Multiple Adaptive Substitutions During Evolution in Novel Environments

Kavita Jain^{*,†,1} and Sarada Seetharaman^{*}

^{*}Theoretical Sciences Unit and [†]Evolutionary and Organismal Biology Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bangalore 560064, India

ABSTRACT We consider an asexual population under strong selection–weak mutation conditions evolving on rugged fitness landscapes with many local fitness peaks. Unlike the previous studies in which the initial fitness of the population is assumed to be high, here we start the adaptation process with a low fitness corresponding to a population in a stressful novel environment. For generic fitness distributions, using an analytic argument we find that the average number of steps to a local optimum varies logarithmically with the genotype sequence length and increases as the correlations among genotypic fitnesses increase. When the fitnesses are exponentially or uniformly distributed, using an evolution equation for the distribution of population fitness, we analytically calculate the fitness distribution of fixed beneficial mutations and the walk length distribution.

ADAPTATION is an evolutionary process during which a population improves its fitness by accumulating beneficial mutations. A population of genotypic sequences produces a suite of mutants and if better mutants become available, a maladapted population may acquire one of the beneficial mutations provided it does not get lost due to genetic drift. The fitter population in turn may acquire another advantageous mutation and the process goes on until the supply of beneficial mutations gets exhausted. A number of models with variable degrees of biological consistency have been proposed and investigated to understand the process of adaptation (Miller *et al.* 2011). One of the simplest mathematical models was introduced by Gillespie in which beneficial mutations arise sequentially and fix rapidly (Gillespie 1991). If the mutation rate is small and the selection coefficient is large (compared to the inverse population size), it is a good approximation to assume that only the one-step mutants are accessible at any time and the population is localized at a single genotype. Such a monomorphic population performs an adaptive walk by moving uphill on a fitness landscape until no more beneficial mutations can be found.

In the last few years, much of the work on Gillespie's model has focused on the first step in the adaptation process. If the fitnesses of the wild type and its one-mutant neighbors are rank ordered with the fittest sequence at the top, the well established theory of extremes of independent random variables (David and Nagaraja 2003) can be exploited to obtain useful information provided the wild type has a high fitness (rank). For a moderately high-ranked initial fitness, Orr calculated the expected rank at the first step assuming exponential-like fitness distributions (Orr 2002). His prediction has been tested in an experiment using single-stranded DNA and found to be roughly consistent with the experimental data (Rokyta *et al.* 2005). This result has been later generalized for other fitness distributions (Joyce *et al.* 2008) and by including correlations among fitnesses (Orr 2006). However, as the properties of the entire walk are required to design a drug or a biomolecule (Bull and Otto 2005) and as experimental data on multiple adaptive substitutions are becoming available (Rokyta *et al.* 2009; Schoustra *et al.* 2009), it is important to extend the existing theory to address the statistical properties of the entire walk.

With this aim, we study Gillespie's *mutational landscape model* on rugged fitness landscapes with many local fitness optima. An important difference between our work and the previous ones is that here we start the adaptive walk with low fitness to describe the adaptation process in novel environments such as when antibiotics are introduced (MacLean and Buckling 2009; McDonald *et al.* 2010) whereas the

initial fitness is assumed to be high in other studies (Gillespie 1991; Orr 2002, 2006; Joyce *et al.* 2008). Several numerical (Gillespie 1991; Orr 2006) and experimental studies (Rokyta *et al.* 2009; Schoustra *et al.* 2009) have indicated that only a few steps are required to reach a local optimum. In a simple adaptation model that assumes the mutational neighborhood remains unchanged during the entire adaptive walk (Gillespie 1983), the average number of steps to a local fitness peak has been calculated analytically for various fitness distributions and shown to increase logarithmically with the rank of the initial sequence (Neidhart and Krug 2011). However, here we work with a more realistic mutation scheme in which a new suite of mutants is created in each adaptive step. For generic fitness distributions, we argue that the average number of adaptive steps increases logarithmically with sequence length with a prefactor that depends on the choice of fitness distribution. Although our argument does not capture the proportionality constant correctly, the logarithmic dependence is seen to be in excellent agreement with the simulation results. We also present detailed results on the statistical properties of the entire walk for exponentially and uniformly distributed fitnesses as these two distributions lend themselves to an analytic treatment and are also consistent with the experiments (Eyre-Walker and Keightley 2007; Rokyta *et al.* 2008). Following the approach of Flyvbjerg and Lautrup (1992), we write a recursion relation for the fitness distribution of *fixed* beneficial mutations at an adaptive step that is valid for long sequences and fitness distributions with a finite mean. A similar distribution has been calculated in the clonal interference regime in which multiple mutants are produced per generation (Rozen *et al.* 2002) while here we work in the weak mutation regime. For the above-mentioned distributions, we also find the distribution of walk length. The average walk length calculated using this approach gives a prefactor consistent with the numerical results.

Although for most of the article we work with uncorrelated fitnesses and assume that the distribution of the fitness does not change during the course of evolution, the effect of correlations is also discussed. As experiments support an intermediate degree of correlations in fitness landscapes (Carneiro and Hartl 2010; Miller *et al.* 2011) and changing fitness distributions may be modeled by correlated fitnesses (Orr 2006), we calculate the average number of steps to an optimum on a fitness landscape generated by the block model of correlated fitnesses in which a sequence is divided into several independent blocks and correlations arise when two sequences share some blocks (Perelson and Macken 1995). The average walk length has been measured using numerical simulations in a block model in Orr (2006) and it was speculated that the average number of adaptive steps is independent of the underlying fitness distribution and increases linearly with the number of blocks. We show that while the latter result is roughly correct, the average number of steps to a local optimum is

not independent of the fitness distribution, which is a consequence of the result discussed above for the uncorrelated fitness landscapes.

Models and Methods

Uncorrelated and correlated fitness landscapes

An uncorrelated fitness landscape can be generated by assigning a fitness to a sequence independent of that of other sequences. The fitnesses are sampled from a common distribution $p(f)$ with support on the interval $[l, u]$. Although the full distribution of absolute fitness is unknown, one can obtain an insight into its nature through the distribution of beneficial mutations that has been measured in several theoretical and experimental studies (Eyre-Walker and Keightley 2007). A theoretical argument suggests that since good mutations are rare, their distribution is governed by the upper tail of the fitness distribution $p(f)$ (Gillespie 1991). It is known from the extreme value theory (EVT) for independent and identically distributed (i.i.d.) random variables that the asymptotic distribution of the extreme value can be one of the following three types (David and Nagaraja 2003): Fréchet for algebraically decaying underlying distributions, Gumbel for unbounded distributions decaying faster than a power law, and Weibull for bounded distributions. To be consistent with this result, we make the following choices for the fitness distributions:

$$p(f) = \begin{cases} (\delta - 1)(1 + f)^{-\delta}, & \delta > 2 & \text{(Fréchet)} & (1) \\ \gamma f^{\gamma-1} e^{-f^\gamma}, & \gamma > 0 & \text{(Gumbel)} & (2) \\ \nu(1-f)^{\nu-1}, & \nu > 0, f < 1 & \text{(Weibull)} & (3). \end{cases}$$

The condition $\delta > 2$ in (1) is imposed to keep the transition rate (6) finite (as explained later). The last two fitness functions (2) and (3) are of particular interest as several experimental results on the distribution of beneficial mutations have been found to lie in the Gumbel domain (Imhof and Schlotterer 2001; Sanjuán *et al.* 2004; Rokyta *et al.* 2005; Kassen and Bataillon 2006; MacLean and Buckling 2009) and a recent work finds a best fit for the distribution of beneficial effects to a uniform distribution that lies in the Weibull domain (Rokyta *et al.* 2008).

We also study adaptive walks on correlated fitness landscapes that are generated using a block model (Perelson and Macken 1995) in which a sequence of length L is divided into B blocks of equal size $L_B = L/B$. The block fitness is an i.i.d. random variable chosen from the distribution $p(f)$ and the sequence fitness is obtained by averaging over the fitnesses of the blocks in the sequence. If two sequences share one or more blocks, their fitnesses are correlated. The correlations can be tuned by changing the number of blocks: If the number of blocks $B = 1$, sequence fitnesses are completely uncorrelated while $B = L$ gives strongly correlated fitnesses. It should be noted that the extreme value distribution of correlated fitnesses may change from the

corresponding i.i.d. class even if correlations are weak (Jain *et al.* 2009; Jain 2011). In the following discussion, we assume that the sequence fitnesses are uncorrelated and deal with the correlated fitnesses in the last subsection of next section.

Adaptive walk model for long sequences

We work with haploid binary sequences of length L in the strong selection–weak mutation (SSWM) regime. If N is the population size, the SSWM regime corresponds to $Ns \gg 1$, $N\mu \ll 1$ where s is the selection coefficient and μ is the mutation probability per locus per generation. Since the expected number of mutants produced per generation is much smaller than one, mutations occur sequentially and double and higher mutations may be neglected. Thus the mutational neighborhood of a sequence is limited to L mutants that are a single mutation away from it. If the fitnesses of the wild-type sequence and its L one-mutant neighbors are arranged in a descending order with the best fitness assigned the rank 1, the transition probability that the population moves from the wild type with fitness rank i and value f_i to a mutant with rank $j < i$ and value f_j is proportional to the fixation probability, which is well approximated by $2(f_j - f_i)/f_i$ in the strong selection limit (Gillespie 1991). The normalized transition probability from fitness f_i to fitness f_j is given by

$$T(f_j \leftarrow f_i) = \frac{f_j - f_i}{\sum_{k=1}^{i-1} f_k - f_i}, \quad 1 \leq j \leq i-1. \quad (4)$$

Once the population has moved to a mutant sequence with fitness f_j with probability $T(f_j \leftarrow f_i)$, it produces a set of new mutants that are rank ordered and chosen according to (4) and the process repeats itself until the population reaches a local optimum whose nearest neighbors are all less fit than itself. Note that the parameters N and μ have dropped out of the picture and the properties of the model depend on the sequence length (or the initial rank) and the distribution of sequence fitnesses.

The model described above has been studied using (4) and EVT in previous works (Gillespie 1991; Orr 2002, 2006; Joyce *et al.* 2008) assuming the initial fitness to be high (small i). In contrast, we start with a low fitness and write a recursion relation for the probability $P_J(f)$ that an adaptive walk has at least J steps and the fitness is f at the J th step, following Flyvbjerg and Lautrup (1992) who studied this distribution for random adaptive walks (see Appendix A). In the following discussion, it is assumed that the sequence length is large, which allows the following two simplifications: First, the events in which a sequence is backtracked can be ignored and second, the transition rates can be written in terms of absolute fitnesses instead of fitness ranks. Consider a population at the J th adaptive step and with fitness h . It can proceed to the next step provided at least one fitter mutant is available. If $q^L(h) = \int_h^L dg p(g)$, this event occurs with a probability $1 -$

$q^L(h)$, where it is assumed that at each step in the evolutionary process, L novel mutants are available that have not been encountered before. While this is true at the first step, the number of novel mutants is $L - 1$ at the second step since one of the mutants is the parent sequence itself that is not an allowed descendant as the walk always proceeds uphill. In fact for any $J \geq 2$, some of the mutants have already been probed but the error introduced by ignoring this complication is of the order of $1/L$, which is negligible for large L (Flyvbjerg and Lautrup 1992). Then for long sequences we can write

$$P_{J+1}(f) = \int_l^f dh p(f) T(f \leftarrow h) (1 - q^L(h)) P_J(h), \quad J \geq 0, \quad (5)$$

where $p(f) T(f \leftarrow h)$ gives the probability that a mutant with fitness $f > h$ is chosen. Furthermore for large L , it is a good approximation to replace the sum in the denominator of (4) by an integral and we may write

$$T(f \leftarrow h) = \frac{f - h}{\int_h^u dg (g - h) p(g)}, \quad f > h. \quad (6)$$

Thus we work with absolute fitnesses instead of fitness ranks. Since the transition probability (6) vanishes for slowly decaying fitness distributions $p(f) \sim f^{-\delta}$, $\delta \leq 2$, we restrict $\delta > 2$ in (1). Using (6) in (5), we finally obtain

$$P_{J+1}(f) = \int_l^f dh \frac{(f - h) p(f)}{\int_h^u dg (g - h) p(g)} (1 - q^L(h)) P_J(h), \quad J \geq 0. \quad (7)$$

Equation 7 is the central equation of this article and we employ it to obtain various results on the statistical properties of adaptive walks. In the following, we assume the initial condition $P_0(f) = \delta(f)$ corresponding to zero initial fitness. As $P_J(f)$ obeys an integral equation that is harder to analyze, we may try to write a differential equation for $P_J(f)$. Differentiating (7) with respect to f , we get

$$P'_{J+1}(f) = \int_l^f dh \frac{(f - h) p'(f) + p(f)}{\int_h^u dg (g - h) p(g)} (1 - q^L(h)) P_J(h), \quad J \geq 0 \quad (8)$$

$$P''_{J+1}(f) = \int_l^f dh \frac{(f - h) p''(f) + 2p'(f)}{\int_h^u dg (g - h) p(g)} (1 - q^L(h)) P_J(h) + \frac{p(f)(1 - q^L(f))}{\int_f^u dg (g - f) p(g)} P_J(f), \quad J \geq 1, \quad (9)$$

where the prime denotes an f -derivative. On using (7) and (8) in (9), we find

$$P_{J+1}''(f) = 2 \frac{p'(f)}{p(f)} P_{J+1}'(f) + \left[\frac{p''(f)}{p(f)} - 2 \left(\frac{p'(f)}{p(f)} \right)^2 \right] P_{J+1}(f) + \frac{p(f)(1-q^L(f))}{\int_f^u dg (g-f)p(g)} P_J(f), \quad J \geq 1. \quad (10)$$

The first derivative term in the above equation can be eliminated by writing $P_J(f) = p(f)\tilde{P}_J(f)$, which finally yields

$$\tilde{P}_{J+1}''(f) = \frac{p(f)(1-q^L(f))}{\int_f^u dg (g-f)p(g)} \tilde{P}_J(f), \quad J \geq 1. \quad (11)$$

In this article, we restrict our attention to exponentially and uniformly distributed fitnesses as these two fitness distributions are consistent with the available empirical data. We show that due to (9), a second-order ordinary differential equation is obeyed by a generating function of $P_J(f)$ for these two distributions, which can be solved within an approximation subject to the boundary conditions

$$P_J(f)|_{f=l} = 0, \quad J \geq 1 \quad (12)$$

$$P_J(f)|_{f=l} = \frac{p(l)}{\int_l^u dg g p(g)} \delta_{J,1}, \quad (13)$$

where (12) is a direct consequence of (7) and (13) arises on using the initial condition in (8).

Besides $P_J(f)$, we also find the walk length distribution Q_J and the average fitness f_J at the J th step, which can be related to $P_J(f)$ as explained below. Integrating over f on both sides of (7), we get

$$P_{J+1} = \int_l^u df P_{J+1}(f) \quad (14)$$

$$= \int_l^u dh \int_h^u df \frac{(f-h)p(f)}{\int_h^u dg (g-h)p(g)} (1-q^L(h)) P_J(h) \quad (15)$$

$$= \int_l^u dh (1-q^L(h)) P_J(h) = P_J - \int_l^u dh q^L(h) P_J(h). \quad (16)$$

Figure 1 Evolution of average fitness with the number of adaptive steps starting from zero initial fitness obtained numerically (points) and compared with the average fitness in infinite sequence length limit (lines) for (A) power law-distributed fitness with $\delta = 6$, Equation 22; (B) exponentially distributed fitness, Equation 23; and (C) uniformly distributed fitness, Equation 25.

Then the walk length probability Q_J that exactly J steps are taken is given by

$$Q_J = P_J - P_{J+1} = \int_l^u dh q^L(h) P_J(h) \quad (17)$$

with $Q_0 = 0$ since the initial fitness is zero. The above equation has a simple interpretation: Since $P_J(h)$ is the probability that at least J steps are taken and the fitness at the J th step is h , exactly J steps will be taken if all the L mutants of the sequence at the J th step carry a fitness smaller than h from which (17) follows. The average walk length $\bar{J} = \sum_{J=0}^{2^L} J Q_J \approx \sum_{J=0}^{\infty} J Q_J$ for large L . The average fitness $\bar{f}_J$ is defined as $\bar{f}_J = \int_l^u df f P_J(f)$. Using (7), we can write

$$\bar{f}_{J+1} = \int_l^u df f \int_l^f dh \frac{(f-h)p(f)}{\int_h^u dg(g-h)p(g)} (1 - q^L(h)) P_J(h) \quad (18)$$

$$= \int_l^u dh \frac{(1 - q^L(h)) P_J(h)}{\int_h^u dg(g-h)p(g)} \int_h^u df f(f-h)p(f). \quad (19)$$

Note that neither (17) nor (19) is a closed equation.

Our analytical results are also compared with numerical simulations that were performed using an exact procedure for $L \leq 10$ and an approximate method outlined in Orr (2002) for larger L . We refer the reader to Appendix B for details.

Results

Average fitness and walk length for general fitness distributions

For a broad class of fitness distributions, the average fitness for an infinitely long sequence can be computed. Although this limit is biologically unrealistic, it provides a good approximation to the average fitness $\bar{f}_J$ for small J (see Figure 1) as the population cannot sense the finiteness of sequence length far from the local optimum. On taking the limit $L \rightarrow \infty$ in (19) and denoting the average fitness in this limit by F_J , we obtain

$$F_{J+1} = \int_l^u dh \frac{\int_h^u df f(f-h)p(f)}{\int_h^u dg(g-h)p(g)} P_J(h)|_{L \rightarrow \infty}. \quad (20)$$

Algebraically decaying fitness distributions: On substituting (1) in (20) and performing the integrals involving $p(f)$, we get

$$F_{J+1} = \int_0^\infty dh \frac{2 + (\delta - 1)h}{\delta - 3} P_J(h)|_{L \rightarrow \infty} = \frac{2}{\delta - 3} + \frac{\delta - 1}{\delta - 3} F_J, \quad \delta > 3, \quad (21)$$

where we have used that $P_J|_{L \rightarrow \infty} = 1$ due to (16) and the initial condition $P_0 = 1$. Repeated iteration with $F_0 = 1$ yields

$$F_J = \left(\frac{\delta - 1}{\delta - 3} \right)^J - 1, \quad (22)$$

which increases geometrically with J . This result is compared in Figure 1A with the average fitness for finite sequences, which shows that the number of steps up to which $\bar{f}_J$ and F_J match increases with L .

Exponential fitness distribution: For fitness distributions given by (2), the equation for F_J does not close except for $\gamma = 1$. For $p(f) = e^{-f}$, we get $F_J = 2 + F_{J-1}$, which gives

$$F_J = 2J. \quad (23)$$

Figure 1B shows that the rate of increase of fitness $\bar{f}_J$ is slower than a constant at larger J 's.

Bounded fitness distributions: A calculation similar to that above for $p(f)$ in (3) gives

$$F_{J+1} = \frac{2 + \nu F_J}{2 + \nu} \quad (24)$$

and therefore

$$F_J = 1 - \left(\frac{\nu}{2 + \nu} \right)^J. \quad (25)$$

For uniformly distributed fitness ($\nu = 1$), we find that $1 - F_J = 3^{-J}$ in good agreement with the numerical data in Figure 1 for small J .

We now give an argument to estimate the average walk length $\bar{J}$ using the above results for the average fitness F_J and the EVT (Flyvbjerg and Lautrup 1992). We first note that since $P_J|_{L \rightarrow \infty} = 1$ for all J , every step in the adaptive walk is definitely taken for infinitely long sequences and hence the average walk length is expected to diverge with L . For a sequence of finite length, the adaptive walk stops when the population has reached a local optimum whose fitness is the largest among $L + 1$ i.i.d. random variables. But since the average number of fitnesses with value $\geq f$ is given by $(L + 1)(1 - q(f))$, at a local optimum we have

$$(L + 1) \int_{F_J}^u df p(f) = 1 \quad (26)$$

(Sornette 2000), where we have approximated $\bar{f}_J$ by F_J . The above equation yields

$$F_J \approx \begin{cases} L^{1/(\delta-1)} - 1 & \text{(Algebraic)} \\ \ln L & \text{(Exponential)} \\ 1 - L^{-(1/\nu)} & \text{(Bounded)}. \end{cases} \quad (27) \quad (28) \quad (29)$$

Figure 2 Average number $\bar{J}$ of adaptive steps as a function of sequence length L for various fitness distributions when the fitnesses are uncorrelated. The points show the data obtained using numerical simulations and the lines are the best fit to the function $\bar{J} = \alpha \ln L + \beta$. The results for the greedy walk and the random adaptive walk (up to an additive constant) are also shown. The numerical fit for the prefactor α for exponential and uniform fitness distribution matches well with the analytical results given by (46) and (56), respectively.

On matching the expected fitness $F_{\bar{J}}$ with the F_J obtained in the above discussion for various distributions, we get

$$\bar{J} \approx \begin{cases} \frac{1}{\delta-1} \frac{\ln L}{\ln((\delta-1)/(\delta-3))} & \text{(Algebraic)} & (30) \\ \frac{1}{2} \ln L & \text{(Exponential)} & (31) \\ \frac{1}{\nu} \frac{\ln L}{\ln((2+\nu)/\nu)} & \text{(Bounded).} & (32) \end{cases}$$

Thus the above argument shows that for large L ,

$$\bar{J} \approx \alpha \ln L, \quad (33)$$

where the prefactor α depends on $p(f)$. We note that $\alpha_{\text{algebraic}} < \alpha_{\text{exponential}} < \alpha_{\text{bounded}}$, which implies that smaller numbers of substitutions occur for fat-tailed fitness distributions than for the bounded ones. To understand this qualitative trend, consider the transition probability for the first step given by $T(f \leftarrow 0)p(f) \sim fp(f)$. At large f , this probability is higher for slowly decaying distributions and thus a large fitness gain occurs initially. But as the probability to exceed the high fitness achieved at the first step is small, the walk terminates sooner for broad distributions.

The results of our numerical simulations for $\bar{J}$ shown in Figure 2 are in agreement with the logarithmic dependence on L but the value of the prefactor does not match with that obtained above [except for $p(f) = e^{-f}$]. The prefactor α is expected to interpolate between the two limiting cases of adaptive walks, namely greedy walk in which the best mutant is chosen with probability one and random adaptive walk in which all better mutants are chosen with equal probability. The former limit is obtained when $\delta \rightarrow 1$ in (1) and the latter when $\nu \rightarrow 0$ in (3) (Joyce *et al.* 2008).

Since the average walk length for a greedy walker is a finite constant equal to $e - 1 \approx 1.718$ for infinitely long sequences (Orr 2003), the prefactor $\alpha = 0$ while $\alpha = 1$ for the random adaptive walk (see Appendix A). In the following sections, we find that $\alpha = \frac{1}{2}$ for exponentially distributed fitness and $\alpha = \frac{2}{3}$ for the uniform case, which are consistent with the results in Figure 2 and the analytical results of Neidhart and Krug (2011), which are obtained using a simpler version of the adaptive walk model considered here.

Fitness distribution at the first step for general distributions

If the whole population is assumed to have an initial fitness f_0 , using $P_0(f) = \delta(f - f_0)$ in (7) we have

$$P_1(f) = \frac{(f - f_0)p(f)(1 - q^L(f_0))}{\int_{f_0}^u dg(g - f_0)p(g)} \propto (f - f_0)p(f). \quad (34)$$

The above fitness distribution at the first step is nonmonotonic for all fitness distributions in (1–3) except for truncated distributions with $\nu \leq 1$. The implications of this result are examined in the *Discussion*.

Entire walk with exponentially distributed fitness

For $p(f) = e^{-f}$, from (11) we obtain

$$\tilde{P}_{J+1}'(f) = (1 - q^L(f))\tilde{P}_J(f), \quad J \geq 1, \quad (35)$$

where $q(f) = 1 - e^{-f}$. Due to (12) and (13), the boundary conditions are $P_J(0) = 0$ and $P_J'(0) = \delta_{J,1}$.

We define a generating function $G(x, f) = \sum_{J=1}^{\infty} \tilde{P}_J(f)x^J$, $x < 1$, which obeys the following second-order ordinary differential equation:

$$G''(x, f) = x(1 - q^L(f))G(x, f). \quad (36)$$

To arrive at the above equation, we have used that $\tilde{P}_1(f) = f$, which is obtained by using the initial condition in (7). The generating function $G(x, f)$ obeys a Schrödinger equation for the wave function of a particle in a one-dimensional potential $V(f) \sim 1 - q^L(f)$ and energy zero (Mathews and Walker 2004). Since $1 - q^L(f) \approx 1 - e^{-Le^{-f}}$ is close to unity for $f \ll \ln L$ and vanishes for $f \gg \ln L$, the potential $V(f)$ decreases smoothly from one to zero and moves rightward with increasing L . Similar potentials also arise when two materials with different transport properties are joined together and in such systems, an analytical solution is obtained within a step function potential approximation (Blonder *et al.* 1982; Schaeybroeck and Lazarides 2009). We follow this approach here and approximate the distribution $1 - q^L(f)$ by the Heaviside theta function $\Theta(\tilde{f} - f)$, where $\tilde{f} = \ln L$. Within this *step distribution approximation*, we have

$$G''(x, f) = \begin{cases} xG(x, f), & f < \tilde{f} \\ 0, & f > \tilde{f} \end{cases} \quad (37)$$

Figure 3 Comparison of the distribution $P_J(f)$ for $J = 1, 2, 3, 5$ obtained numerically (points) and analytically (lines) given by (43) for exponentially distributed fitness and sequence length $L = 1000$. Inset: Numerical data for $P_J(f)$ for $J = 4, 5, 6$ to show that the fitness distribution does not shift appreciably beyond $\bar{f} \approx 4.6$ as the local optimum with average fitness ≈ 7 is approached.

For $f < \bar{f}$, the differential Equation 37 has a solution of the form $G_<(x, f) = a_+ e^{\sqrt{x}f} + a_- e^{-\sqrt{x}f}$, which reduces to $G_<(x, f) = c \sinh(\sqrt{x}f)$ since $G(x, 0) = 0$ due to $P_J(0) = 0$. Since the solution for $f < \bar{f}$ cannot depend on $\bar{f}$, we appeal to the infinite sequence length limit to fix the proportionality constant c . As noted earlier, the distribution $P_J|_{L \rightarrow \infty} = 1$ for all $J \geq 0$, which implies that

$$\int_0^\infty df e^{-f} G_<(x, f) = \frac{x}{1-x} \quad (38)$$

and therefore

$$G_<(x, f) = \sqrt{x} \sinh(\sqrt{x}f). \quad (39)$$

We check that the boundary condition $P'_J(0) = \tilde{P}'_J(0) = \delta_{J,1}$, which is equivalent to $G'(x, 0) = x$, is also satisfied by the above solution.

For $f > \bar{f}$, the solution $G_>(x, f) = af + b$, where the constants of integration a, b can be fixed by matching the solutions $G_<$ and $G_>$ and their first derivative at $f = \bar{f}$. Thus the constants a and b are determined by the following conditions:

$$G_<(x, \bar{f}) = G_>(x, \bar{f}) = a\bar{f} + b \quad (40)$$

$$G'_<(x, f)|_{f=\bar{f}} = G'_>(x, f)|_{f=\bar{f}} = a. \quad (41)$$

A simple algebra shows that

$$G_>(x, f) = x \cosh(\sqrt{x}\bar{f})(f - \bar{f}) + \sqrt{x} \sinh(\sqrt{x}\bar{f}). \quad (42)$$

Using the above expressions for $G(x, f)$, the fitness distribution $P_J(f)$ for the fixed beneficial mutations can be calculated. On expanding (39) and (42) in a power series about $x = 0$ and picking the coefficient of x^J , we have

Figure 4 Walk length distribution Q_J for $p(f) = e^{-f}$ comparing numerical (points) and analytical results (lines) given by (45).

$$P_J(f) = \frac{e^{-f} f^{2J-1}}{(2J-1)!} \times \begin{cases} 1, & r \leq 1 \\ \frac{(2J-1)r - (2J-2)}{r^{2J-1}}, & r > 1, \end{cases} \quad (43)$$

where $r = f/\bar{f}$. Figure 3 shows our numerical results for $P_J(f)$ for the first few adaptive steps. As the walk proceeds, the distribution moves rightward as expected and its amplitude decreases since the probability $q^L(f)$ that the walker cannot find a better neighbor approaches unity with increasing f . Our analytical result (43) is also shown in Figure 3 for comparison. For $L = 10^3$, the step distribution approximation used to find (43) gives $1 - q^L(f) \approx 1$ for $f < \ln L = 6.9$ and zero otherwise. However, as the probability $1 - q^L(f)$ stays close to unity for $f \leq 5$ and decreases gradually to zero when $f \approx 12$, the distribution (43) in the region $5 < f < 12$ does not match well with the simulation results but outside this crossover region, we see a good quantitative agreement. We also note that the fitness distribution does not move appreciably for $J \geq 4$ and is centered around $f \approx 7$ (see inset in Figure 3). This is because the average walk length for $L = 10^3$ is ~ 4.6 steps (refer to Figure 2) and as the local optimum is approached, the fitness distribution of fixed beneficial mutation remains centered close to the typical fitness of the local optimum given by (26), which is $\ln L \approx 6.9$. This also explains the initial linear rise in the average fitness followed by a slower increase in Figure 1.

We next calculate the walk length distribution Q_J defined by (17). Since $q^L(f) = \Theta(f - \bar{f})$ within the step distribution approximation discussed above, (17) reduces to

$$Q_J = \int_{\bar{f}}^\infty df P_J(f). \quad (44)$$

On integrating $P_J(f)$ given in (43), we get

$$Q_J = e^{-\ln L} \left[\frac{(\ln L)^{2J-2}}{(2J-2)!} + \frac{(\ln L)^{2J-1}}{(2J-1)!} \right], \quad J > 0. \quad (45)$$

Figure 5 Comparison of the distribution $P_J(f)$ for $J = 1, 2, 3, 4$ obtained numerically (points) and analytically (lines) given by (C6–C9) for uniformly distributed fitness and sequence length $L = 100$. The distribution for $f \leq \tilde{f}$ is shown in the main plot and for $f > \tilde{f}$ in the inset.

This expression is compared with numerical results in Figure 4 and shows a reasonable agreement. The average number of adaptive steps calculated using (45) is given by

$$\bar{J} = \sum_{J=1}^{\infty} J Q_J \approx \frac{1}{2} \ln L, \quad (46)$$

which is in good agreement with the simulation result in Figure 2. The width of the distribution Q_J measured using the variance $\sigma^2 = \bar{J}^2 - \bar{J}^2 \approx \ln L / 4$ also increases with L .

Entire walk with uniformly distributed fitness

For $p(f) = 1$, since $P_J(f) = \tilde{P}_J(f)$, the differential equation (11) reduces to

$$P_{J+1}''(f) = \frac{1-f^L}{\int_1^f dg(g-f)} P_J(f) = \frac{2(1-f^L)}{(1-f)^2} P_J(f), \quad J \geq 1 \quad (47)$$

with boundary conditions $P_J(0) = 0$ and $P_J'(0) = 2\delta_{J,1}$. As before, we define a generating function $G(x, f) = \sum_{J=2}^{\infty} x^{J-2} P_J(f)$ that obeys the second-order ordinary differential equation

$$G''(x, f) = \frac{2(1-f^L)}{(1-f)^2} (xG(x, f) + 2f), \quad (48)$$

where we have used that $P_1(f) = 2f$. We treat this case also within the step distribution approximation discussed earlier. Since the probability $1-f^L \approx 1 - e^{-L(1-f)}$, we approximate it by a step function $\Theta(\tilde{f}-f)$, where $\tilde{f} = (L-1)/L$. For $f < \tilde{f}$, we obtain an inhomogeneous second-order ordinary differential equation with variable coefficients:

$$G_{<}''(x, f) = \frac{2x}{(1-f)^2} G_{<}(x, f) + \frac{4f}{(1-f)^2}. \quad (49)$$

Figure 6 Walk length distribution Q_J for uniformly distributed fitness comparing simulation (points) and analytical results (lines) in (C27).

This equation can be solved by standard methods (detailed in Appendix C) to yield

$$G_{<}(x, f) = a_+(1-f)^{\alpha_+} + a_-(1-f)^{\alpha_-} + u_+(f)(1-f)^{\alpha_+} + u_-(f)(1-f)^{\alpha_-}, \quad (50)$$

where the exponents

$$\alpha_{\pm} = \frac{1 \pm \sqrt{1+8x}}{2}. \quad (51)$$

The first two terms on the right-hand side give the solution of the homogeneous equation and the last two terms are the particular integral involving the variational parameters $u_{\pm}(f)$ given in Appendix C. The constants of integration $a_{\pm}$ can be obtained using the boundary conditions $G(x, 0) = 0$ and $\int_0^1 df G_{<}(x, f) = (1-x)^{-1}$. After some straightforward algebra, we find that

$$G_{<}(x, f) = \frac{-2}{x} \left[\frac{(1-f)^{\alpha_+} - (1-f)^{\alpha_-}}{\alpha_+ - \alpha_-} + f \right]. \quad (52)$$

We verify that the condition $P_J'(0) = 0$ for $J > 1$ that amounts to $G'(x, 0) = 0$ is also satisfied. For $f > \tilde{f}$, as $G_{>}''(x, f) = 0$, the solution $G_{>}(x, f) = af + b$, where a , b can be determined using (40) and (41) to give

$$G_{>}(x, f) = \frac{-2}{x} \left[\frac{\alpha_-(1-\tilde{f})^{\alpha_- - 1} - \alpha_+(1-\tilde{f})^{\alpha_+ - 1}}{\alpha_+ - \alpha_-} + 1 \right] f - \frac{2}{x} \left[\frac{(1-\tilde{f})^{\alpha_+} - (1-\tilde{f})^{\alpha_-} - \alpha_-\tilde{f}(1-\tilde{f})^{\alpha_- - 1} + \alpha_+\tilde{f}(1-\tilde{f})^{\alpha_+ - 1}}{\alpha_+ - \alpha_-} \right]. \quad (53)$$

Explicit expressions for $P_J(f)$ for first few adaptive steps are given in Appendix C and a comparison between the analytical and the simulation results is shown in Figure 5.

To find the walk length distribution $Q_J = \int_0^1 df P_J(f)$, we define

Figure 7 Average number $\bar{J}_B$ of adaptive steps as a function of block number B for fixed $L/B = 100$. The numerical data are in excellent agreement with (59), shown by solid lines.

$$H(x) = \sum_{J=1}^{\infty} x^J Q_J = xQ_1 + x^2 \int_{\tilde{f}}^1 df G_{>}(x, f) \quad (54)$$

$$= \frac{x(1-\tilde{f})}{\alpha_- - \alpha_+} \left[(2 - \alpha_+)(1-\tilde{f})^{\alpha_+} - (2 - \alpha_-)(1-\tilde{f})^{\alpha_-} \right]. \quad (55)$$

As an explicit expression for Q_J is rather unwieldy, its derivation and the expression itself are given in *Appendix C* and a comparison with the simulations is shown in *Figure 6*. The average number of steps is given by

$$\bar{J} = \left. \frac{dH(x)}{dx} \right|_{x=1} = \frac{-6 \ln(1-\tilde{f})}{9}, \quad (56)$$

which shows that for large L , the number of adaptive steps grows as $(2/3) \ln L$ in agreement with the numerical results shown in *Figure 2*. The higher moments can also be found straightforwardly and we find that the variance $\bar{J}^2 - \bar{J}^2 \approx (10/27) \ln L$ and the skewness of the distribution decays slowly as $(\ln L)^{-1/2}$.

Effect of correlations on the number of adaptive steps

We now turn to a discussion of adaptive walk properties when the fitnesses are correlated and given by a block model. We compute the average number $\bar{J}_B(L)$ of adaptive steps given by $\sum_{J=1}^{\infty} J Q_J(L, B)$, where $Q_J(L, B)$ is the probability that exactly J adaptive mutations occur when a sequence of length L is divided into B blocks.

Consider the distribution $Q(m_1, \dots, m_B)$, which gives the joint probability that the i th block of length L_B in a sequence of length L carries m_i adaptive mutations, where $i = 1, \dots, B$. An important property of the block model is that this joint distribution factorizes; that is,

$$Q(m_1, \dots, m_B) = \prod_{b=1}^B Q_{m_b}(L_B, 1) \quad (57)$$

(Perelson and Macken 1995), where $Q_J(L_B, 1) \equiv Q_J(L_B)$ is the walk length probability when the fitnesses are uncorrelated and the sequence length is L_B . The above equation expresses the fact that the block fitnesses evolve independently. As only one mutation occurs in the sequence at any step so that all but one block sequence remains unchanged and since the block fitnesses are i.i.d. random variables, (57) holds.

Since the distribution $Q_J(L, B)$ is given by

$$Q_J(L, B) = \sum_{m_1, \dots, m_B=0}^J Q(m_1, \dots, m_B) \delta(m_1 + \dots + m_B - J), \quad (58)$$

it follows that

$$\begin{aligned} \bar{J}_B(L) &= \sum_{J=1}^{\infty} J \sum_{m_B=0}^J Q_{m_B}(L_B) \sum_{m_1, \dots, m_{B-1}=0}^{J-m_B} \prod_{b=1}^{B-1} Q_{m_b}(L_B) \\ &\quad \times \delta\left(\sum_{b=1}^{B-1} m_b - (J - m_B)\right) \\ &= \sum_{J=1}^{\infty} J \sum_{m_B=0}^J Q_{m_B}(L_B) Q_{J-m_B}(L - L_B, B-1) \\ &= \sum_{m=0}^{\infty} Q_m(L - L_B, B-1) \sum_{n=0}^{\infty} (n+m) Q_n(L_B) \\ &= \bar{J}(L_B) + \sum_{m=1}^{\infty} m Q_m(L - L_B, B-1) \\ &= \bar{J}(L_B) + \bar{J}_{B-1}(L - L_B) \\ &= B \bar{J}(L_B), \end{aligned} \quad (59)$$

where we have used that $\sum_{J=0}^{\infty} Q_J(L, B) = 1$ and $\bar{J}$ is the average number of steps in the adaptive walk for uncorrelated fitnesses. *Figure 7* shows the results of our numerical simulations for average walk length when the block length $L_B = L/B$ is kept fixed and the block fitnesses are exponentially and uniformly distributed. For fixed L_B , (59) predicts that $\bar{J}_B$ increases linearly with B , which is in excellent agreement with the numerical data.

For large L , due to (33) we have

$$\bar{J}_B(L) \approx \alpha B \ln \frac{L}{B}. \quad (60)$$

For small B , a linear rise in the average number of steps with the number of blocks has been seen numerically for exponential-like distributions and it was inferred that the mean walk length is independent of underlying fitness distributions (Orr 2006). However, as discussed in the previous sections, the average number $\bar{J}$ depends on the fitness distribution $p(f)$ and therefore the average $\bar{J}_B$ is also nonuniversal.

Discussion

In the last few years, several analytical results have been obtained for the mutational landscape model (Gillespie 1991). However, many of these results deal with the first step in the adaptation process (Orr 2002, 2006; Joyce *et al.* 2008) and an extension of the theory to full adaptive walk is necessary. Previous studies also assume that the process of adaptation starts from a highly fit sequence that is not applicable to situations in which the population is subjected to high stress and hence has a very low initial fitness (MacLean and Buckling 2009; McDonald *et al.* 2010). In this article, we have obtained results for the entire adaptive walk starting from a low initial fitness but as discussed below, we expect some of these results to hold for moderately high initial fitness also.

Walk length distribution and average walk length

In previous works, the walk length distributions for the greedy walk and the random adaptive walk have been studied and found to be universal in that they are independent of the underlying fitness distribution $p(f)$. The origin of this universality property is clear in the light of the results of Joyce *et al.* (2008) who pointed out that these two models can be obtained as a limit of (4), which defines the mutational landscape model. For the random adaptive walk, the distribution Q_J for infinitely long sequence vanishes (see Equation A3) and the average walk length diverges with sequence length. In contrast, for the greedy walk, the walk length distribution in the $L \rightarrow \infty$ limit decreases exponentially fast with J for the greedy walk as a result of which the average number of steps turns out to be a constant (Orr 2003; Rosenberg 2005).

In this article, we have calculated the walk length distribution for exponentially and uniformly distributed fitnesses and found the average walk length for general fitness distributions. An important conclusion of our study is that the average number of adaptive steps increases logarithmically with the sequence length with a prefactor smaller than unity if the walk starts from zero fitness. Our simulations (not shown) also indicate that if the initial rank is of order L , the average number of steps increases logarithmically with the rank and with the same proportionality constant as that for the zero initial fitness case. Thus for a wild-type sequence with initial rank (or L) of the order 100, the number of substitutions is expected to be less than five. Although short adaptive walks have been observed in experiments (Rokyta *et al.* 2009; Schoustra *et al.* 2009), more detailed experimental studies testing the logarithmic dependence would be desirable. Although a test of the L dependence of the average walk length may not be experimentally viable, it should be possible to study the average walk length as a function of the initial rank.

Besides the sequence length, the number of steps to a local optimum depends on the underlying fitness distribution and the fitness correlations also. If the fitnesses are uncorrelated, as the numerical data in Figure 2 show, the

Figure 8 Distribution $P(s_j)$ of selection coefficient s_j for $L = 1000$ and $p(f) = e^{-f}$. The inset shows the decay in average selection coefficient $\bar{s}_J$ as a function of J . The points are joined by line to guide the eye.

prefactor α in (33) depends on the shape of the fitness distribution and therefore a rather detailed knowledge of the full fitness distribution (how fast it decays) is required to test this, which is presently unavailable. However, one can discern a trend in the value of α : It decreases as the fitness distribution broadens. This suggests that systems with fitness distribution in the Gumbel class (Imhof and Schlotterer 2001; Sanjuán *et al.* 2004; Rokyta *et al.* 2005; Kassen and Bataillon 2006; MacLean and Buckling 2009) will register shorter walks than those in the Weibull domain (Rokyta *et al.* 2008). As shown here in the block model of correlated fitnesses, the average number of adaptive steps increases as the number of blocks (and hence fitness correlations) increases. This is in accordance with the expectation that on a smooth correlated fitness landscape, as the local optima are less common (Perelson and Macken 1995), there is less chance to get trapped and therefore the uphill walk can last longer (Weinberger 1991; Kauffman 1993; Orr 2006).

Distribution of fixed beneficial mutations during the walk

The fitness distribution $P_J(f)$ has not been studied in previous theoretical studies of adaptive walks in the SSWM limit and here we have computed this fitness distribution analytically using the recursion relation (7). The fitness distribution at the first step given by (34) can give a qualitative idea about the shape of $p(f)$. For most fitness distributions, $P_1(f)$ is expected to be nonmonotonic but for bounded distributions that diverge at the upper limit or the uniform distribution, $P_1(f)$ increases monotonically toward the upper bound. An inspection of the experimental data of Rokyta *et al.* (2005) shows the fitness distribution at the first step to be nonmonotonic, which is consistent with their assumption of exponentially decreasing distribution of beneficial effects. It would be interesting to check whether the distribution $P_1(f)$ in Rokyta *et al.* (2008) is monotonic as the data in this study are consistent with a uniformly distributed

fitness. The above behavior of $P_1(f)$ is expected to be robust in the presence of correlations as at the first step in evolution, the population has not sensed the correlations in the fitness landscape (Orr 2006).

For the fitness distribution for the entire walk, we presented an analysis for two distributions, namely exponential and uniform, that are consistent with the available experimental data. The distribution $P_J(f)$ is obtained within a step distribution approximation that captures the shape of the fitness distribution correctly for the first few steps and leads to an accurate estimate of the number of average steps. Our approximation consists of replacing the probability $1 - q^L(f)$ by a step function $\Theta(\hat{f} - f)$, where $\hat{f}$ is given by (28) for exponentially and by (29) for uniformly distributed fitnesses. For $f \ll \hat{f}$ and $f \gg \hat{f}$, our approximate solution matches the simulation results well for any J . With increasing J , the distribution $P_J(f)$ shifts toward higher fitnesses and peaks about $\hat{f}$ for larger J s. As explained earlier, the fitness $\hat{f}$ is reached when J is close to $\bar{J} \propto \ln L$ and therefore we expect our approximation to work well for $J \ll \ln L$.

When the underlying fitness distribution is exponential, we find that the fitness distribution of the fixed beneficial mutation also has an exponential tail (see Equation 43). The robustness of this result, *i.e.*, whether any fitness distribution in the Gumbel class exhibits an exponential tail for $P_J(f)$, is, however, not clear. For uniformly distributed fitnesses, as the width of the distribution $1 - q^L(f)$ decreases with increasing L , the step distribution approximation works better in this case than in the exponential case where the width is a constant (compare Figures 4 and 6). The properties of multiple steps in an adaptive walk have been measured in some recent experiments (Rokyta *et al.* 2009; Schoustra *et al.* 2009) and a detailed analysis of the experimental results would be very welcome. On the theoretical front, an extension of the results described above to distributions other than uniform and exponential would be desirable. We have recently made some progress in this direction and the results will appear elsewhere.

Another interesting question concerns the distribution $P(s_J)$ of the selection coefficient $s_J = (f_J - f_{J-1})/f_{J-1}$ at the J th step in the adaptive walk. As we start with zero fitness, the selection coefficient is defined for $J \geq 2$. Our preliminary numerical results for $P(s_J)$ are shown in Figure 8 for the first few steps in the walk and we observe that the typical selection coefficient decreases as the walk proceeds. This behavior matches qualitatively with the experimental results of Schoustra *et al.* (2009). A theoretical analysis of the distribution $P(s_J)$ requires the joint distribution of the fitness at step $J - 1$ and J and we hope to address this question in a future work.

Acknowledgments

K.J. thanks J. R. David for helpful suggestions and J. Krug for comments on an earlier version of the manuscript and useful correspondence. The authors also thank L. Wahl for

suggestions to improve the manuscript. K.J. thanks Kavli Institute of Theoretical Physics, Santa Barbara for hospitality and support under National Science Foundation grant PHY05-51164.

Literature Cited

- Blonder, G. E., M. Tinkham, and T. M. Klapwijk, 1982 Transition from metallic to tunneling regimes in superconducting micro-constrictions: excess current, charge imbalance, and supercurrent conversion. *Phys. Rev. B* 25: 4515.
- Bull, J. J., and S. P. Otto, 2005 The first steps in adaptive evolution. *Nat. Genet.* 37: 342–343.
- Carneiro, C., and D. Hartl, 2010 Adaptive landscapes and protein evolution. *Proc. Natl. Acad. Sci. USA* 107: 1747–1751.
- David, H., and H. Nagaraja, 2003 *Order Statistics*. Wiley, New York.
- Eyre-Walker, A., and P. Keightley, 2007 The distribution of fitness effects of new mutations. *Nat. Rev. Genet.* 8: 610.
- Flyvbjerg, H., and B. Lautrup, 1992 Evolution in a rugged fitness landscape. *Phys. Rev. A* 46: 6714–6723.
- Gillespie, J. H., 1983 A simple stochastic gene substitution process. *Theor. Popul. Biol.* 23: 202–215.
- Gillespie, J. H., 1991 *The Causes of Molecular Evolution*. Oxford University Press, Oxford.
- Imhof, M., and C. Schlotterer, 2001 Fitness effects of advantageous mutations in evolving *Escherichia coli* populations. *Proc. Natl. Acad. Sci. USA* 98: 1113–1117.
- Jain, K., 2011 Extreme value distributions for weakly correlated fitnesses in block model. *J. Stat. Mech.* 2011: P04020.
- Jain, K., A. Dasgupta, and G. Das, 2009 Exact and limit distributions of the largest fitness on correlated fitness landscapes. *J. Stat. Mech.* 2009: L10001.
- Joyce, P., D. R. Rokyta, C. J. Beisel, and H. A. Orr, 2008 A general extreme value theory model for the adaptation of DNA sequences under strong selection and weak mutation. *Genetics* 180: 1627–1643.
- Kassen, R., and T. Bataillon, 2006 Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. *Nat. Genet.* 38: 484–488.
- Kauffman, S. A., 1993 *The Origins of Order*. Oxford University Press, New York.
- Macken, C. A., and A. S. Perelson, 1989 Protein evolution on rugged landscapes. *Proc. Natl. Acad. Sci. USA* 86: 6191–6195.
- MacLean, R., and A. Buckling, 2009 The distribution of fitness effects of beneficial mutations in *Pseudomonas aeruginosa*. *PLoS Genet.* 5: e1000406.
- Mathews, J., and R. L. Walker, 2004 *Mathematical Methods of Physics*. Pearson Education, Delhi, India.
- McDonald, M., T. F. Cooper, H. J. E. Beaumont, and P. B. Rainey, 2010 The distribution of fitness effects of new beneficial mutations in *Pseudomonas fluorescens*. *Biol. Lett.* 7: 98–100.
- Miller, C. R., P. Joyce, and H. Wichman, 2011 Mutational effects and population dynamics during viral adaptation challenge current models. *Genetics* 187: 185–202.
- Neidhart, J., and J. Krug, 2011 Adaptive walks and extreme value theory. *Phys. Rev. Lett.* (in press).
- Orr, H., 2003 A minimum on the mean number of steps taken in adaptive walks. *J. Theor. Biol.* 220: 241–247.
- Orr, H. A., 2002 The population genetics of adaptation: the adaptation of DNA sequences. *Evolution* 56: 1317–1330.
- Orr, H. A., 2006 The population genetics of adaptation on correlated fitness landscapes: the block model. *Evolution* 60: 1113.
- Perelson, A., and C. Macken, 1995 Protein evolution on partially correlated landscapes. *Proc. Natl. Acad. Sci. USA* 92: 9657–9661.

- Rokyta, D., P. Joyce, S. Caudle, and H. Wichman, 2005 An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. *Nat. Genet.* 37: 441–444.
- Rokyta, D., C. J. Beisel, P. Joyce, M. T. Ferris, C. L. Burch *et al.*, 2008 Beneficial fitness effects are not exponential for two viruses. *J. Mol. Evol.* 69: 229.
- Rokyta, D., Z. Abdo, and H. Wichman, 2009 The genetics of adaptation for eight microvirid bacteriophages. *J. Mol. Evol.* 69: 229.
- Rosenberg, N., 2005 A sharp minimum on the mean number of steps taken in adaptive walks. *J. Theor. Biol.* 237: 17–22.
- Rozen, D., J. de Visser, and P. J. Gerrish, 2002 Fitness effects of fixed beneficial mutations in microbial populations. *Curr. Biol.* 12: 1040–1045.
- Sanjuán, R., A. Moya, and S. Elena, 2004 The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proc. Natl. Acad. Sci. USA* 101: 8396–8401.
- Schaeybroeck, B., and A. Lazarides, 2009 Normal-superfluid interface for polarized fermion gases. *Phys. Rev. A* 79: 053612.
- Schoustra, S., T. Bataillon, D. Gifford, and R. Kassen, 2009 The properties of adaptive walks in evolving populations of fungus. *PLoS Biol.* 7(11): e1000250.
- Sornette, D., 2000 *Critical Phenomena in Natural Sciences*. Springer-Verlag, Berlin.
- Weinberger, E. D., 1991 Local properties of Kauffman's N-k model: a tunably rugged energy landscape. *Phys. Rev. A* 44: 6399–6413.

Communicating editor: L. M. Wahl

Appendix A: Random Adaptive Walk

In this *Appendix*, we briefly review the known results for random adaptive walk in which all better mutants are chosen with equal probability (Macken and Perelson 1989; Flyvbjerg and Lautrup 1992; Kauffman 1993). The probability distribution $P_J(f)$ obeys the recursion relation

$$P_{J+1}(f) = \int_l^f dh \frac{p(f)}{\int_h^u dg p(g)} [1 - q^L(h)] P_J(h) \quad (A1)$$

(Flyvbjerg and Lautrup 1992), where $q(f) = \int_l^f dg p(g)$. A change of variable from the fitness f to the cumulative probability $q(f)$ gives

$$P_{J+1}(q) = \int_0^q dq' \frac{1 - q'^L}{1 - q'} P_J(q'). \quad (A2)$$

Since the walk length distribution for the random adaptive walk also obeys (17), we have

$$Q_J = \int_l^u dh q^L(h) P_J(h) = \int_l^u dq q^L P_J(q), \quad (A3)$$

which shows that Q_J is a *universal distribution* in that it is independent of the underlying fitness distribution $p(f)$. Note that for infinitely long sequences, the probability $Q_J = 0$ as in the mutational landscape model. Differentiating (62) with respect to q immediately gives

$$\frac{dP_{J+1}(q)}{dq} = \frac{1 - q^L}{1 - q} P_J(q) = \sum_{n=0}^L q^n P_J(q). \quad (A4)$$

The generating function $G(x, q) = \sum_{J=1}^{\infty} x^J P_J(q)$ then obeys the following *first-order* differential equation:

$$G'(x, q) - x P_1'(q) = x \frac{1 - q^L}{1 - q} G(x, q). \quad (A5)$$

For the initial condition $P_0(f) = \delta(f)$, we have $P_1(q) = 1$ and due to (A2), the distribution $P_J(0) = 0$. Solving the above differential equation using these boundary conditions gives $G(x, q) = x e^{x H_L(q)}$, where $H_L(q) = \sum_{k=1}^L q^k / k$ and hence the distribution $P_J(q)$ is given by

$$P_J(q) = \frac{H_L^{J-1}(q)}{(J-1)!} \quad (A6)$$

(Flyvbjerg and Lautrup 1992). Since the product $q^L P_J(q)$ in (A3) peaks around $q = 1$, using $H_L(q) \approx \ln L$ for q close to unity for finite but long sequences and performing the integral in (A3), we get

$$Q_J \approx e^{-\bar{J}} \frac{\bar{J}^{J-1}}{(J-1)!}, \quad (\text{A7})$$

where $\bar{J} = \ln L$. Thus the walk length distribution is a Poisson distribution (in J) with mean $\bar{J} = \ln L$ (Flyvbjerg and Lautrup 1992).

Appendix B: Simulation Procedure

For short sequences of length $L \leq 10$ and uncorrelated fitnesses, a randomly chosen sequence was assigned a fitness equal to zero. Then the rest of the fitness landscape composed of $2^L - 1$ fitnesses was generated by drawing random variables independently from a common distribution $p(f)$. The transition probability from the initial sequence to each of the better sequences among the L nearest neighbors was calculated according to (4) and the fixed sequence at the first step in the adaptive walk was chosen. Then the transition probability from the chosen mutant sequence to its better neighbors was calculated and this process was repeated until a fitter sequence was not available.

To simulate sequences with length $L \geq 10^2$, we followed an approximate procedure outlined in Orr (2002) as the total number of sequences 2^L is prohibitively large for long sequences. Starting with zero fitness, L i.i.d. random variables were generated and a higher fitness f was chosen according to the transition probability (4). During the next step in the process, L new i.i.d. random variables were generated and the transition probability from f to a better fitness was calculated. These steps were repeated until the new set of random fitnesses did not exceed the currently fixed fitness. The block model was simulated to generate weakly correlated fitnesses by assigning independent fitnesses to each block sequence. In all the simulations, the data were collected using 10^6 independent realizations of the fitness landscape.

Appendix C: Derivations for Uniformly Distributed Fitness

Solution of Differential Equation 49

The generating function $G_{<}(x, f)$ obeys the inhomogeneous second-order differential equation

$$G''(x, f) - \frac{2x}{(1-f)^2} G(x, f) = \frac{4f}{(1-f)^2}, \quad (\text{C1})$$

where we have dropped the subscript for brevity. The general solution of such differential equations is a linear combination of the general solution $G_H(x, f)$ of the homogeneous equation obtained by setting the right-hand side equal to zero and the particular solution G_P of the inhomogeneous equation (Mathews and Walker 2004). The homogeneous solution is of the form

$$G_H(x, f) = a_+(1-f)^{\alpha_+} + a_-(1-f)^{\alpha_-}, \quad (\text{C2})$$

where $\alpha_{\pm}$ are the solutions of the quadratic equation $\alpha^2 - \alpha - 2x = 0$ and given by (51). The particular solution is found using the method of variation of parameters and is of the form $G_P(x, f) = u_+(x)(1-f)^{\alpha_+} + u_-(x)(1-f)^{\alpha_-}$, where the functions $u_{\pm}(f)$ obey the following first-order differential equations:

$$u'_+(f)(1-f)^{\alpha_+} + u'_-(f)(1-f)^{\alpha_-} = 0 \quad (\text{C3})$$

$$\alpha_+ u'_+(f)(1-f)^{\alpha_+-1} + \alpha_- u'_-(f)(1-f)^{\alpha_--1} = \frac{4f}{(1-f)^2} \quad (\text{C4})$$

(Mathews and Walker 2004). On solving the above equations, we obtain

$$G_P(x, f) = \frac{4}{\alpha_+ \alpha_-} - \frac{4(1-f)}{(1-\alpha_+)(1-\alpha_-)} = \frac{-2f}{x}. \quad (\text{C5})$$

Finally, using the boundary conditions in the general solution $G_{<}(x, f) = G_P(x, f) + G_H(x, f)$, the desired result (Equation 52) is obtained.

Distribution of Fixed Beneficial Mutations

The fitness distribution found using (52) and (53) is given below for the first few adaptive steps:

$$P_1(f) = 2f, \quad f \leq 1 \quad (\text{C6})$$

$$P_2(f) = \begin{cases} -8f + 4(f-2)\ln(1-f), & f \leq \tilde{f} \\ \frac{4\tilde{f}(f+\tilde{f}-2)}{1-\tilde{f}} + 4(f-2)\ln(1-\tilde{f}), & f > \tilde{f} \end{cases} \quad (C7)$$

$$P_3(f) = 4 \begin{cases} 12f + \ln(1-f)(12-6f+f\ln(1-f)), & f \leq \tilde{f} \\ \frac{1}{1-\tilde{f}} [6\tilde{f}(2-f-\tilde{f}) + 2(6-(6-\tilde{f})\tilde{f}-f(3-2\tilde{f}))\ln(1-\tilde{f}) \\ + f(1-\tilde{f})\ln^2(1-\tilde{f})], & f > \tilde{f} \end{cases} \quad (C8)$$

$$P_4(f) = \frac{-8}{3} \begin{cases} 120f + 60(2-f)\ln(1-f) + 12f\ln^2(1-f) + (2-f)\ln^3(1-f), & f \leq \tilde{f} \\ \frac{1}{1-\tilde{f}} [60\tilde{f}(2-f-\tilde{f}) + 12(f(5-3\tilde{f}) - 2(5-(5-\tilde{f})\tilde{f}))\ln(1-\tilde{f}) \\ + 3(f(2-3\tilde{f}) + (2-\tilde{f})\tilde{f})\ln^2(1-\tilde{f}) + (2-f)(1-\tilde{f})\ln^3(1-\tilde{f})], & f > \tilde{f}. \end{cases} \quad (C9)$$

Walk Length Distribution

On matching powers of x^J on both sides in (55), we get

$$Q_1 = e^{-2\ell}(-1 + 2e^\ell) \quad (C10)$$

$$Q_2 = 2e^{-2\ell}(3 + \ell + (-3 + 2\ell)e^\ell) \quad (C11)$$

$$Q_3 = e^{-2\ell}[-2(18 + 8\ell + \ell^2) + 4e^\ell(9 - 5\ell + \ell^2)] \quad (C12)$$

$$Q_4 = \frac{4e^{-2\ell}}{3}[180 + 84\ell + 15\ell^2 + \ell^3 + e^\ell(-180 + 96\ell - 21\ell^2 + 2\ell^3)], \quad (C13)$$

where $\ell = \ln L$. A general solution of Q_J by this method does not seem possible but an approximate analytic expression for Q_J can be obtained as explained below.

From the definition of the generating function $H(x)$ in (55), it follows that

$$Q_J = \frac{1}{J!} \left. \frac{d^J H(x)}{dx^J} \right|_{x=0}. \quad (C14)$$

By the residue theorem for complex variables, we have

$$\frac{1}{2\pi i} \int_C dz f(z) = \frac{1}{n!} \left. \frac{d^n}{dz^n} ((z-z_0)^{n+1} f(z)) \right|_{z=z_0} \quad (C15)$$

(Mathews and Walker 2004), where z_0 is a pole of order $n + 1$ of the function $f(z)$ and the contour C encloses the singularities of $f(z)$. From (C14) and (C15), we can write

$$Q_J = \frac{1}{2\pi i} \int_C dz \frac{H(z)}{z^{J+1}} = \frac{1}{2\pi i} \int_C dz e^{K(z)}, \quad (C16)$$

where $K(z) = \ln H(z) - (J + 1) \ln z$. We solve this integral by the method of steepest descent, which for large J gives

$$Q_J \approx \sqrt{\frac{1}{2\pi K''(z_s)}} e^{K(z_s)} = \sqrt{\frac{1}{2\pi K''(z_s)}} \frac{H(z_s)}{z_s^{J+1}} \quad (C17)$$

(Mathews and Walker 2004), where the prime refers to the derivative with respect to z . In the above equation, z_s is a solution of the equation

$$\frac{H'(z_s)}{H(z_s)} = \frac{J}{z_s} \quad (\text{C18})$$

and

$$K''(z_s) = \left(\frac{H'(z)}{H(z)} \right)' \bigg|_{z=z_s} + \frac{J}{z_s^2} \quad (\text{C19})$$

$$= \left(\frac{H'(z)}{H(z)} \right)' \bigg|_{z=z_s} + \frac{1}{z_s} \frac{H'(z_s)}{H(z_s)}, \quad (\text{C20})$$

where the prime denotes a derivative with respect to z . Since $\alpha_+ > 0$, neglecting the exponentially small term in $(1-\tilde{f})^{\alpha_+}$ in (55), we get

$$H(z) \approx \frac{e^{-3\ell/2} e^{y/2} (3+y)(y^2-1)}{16y}, \quad (\text{C21})$$

where $y = \sqrt{1+8z}$. Differentiating $H(z)$ once with respect to z gives

$$\frac{H'(z)}{H(z)} \approx \frac{8(y+3) + 4(2y+3)(y^2-1) + 2y(y+3)(y^2-1)\ell}{y^2(y^2-1)(y+3)}. \quad (\text{C22})$$

Using the above expression in (C18) for large y , we get $y_s \approx 4J/\ell$ and therefore

$$z_s \approx \frac{2J^2}{\ell^2}. \quad (\text{C23})$$

On differentiating (C22) once, we have

$$\left(\frac{H'(z)}{H(z)} \right)' \approx \frac{4}{y} \left[\frac{4}{3(y+3)^2} + \frac{4}{(1+y)^2} - \frac{4+6\ell}{3y^2} + \frac{8}{y^3} - \frac{4}{(1-y)^2} \right]. \quad (\text{C24})$$

Using (C22) and (C24) in (C20), we obtain

$$K''(z_s) \approx \frac{8 \left[-36 + 6y_s(y_s^2-3) + y_s(y_s+3)^2(1+y_s)^2\ell \right]}{y_s^4(y_s+3)^2(y_s^2-1)} \quad (\text{C25})$$

$$\approx \frac{8\ell}{y_s^3} = \frac{a^4}{8J^3}. \quad (\text{C26})$$

Thus we have

$$Q_J \approx \frac{2J^{3/2}}{\sqrt{\pi}\ell^2} \times \frac{2-\alpha_-(z_s)}{\alpha_+(z_s)-\alpha_-(z_s)} \times \frac{(1-\tilde{f})^{1+\alpha_-(z_s)}}{z_s^J}, \quad (\text{C27})$$

where $\alpha_{\pm}$ is given by (51).