponding to the stationary distribution of P . Thus:

$$\begin{aligned} &\sum_{m=1}^{t-1} \left(1 - \frac{m}{t} \right) (c^T P^{m-1} g - E_\pi g) \\ &= c^T \left(\sum_{m=1}^{t-1} \left(1 - \frac{m}{t} \right) (P^{m-1} - \Pi) \right) g. \end{aligned} \tag{A27}$$

Pulling apart the sum in parentheses, we have

$$\begin{aligned} &\sum_{m=1}^{t-1} \left(1 - \frac{m}{t} \right) (P^{m-1} - \Pi) \\ &= \left(\sum_{m=1}^{t-1} (P^{m-1} - \Pi) \right) - \left(\sum_{m=1}^{t-1} \frac{m}{t} \cdot (P^{m-1} - \Pi) \right). \end{aligned} \tag{A28}$$

Furthermore,

$$\lim_{t \rightarrow \infty} \sum_{m=1}^{t-1} (P^{m-1} - \Pi) = Z \tag{A29}$$

by the definition of Z and

$$\lim_{t \rightarrow \infty} \left(\sum_{m=1}^{t-1} \frac{m}{t} \cdot (P^{m-1} - \Pi) \right) = 0 \tag{A30}$$

since P^t converges at least geometrically quickly to Π entry-wise for sufficiently large t under Perron–Frobenius theory (see, e.g., Theorem 1.2 of Seneta 2006) using the facts that P is primitive and that the Perron eigenvalue of P is 1.

Putting this all together, we have

$$\begin{aligned} R_\infty \equiv \lim_{t \rightarrow \infty} \frac{\text{Var}_\pi S_t}{E_\pi S_t} &= \lim_{t \rightarrow \infty} \left(1 - E_\pi g + 2 \sum_{m=1}^{t-1} \left(1 - \frac{m}{t} \right) (c^T P^{m-1} g - E_\pi g) \right) \\ &= 1 - E_\pi g + 2c^T Z g \end{aligned} \tag{A31}$$

as required.

(E) Sufficient condition for overdispersion

If $DV\mathbf{1} = (DV)^T \mathbf{1}$ then $c = Dg/E_\pi g$. This means that we can write:

$$\begin{aligned} c^T (\tau_k r_k l_k^T) g &= \tau_k g^T D (r_k l_k^T) g / (E_\pi g) \\ &= \tau_k g^T (l_k l_k^T) g / (E_\pi g) \\ &= \tau_k (l_k^T g)^2 / (E_\pi g) \end{aligned} \tag{A32}$$

which is nonnegative because the numerator is nonnegative and the denominator is positive. Thus, if $DV\mathbf{1} = (DV)^T \mathbf{1}$, $c^T Z g$ is also nonnegative, by Eq. (15).

References

- Aguilar-Rodríguez J, Payne JL, Wagner A (2017) A thousand empirical adaptive landscapes and their navigability. *Nat Ecol Evol* 1:0045
- Aldous D (1989) Probability approximations via the Poisson clumping heuristic. Springer-Verlag, New York, NY
- Aldous D, Fill JA (2002) Reversible markov chains and random walks on graphs. Unfinished monograph, recompiled 2014. Available at <http://www.stat.berkeley.edu/aldous/RWG/book.html>
- Bank C, Matuszewski S, Hietpas RT, Jensen JD (2016) On the (un) predictability of a large intragenic fitness landscape. *Proc Natl Acad Sci* 113(49):14085–14090
- Bastolla U, Porto M, Roman HE, Vendruscolo M (2007) The structurally constrained neutral model of protein evolution. In: Ugo B, Markus P, H Eduardo R, Michele V (eds) Structural approaches to sequence evolution, biological and medical physics, biomedical engineering. Springer, Berlin, Heidelberg, p 75–112
- Bazykin GA (2015) Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins. *Biol Lett* 11(10):20150315
- Beerenwinkel N, Pachter L, Sturmfels B, Elena S, Lenski R (2007) Analysis of epistatic interactions and fitness landscapes using a new geometric approach. *BMC Evol Biol* 7(1):60
- Berg J, Willmann S, Lassig M (2004) Adaptive evolution of transcription factor binding sites. *BMC Evol Biol* 4(1):42
- Bloom JD, Raval A, Wilke CO (2007) Thermodynamics of neutral protein evolution. *Genetics* 175(1):255–266
- Bornberg-Bauer E, Chan HS (1999) Modeling evolutionary landscapes: mutational stability, topology, and superfunnels in sequence space. *Proc Natl Acad Sci USA* 96(19):10689–10694
- Carneiro M, Hartl DL (2010) Adaptive landscapes and protein evolution. *Proc Natl Acad Sci USA* 107(suppl 1):S1747–S1751
- Cowperthwaite MC, Meyers LA (2007) How mutational networks shape evolution: lessons from RNA models. *Annu Rev Ecol, Evol Syst* 38(1):203–230
- Crona K, Gavryushkin A, Greene D, Beerenwinkel N (2017) Inferring genetic interactions from comparative fitness data. *eLife* 6: e28629. <https://doi.org/10.7554/eLife.28629>
- Cutler DJ (2000) Understanding the overdispersed molecular clock. *Genetics* 154(3):1403–1417
- de Visser JAGM, Park SC, Krug J (2009) Exploring the effect of sex on empirical fitness landscapes. *Am Nat* 174(Suppl 1):S15–S30
- Ferretti L, Schmiegel B, Weinreich D, Yamauchi A, Kobayashi Y, Tajima F, Achaz G (2016) Measuring epistasis in fitness landscapes: the correlation of fitness effects of mutations. *J Theor Biol* 396:132–143
- Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, London
- Fontana W (2002) Modelling “evo-devo” with RNA. *Bioessays* 24(12):1164–1177
- Fontana W, Schuster P (1998a) Continuity in evolution: on the nature of transitions. *Science* 280(5368):1451–1455
- Fontana W, Schuster P (1998b) Shaping space: the possible and the attainable in RNA genotype–phenotype mapping. *J Theor Biol* 194(4):491–515
- Fontana W, Stadler PF, Bornberg-Bauer EG, Griesmacher T, Hofacker IL, Tacker M, Tarazona P, Weinberger ED, Schuster P (1993) RNA folding and combinatory landscapes. *Phys Rev E Stat Nonlinear Soft Matter Phys* 47:2083–2099
- Gavrilets S (1997) Evolution and speciation on holey adaptive landscapes. *Trends Ecol Evol* 12(8):307–312
- Gavrilets S (2003) Evolution and speciation in a hyperspace: The roles of neutrality, selection, mutation and random drift. In: Crutchfield J, Schuster P (eds) Evolutionary dynamics: exploring the interplay of selection, accident, neutrality and function. Oxford University Press, Oxford, pp 135–162
- Gavrilets S (2004) Fitness landscapes and the origin of species. Princeton University Press, Princeton, NJ
- Gavrilets S, Gravner J (1997) Percolation on the fitness hypercube and the evolution of reproductive isolation. *J Theor Biol* 184(1):51–64
- Gillespie JH (1983) A simple stochastic gene substitution model. *Theor Popul Biol* 23(2):202–215
- Gillespie JH (1984) Molecular evolution over the mutational landscapes. *Evolution* 38(5):1116–1129
- Gillespie JH (1991) The causes of molecular evolution. Oxford University Press, New York, NY
- Goyal S, Balick DJ, Jerison ER, Neher RA, Shraiman BI, Desai MM (2012) Dynamic mutation–selection balance as an evolutionary attractor. *Genetics* 191(4):1309–1319
- Gravner J, Pitman D, Gavrilets S (2007) Percolation on fitness landscapes: effects of correlation, phenotype, and incompatibilities. *J Theor Biol* 248(4):627–645
- Gruner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker I, Stadler P, Schuster P (1996) Analysis of RNA sequence structure maps by exhaustive enumeration I. Neutral networks. *Chem Mon* 127(4):355–374
- Hartl DL, Taubes CH (1996) Compensatory nearly neutral mutations: selection without adaptation. *J Theor Biol* 182(3):303–309
- Hartl DL, Dykhuizen DE, Dean AM (1985) Limits of adaptation: the evolution of selective neutrality. *Genetics* 111(3):655–674
- Hinkley T, Martins J, Chappey C, Haddad M, Stawiski E, Whitcomb JM, Petropoulos CJ, Bonhoeffer S (2011) A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat Genet* 43(5):487–489
- Houchmandzadeh B, Vallade M (2016) A simple, general result for the variance of substitution number in molecular evolution. *Mol Biol Evol* 33(7):1858–1869
- Hunter JJ (1983) Mathematical techniques of applied probability. Discrete time models: techniques and applications, Vol 2. Academic Press, New York, NY
- Iwasa Y (1988) Free fitness that always increases in evolution. *J Theor Biol* 135(3):265–281
- Iwasa Y (1993) Overdispersed molecular evolution in constant environments. *J Theor Biol* 164(3):373–393
- Iwasa Y, Michor F, Nowak MA (2004) Stochastic tunnels in evolutionary dynamics. *Genetics* 166(3):1571–1579
- Jain K, Krug J (2007) Deterministic and stochastic regimes of asexual evolution on rugged fitness landscapes. *Genetics* 175(3):1275–1288
- Jain K, Seetharaman S (2011) Multiple adaptive substitutions during evolution in novel environments. *Genetics* 189(3):1029–1043
- Kauffman S, Levin S (1987) Towards a general theory of adaptive walks on rugged landscapes. *J Theor Biol* 128(1):11–45
- Kauffman SA (1993) The origins of order: self organization and selection in evolution. Oxford University Press, New York, NY
- Kelly FP (1979) Reversibility and stochastic networks. Wiley, Chichester
- Kemeny JG, Snell JL (1976) Finite Markov chains. Springer, New York, NY
- Kielson J (1979) Markov chain models—rarity and exponentiality. Springer-Verlag, New York, NY
- Kimura M (1962) On the probability of fixation of mutant genes in a population. *Genetics* 47(6):713–719
- Kondrashov FA, Kondrashov AS (2001) Multidimensional epistasis and the disadvantage of sex. *Proc Natl Acad Sci USA* 98(21):12089–12092
- Koonin EV, Novozhilov AS (2009) Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* 61(2):99–111

- Koonin EV, Novozhilov AS (2017) Origin and evolution of the universal genetic code. *Annu Rev Genet* 51:45–62
- Lobkovsky AE, Wolf YI, Koonin EV (2011) Predictability of evolutionary trajectories in fitness landscapes. *PLoS Comput Biol* 7(12):e1002302
- Lozovsky ER, Chookajorn T, Brown KM, Imwong M, Shaw PJ, Kamchonwongpaisan S, Neafsey DE, Weinreich DM, Hartl DL (2009) Stepwise acquisition of pyrimethamine resistance in the malaria parasite. *Proc Natl Acad Sci USA* 106(29):12025–12030
- Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, Thomas WK (2008) A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci USA* 105(27):9272–9277
- Macken CA, Perelson AS (1989) Protein evolution on rugged landscapes. *Proc Natl Acad Sci USA* 86(16):6191–6195
- McCandlish DM (2011) Visualizing fitness landscapes. *Evolution* 65(6):1544–1558
- McCandlish DM (2013) On the findability of genotypes. *Evolution* 67:2592–2603
- McCandlish DM, Stoltzfus A (2014) Modeling evolution using the probability of fixation: History and implications. *Q Rev Biol* 89(3):225–252
- McCandlish DM, Epstein CL, Plotkin JB (2014) The inevitability of unconditionally deleterious substitutions during adaptation. *Evolution* 68:1351–1364
- McCandlish DM, Epstein CL, Plotkin JB (2015a) Formal properties of the probability of fixation: identities, inequalities and approximations. *Theor Popul Biol* 99:98–113
- McCandlish DM, Otwinowski J, Plotkin JB (2015b) Detecting epistasis from an ensemble of adapting populations. *Evolution* 69(9):2359–2370
- McCandlish DM, Shah P, Plotkin JB (2016) Epistasis and the dynamics of reversion in molecular evolution. *Genetics* 203(3):1335–1351
- Meyer Jr CD (1975) The role of the group generalized inverse in the theory of finite Markov chains. *SIAM Rev* 17(3):443–464
- Neidhart J, Szendro IG, Krug J (2013) Exact results for amplitude spectra of fitness landscapes. *J Theor Biol* 332:218–227
- Orr HA (1998) The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution* 52(4):935–949
- Østman B, Hintze A, Adami C (2011) Impact of epistasis and pleiotropy on evolutionary adaptation. *Proc R Soc Lond B Biol Sci* <https://doi.org/10.1098/rspb.2011.0870>
- Phillips PC (2008) Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* 9(11):855–867
- Pitt JN, Ferré-D'Amaré AR (2010) Rapid construction of empirical RNA fitness landscapes. *Science* 330(6002):376–379
- Poelwijk FJ, Krishna V, Ranganathan R (2016) The context-dependence of mutations: a linkage of formalisms. *PLoS Comput Biol* 12(6):e1004771
- Raval A (2007) Molecular clock on a neutral network. *Phys Rev Lett* 99(13):138104
- Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, Ivankov DN, Bozhanova NG, Baranov MS, Soylemez O et al. (2016) Local fitness landscape of the green fluorescent protein. *Nature* 533(7603):397–401
- Schaper S, Louis AA (2014) The arrival of the frequent: how bias in genotype–phenotype maps can steer populations to local optima. *PloS ONE* 9(2):e86635
- Seetharaman S, Jain K (2014) Length of adaptive walk on uncorrelated and correlated fitness landscapes. *Phys Rev E* 90(3):032703
- Sella G, Hirsh AE (2005) The application of statistical physics to evolutionary biology. *Proc Natl Acad Sci USA* 102(27):9541–9546
- Seneta E (2006) *Nonnegative matrices and Markov chains*. Springer Science & Business Media, New York
- Stadler BM, Stadler PF, Wagner GP, Fontana W (2001) The topology of the possible: formal spaces underlying patterns of evolutionary change. *J Theor Biol* 213(2):241–274
- Stadler PF (2003) Spectral landscape theory. In: Crutchfield J, Schuster P (eds) *Evolutionary dynamics: exploring the interplay of selection, accident, neutrality and function*. Oxford University Press, Oxford, pp 231–271
- Stadler PF, Happel R (1999) Random field models for fitness landscapes. *J Math Biol* 38(5):435–478
- Szendro IG, Schenk MF, Franke J, Krug J, de Visser JAG (2013) Quantitative analyses of empirical fitness landscapes. *J Stat Mech Theory Exp* 2013(01):P01005
- Tenaillon O (2014) The utility of Fisher's geometric model in evolutionary genetics. *Annu Rev Ecol, Evol Syst* 45:179–201
- van Nimwegen E, Crutchfield JP (2000) Metastable evolutionary dynamics: crossing fitness barriers or escaping via neutral paths? *Bull Math Biol* 62(5):799–848
- de Visser JAG, Krug J (2014) Empirical fitness landscapes and the predictability of evolution. *Nat Rev Genet* 15:480–490
- Wagner A (1996) Does evolutionary plasticity evolve? *Evolution* 50(3):1008–1023
- Wagner A (2005) *Robustness and evolvability in living systems*. Princeton University Press, Princeton, NJ
- Wagner A (2008) Robustness and evolvability: a paradox resolved. *Proc R Soc Lond B Biol Sci* 275(1630):91–100
- Wagner A (2011) *The origins of evolutionary innovations: a theory of transformative change in living systems*. Oxford University Press, Oxford
- Weinberger E (1990) Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biol Cybern* 63:325–336
- Weinberger E (1991) Fourier and Taylor series on fitness landscapes. *Biol Cybern* 65:321–330
- Weinreich DM (2005) The rank ordering of genotypic fitness values predicts genetic constraint on natural selection on landscapes lacking sign epistasis. *Genetics* 171(3):1397–1405
- Weinreich DM, Chao L (2005) Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. *Evolution* 59(6):1175–1182
- Weinreich DM, Watson RA, Chao L (2005) Perspective: sign epistasis and genetic constraint on evolutionary trajectories. *Evolution* 59(6):1165–1174
- Weinreich DM, Delaney NF, DePristo MA, Hartl DL (2006) Darwinian evolution can follow only very few mutational paths to fitter proteins. *Science* 312(5770):111–114
- Weinreich DM, Lan Y, Wylie CS, Heckendorn RB (2013) Should evolutionary geneticists worry about higher-order epistasis? *Curr Opin Genet Dev* 23(6):700–707
- Weissman DB, Desai MM, Fisher DS, Feldman MW (2009) The rate at which asexual populations cross fitness valleys. *Theor Popul Biol* 75(4):286–300
- Wolf JB, Brodie ED, Wade MJ (2000) *Epistasis and the evolutionary process*. Oxford University Press, Oxford
- Wright S (1931) Evolution in Mendelian populations. *Genetics* 16(2):97–159
- Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. In: *Proceedings of the Sixth International Congress of Genetics*, pp 356–366
- Wu N, Dai L, Olson CA, Lloyd-Smith JO, Sun R (2016) Adaptation in protein fitness landscapes is facilitated by indirect paths. *Elife* 5:e16965