Available online at www.sciencedirect.com
ScienceDirect
Journal of Theoretical Biology 243 (2006) 114–120
Journal of Theoretical Biology
www.elsevier.com/locate/yjtbi

Properties of adaptive walks on uncorrelated landscapes under strong selection and weak mutation

Darin R. Rokyta[a], Craig J. Beisel[b], Paul Joyce[b,c,*]

[a]Department of Biological Sciences, University of Idaho, Moscow, ID 83844, USA
[b]Department of Mathematics, University of Idaho, Moscow, ID 83844, USA
[c]Department of Statistics, University of Idaho, Moscow, ID 83844, USA

Received 30 March 2006; received in revised form 3 June 2006; accepted 7 June 2006
Available online 13 June 2006

Abstract

We examine properties of adaptive walks on uncorrelated (i.e. random) fitness landscapes starting from moderately fit genotypes under strong selection weak mutation. As an extension of Orr's model for a single step in an adaptive walk under these conditions, we show that the fitness rank of the dominant genotype in a population after the fixation of a beneficial mutation is, on average, $(i+6)/4$, where i is the fitness rank of the starting genotype. This accounts for the change in rank due to acquiring a new set of single-mutation neighbors after fixing a new allele through natural selection. Under this scenario, adaptive walks can be modeled as a simple Markov chain on the space of possible fitness ranks with an absorbing state at $i = 1$, from which no beneficial mutations are accessible. We find that these walks are typically short and are often completed in a single step when starting from a moderately fit genotype. As in Orr's original model, these results are insensitive to both the distribution of fitness effects and most biological details of the system under consideration.

Keywords: Mutational landscape; Extreme value theory; Adaptive walk; Adaptive evolution

1. Introduction

Much work has recently emerged describing adaptation in discrete sequence spaces (e.g. DNA or amino acid sequences) that suggests that molecular adaptation may be characterized by general rules that are independent of many of the biological details of the evolving system. Building on the idea of a protein space described by Maynard Smith (1962, 1970), Gillespie (1983, 1984, 1991) incorporated the use of extreme value theory to circumvent the problem of specifying the exact distribution of fitness effects for new mutations. He argued that beneficial mutations represent draws from the extreme right tail of the unknown fitness distribution, thus falling within the purview of extreme value theory, which provides many distribution-independent properties. Orr (2002, 2003a, 2005) expanded significantly on this framework, leading to empirically testable predictions (e.g. Rokyta et al., 2005; Kassen and Bataillon, 2006) for the expected progression of the first step in adaptive evolution.

Work by Gillespie and Orr concerning multiple steps in adaptation or full adaptive walks to local optima has primarily used uncorrelated fitness landscapes. These landscapes are "random" in that each sequence is assigned a fitness at random from the same fitness distribution. Thus, neighbors in sequence space do not tend to have similar fitnesses (i.e. their fitnesses are uncorrelated). This type of landscape has been studied extensively, though typically starting either from a random sequence or the sequence with the lowest fitness, and moving through sequence space by either randomly selecting from among the accessible beneficial mutations or always selecting the most fit (Macken et al., 1991; Flyvbjerg and Lautrup, 1992; Macken and Perelson, 1989; Kauffman and Levin, 1987; Orr, 2003b; Rosenberg, 2005). In contrast, we further explore the model investigated by Gllespie and Orr, the mutational landscape model, that begins from a moderately fit initial genotype that traverses the fitness landscape

*Corresponding author.
E-mail addresses: rokyta@uidaho.edu (D.R. Rokyta), beis2492@uidaho.edu (C.J. Beisel), joyce@uidaho.edu (P. Joyce).

0022-5193/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2006.06.008

according to a more realistic population genetics model. This modeling framework is particularly relevant to microbial experimental evolution studies, where its assumptions can be met, providing reasonable expectations for the outcomes of these experiments (Orr, 2002, 2005). Additionally, microbial systems allow for experimentally feasible testing of both the assumptions and predictions of the model (Rokyta et al., 2005; Kassen and Bataillon, 2006).

2. The model

Following Orr (2002), we consider the adaptation of a population of haploid DNA sequences of length L, representing a gene or a small genome. Under Gillespie's strong selection weak mutation (SSWM) conditions, exploration of sequence space is constrained to those $3L$ sequences differing from the wild type by a single mutation, and the population is effectively fixed for a single sequence at any particular time (Gillespie, 1983, 1984, 1991); adaptation proceeds as the sequential fixation of novel beneficial mutations, and clonal interference does not occur. Selection is considered strong if $Ns \gg 1$ where N is the population size and s is a typical selection coefficient; for weak mutation, it is assumed that $N\mu \ll 1$, where μ is the per site mutation rate. We imagine that the wild-type sequence is at a local fitness optimum; it has a higher fitness than all of its $3L$ single-mutation neighbors. Some environmental change results in the reassignment of fitnesses to these $3L+1$ sequences, and if any have a higher fitness than the wild type, adaptation ensues. We can proceed without knowledge of the exact form of the distribution used to assign fitnesses by assuming that the wild-type sequence remains relatively well adapted to the new environment, i.e. its new rank, i, is small (say <50) (Gillespie, 1983, 1984, 1991). The fitnesses of the $i-1$ beneficial alleles will be from the right tail of the fitness distribution, and in large samples (e.g. $3L+1$ draws for a moderately large L) the extreme values take on distribution-independent properties. Orr (2002) characterized this model for the fixation of a single beneficial mutation.

After the fixation of the first beneficial mutation, a new region of sequence space becomes accessible to the population. We assume that the $3L-1$ new neighboring sequences (this number excludes the original wild-type sequence) are assigned fitnesses from the same distribution used to generate the initial fitnesses. This corresponds to an uncorrelated fitness landscape, where the fitness of neighboring sequences are random with respect to each other. The environment remains constant, but the population is now exploring a previously inaccessible region of sequence space. The sequence currently fixed by the population may now have a new rank, depending upon the number of single-mutation neighbors with fitness values larger than its own (Fig. 1). This process of fixing a new beneficial mutation followed by acquiring new accessible sequences is repeated until the population achieves rank 1 relative to its new neighbors, meaning that no beneficial mutations are available and the population has reached a local optimum. This scenario has been examined through simulations by Gillespie (1991) and Orr (2002).

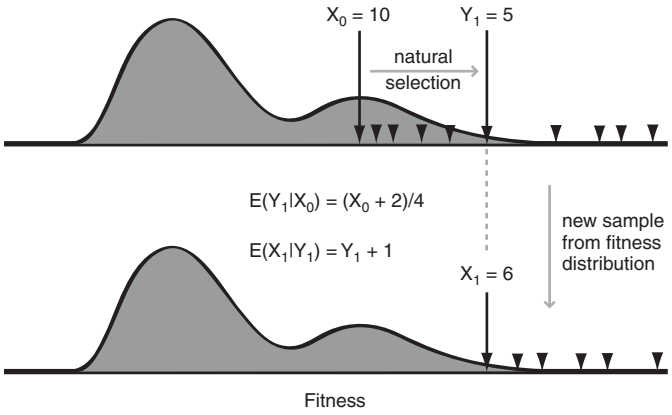

Fig. 1. A schematic illustration of a full step in an adaptive walk on an uncorrelated fitness landscape. After some change in environment, fitnesses are assigned to all accessible sequences from some unknown distribution. In this example, the population begins with initial rank $X_0 = 10$, and the fitnesses of the single-mutation neighboring sequences with higher fitnesses are designated with triangles. The population fixes a new allele through natural selection to decrease its rank to $Y_1 = 5$. This new sequence's neighbors are assigned fitnesses from the same distribution as for the original set, and in this example, 5 sequences have higher fitness, giving the new rank $X_1 = 6$.

3. Results

3.1. Orr's mean transition probabilities

Orr (2002) characterized the expected behavior for the first step in adaptation for a population under the conditions described above. If the initial wild-type sequence has a moderate fitness rank relative to its single-mutation neighbors, the selection coefficients for the accessible beneficial mutations in rank order, $\mathbf{s} = (s_1, s_2, \ldots, s_{i-1})$, can be calculated based on the spacings between the extreme draws from the tail of the fitness distribution. The neighboring sequence with the highest fitness relative to the wild type has rank 1 and selection coefficient s_1, the sequence with the second highest fitness has rank 2 and selection coefficient s_2, etc. Assuming SSWM conditions and that the fixation probability for allele j is approximately $2s_j$ (Haldane, 1927), Orr (2002) found that the expected distribution for the change in fitness rank for the first step in adaptive evolution is given by

$$E_{\mathbf{s}}(P_{ij}(\mathbf{s})) = E_{\mathbf{s}}\left(\frac{s_j}{\sum_{k=1}^{i-1} s_k}\right) = \frac{1}{i-1}\sum_{k=j}^{i-1}\frac{1}{k}, \quad (1)$$

where i is the current fitness rank and $j<i$ are the accessible beneficial mutations. These transition probabilities, $P_{ij}(\mathbf{s})$, describe the probabilities of moving from the wild-type

allele with rank i to the sequences with higher fitnesses (ranks $j<i$) as a function of $\mathbf{s}$. The mean transition probabilities, $E_{\mathbf{s}}(P_{ij}(\mathbf{s}))$, average over the selection coefficients, assuming the fitness values are the extreme order statistics from a fitness distribution in the Gumbel domain of attraction, which includes most commonly encountered distributions such as the Normal, Gamma, and Exponential (Orr, 2002, 2003a; Gumbel, 1958; Gillespie, 1991). On average, fitness rank is decreased in a single step from i to $(i+2)/4$ through natural selection. The variance in rank change is given by $(i-2)(7i+6)/144$. Let Y be a random variable where

$$P(Y=j)=\frac{1}{i-1}\sum_{k=j}^{i-1}\frac{1}{k} \tag{2}$$

and

$$E(Y)=\frac{i+2}{4}, \tag{3}$$

then we say the $Y\sim\mathrm{Orr}(i)$, where i is the current fitness rank.

3.2. The distribution of the number of exceedances

The second component of the model describes the change in fitness rank due to the acquisition of a new set of neighboring sequences. This change depends on the distribution of the number of values larger than the current fitness in a new sample of fitness values, i.e. the number of exceedances in a new sample. The asymptotic form of this distribution was described by Gumbel and von Schelling (1950). Assuming $3L\approx(3L-1)$ is large, the number of exceedances, X, has a negative binomial distribution, regardless of the original distribution, such that

$$P(X=x)=\binom{x+j-1}{x}\left(\frac{1}{2}\right)^{x+j}, \tag{4}$$

where j is the starting fitness rank, and x is the number of draws larger than the current value. Since $E(X)=j$, fitness rank increases by 1 on average, since the rank is the number of exceedances plus 1. The variance is given by $\mathrm{Var}(X)=2j$. We will denote this distribution as $X\sim\mathrm{NegBin}(j)$. This result makes no assumptions about the form of the fitness distribution; it is only necessary that the sample size is large and constant across samples. Interestingly, Eq. (4) has a simple probabilistic interpretation in terms of coin tosses. If a fair coin is repeatedly tossed until exactly j heads occur, it gives the distribution of the number of tails accumulated before the jth head.

To illustrate this idea in terms of a DNA sequence space, consider a sequence of length $L=1000$. Initially, the wild type sequence has some rank, say $i=10$, i.e. if the fitnesses of the wild type and its 3000 single-mutation neighbors were listed in increasing size, the wild-type's fitness would be 10th from the top. Now let us say that natural selection moves the population up the list to the sequence with rank $j=5$. The new sequence had rank 5 among the 3001 sequences including the original wild type and its single-mutation neighbors, but now those sequences, except for the original wild type, differ from the new wild type by two mutations and are thus no longer accessible. It now has a new list of 2999 accessible sequences plus the original wild type, and thus potentially a new fitness rank. We can imagine that these 2999 sequences are all assigned fitnesses from the same distribution as for the original set and are again listed in increasing order. Eq. (4) describes the probability distribution for the new rank given that the current sequence had rank j in the first set of draws from the fitness distribution. In our example, the fitness rank is, on average, expected to increase from 5 to 6. This example is illustrated in Fig. 1. This result is independent of the distribution used to assigned fitness, as long as it remains the same for each set of sequences and the rank remains small relative to the number of draws.

3.3. One complete step

We can begin to characterize the combined process, including both the effects of natural selection and of acquiring new single-mutation neighbors, by calculating the expected change in rank. Define $X_0=i$ as the rank of the initial wild-type sequence. Then let $Y_1\sim\mathrm{Orr}(X_0)$ be the rank after the fixation of a mutation through natural selection and X_1 be the rank of that new sequence relative to its new neighbors in sequence space. Then $E(X_1|Y_1)=Y_1+1$, and thus

$$E(X_1)=E(E(X_1|Y_1))=E(Y_1)+1=\frac{i+6}{4}, \tag{5}$$

where $E(Y_1)$ is given by Eq. (3). Natural selection decreases the fitness rank on average from i to $(i+2)/4$, as described by Orr (2002). In the presence of new neighbors, however, the fitness rank increases by 1, on average (Fig. 1). Note that when the initial rank is 2 (i.e. only a single beneficial mutation is accessible), the expected new rank after a complete step is also 2, and fitness rank does not change on average. Even though the landscape is random, adaptation in terms of fitness ranks follows a very simple rule. It might be expected that the random nature of the landscape might make adaptation highly unpredictable, yet, in fact, a population's fitness rank does not change much on average from acquiring new neighboring sequences. The variance of this process is given by

$$\begin{aligned}\mathrm{Var}(X_1)&=E(\mathrm{Var}(X_1|Y_1))+\mathrm{Var}(E(X_1|Y_1))\\&=\frac{i+2}{2}+\frac{(i-2)(7i+6)}{144}\end{aligned} \tag{6}$$

for $i\geqslant 2$. Note that if $i=2$, then $Y_1=1$ with probability 1, thus there is no variability associated with the action of natural selection, but there is still variability associated with acquiring new accessible sequences. The variance is increased relative to the variance of the rank change due to just natural selection by a factor of $(i+2)/2$.

In addition, we can obtain the full one step transition probabilities for the combined process, which incorporate the change in rank due to both natural selection and to acquiring new neighbors in sequence space (Fig. 1). Using the Chapman–Kolmogorov equation (Karlin and Taylor, 1975), we find that the transition probabilities are given by

$$P(X_1 = x|X_0 = i) = \frac{1}{i-1}\left(\frac{1}{2}\right)^{x-1}\sum_{j=1}^{i-1}\binom{x+j-2}{x-1}\left(\frac{1}{2}\right)^{j}\sum_{k=j}^{i-1}\frac{1}{k}, \quad (7)$$

where i is the initial fitness rank, and x is the new fitness rank after fixing a beneficial mutation through natural selection and acquiring a new set of accessible single-mutation neighbors. To move from rank i to rank x, the population's fitness rank is first decreased through natural selection according to Eq. (2) to some rank j. The rank then changes from j to x according to Eq. (4). To get the total probability of changing from rank i to x, it is necessary to sum the probabilities over all $i-1$ intermediate states j. This describes a simple discrete space, discrete time Markov chain.

The probability that a fitness peak is reached in a single step given the initial fitness rank can be found by setting $x = 1$ in Eq. (7). However, we can also derive this probability using properties of the negative binomial distribution. This probability has an explicit relationship to the probability of fixing the most fit allele through natural selection as derived by Orr (2002) and given in Eq. (2). Let $X_0 = i$ be the rank of the original wild type, and $Y_1 \sim \mathrm{Orr}(X_0)$ be the rank of the first step mutant relative to the original wild type, and X_1 be the rank of the first mutant relative to its new neighbors, then

$$\begin{aligned} P(X_1 = 1|X_0 = i) &= P(Y_1 = 1|X_0 = i) - \frac{\ln 2}{i-1} \\ &\quad + \frac{1}{i-1}\int_0^{1/2}\frac{t^{i-1}}{1-t}\,\mathrm{d}t \\ &= \frac{1}{i-1}\sum_{j=1}^{i-1}\frac{1}{j} - \frac{\ln 2}{i-1} + \frac{1}{i-1}\int_0^{1/2}\frac{t^{i-1}}{1-t}\,\mathrm{d}t. \end{aligned} \quad (8)$$

Note that $\int_0^{1/2} t^{i-1}/(1-t)\,\mathrm{d}t \leqslant (2/i)(1/2)^i$ and $\int_0^{1/2} t^{i-1}/(1-t)\,\mathrm{d}t \geqslant (1/i)(1/2)^i$ which is negligibly small for $i > 10$. Also $P(Y_1 = 1|X_0 = i) = (1/(i-1)) \quad \sum_{j=1}^{i-1}(1/j) \approx (\ln(i-1))/(i-1)$. If we ignore the last term in the above equation and use the log approximation, we get $P(X_1 = 1|X_0 = i) \approx (\ln((i-1)/2))/(i-1)$. See Appendix A for the derivation. Based on Eq. (8), half of the walks starting from rank 2, and thus necessarily fixing the allele of rank 1, will find no new neighboring sequences with a higher fitness. Even with an initial rank of 10, nearly a quarter of adaptive walks will involve only a single substitution, and beginning at rank 50, 8% of walks will be complete after a single step. Thus, we find that adaptive walks consisting of a single step are expected to be common under this model, even when beginning at a moderately high initial rank (Fig. 2). As noted above, under Orr's original formulation of this model (Orr, 2002), it is possible to calculate the probability of reaching rank 1 in a single step using Eq. (2). However, this formulation does not account for the possibility that the newly fixed sequence might have higher fitness neighbors that were inaccessible to the original wild type through a single mutation.

3.4. Multiple steps

Also of interest is the behavior of this process beyond the first step. Define X_n be the rank at the nth step of an adaptive walk. We can describe the conditional distribution of $X_n|X_{n-1}$ by observing that $Y_n|X_{n-1}$ is distributed $\mathrm{Orr}(X_{n-1})$, where Y_n refers to the new rank after natural selection for the nth step, and $X_n|Y_n$ is distributed $1 + \mathrm{NegBin}(Y_n)$ provided that $X_{n-1} > 1$. If $X_{n-1} = 1$ then $X_n = 1$ as well. By using the above conditional distributions and considering the two cases, $X_{n-1} > 1$ and $X_{n-1} = 1$, we find

$$E(X_n) = 1 + \frac{i-1}{4^n} + \frac{3}{4}\sum_{j=0}^{n-1}\left(\frac{1}{4}\right)^{j} P(X_{n-j} > 1), \quad (9)$$

and there exists a constant α, where $\frac{1}{4} < \alpha < 1$

$$0 < \frac{3}{4}\sum_{j=0}^{n-1}\left(\frac{1}{4}\right)^{j} P(X_{n-j} > 1) \leqslant \alpha^{n+1}\frac{4}{4\alpha - 1}. \quad (10)$$

Thus, $\lim_{n\to\infty} E(X_n) = 1$, and, in fact, the convergence is geometrically fast. The full derivation is provided in Appendix B. This geometric rate of convergence of $E(X_n)$ to 1 suggests, albeit indirectly, that rank 1 should be reached quickly. An explicit formula for $E(X_n)$ is more

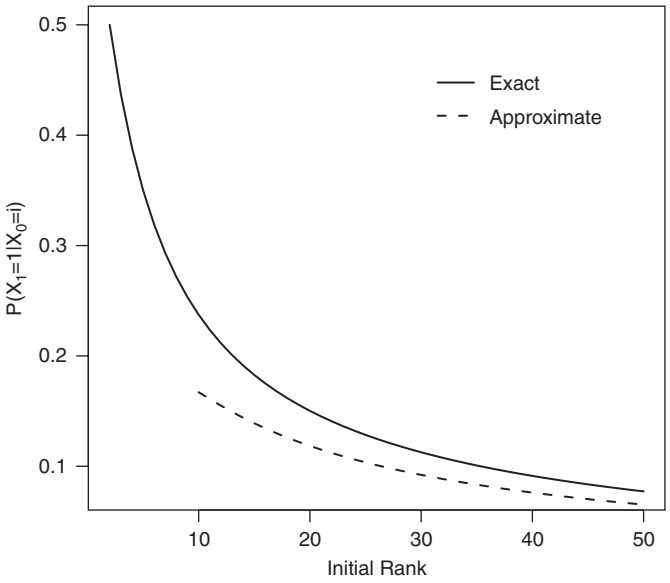

Fig. 2. The probability that an adaptive walk is complete in a single step as a function of the initial fitness rank of the wild type. The solid line gives the exact values based on Eq. (8), and the dashed line gives the values based upon a log approximation. The approximate form is not valid for small initial ranks.

elusive, as it requires that we know $P(X_j = 1)$ for all $j = 1, \ldots, n$. This issue is discussed in more detail in a remark at the end of Appendix B.

3.5. Absorption time

Adaptation will continue until the population becomes fixed for a sequence that has a higher fitness than all of its single-mutation neighbors, i.e. it reaches rank 1. Let T_i be the number of beneficial mutations fixed by a population until a local optimum is reached, starting from rank i, and, as before, X_n is the rank after n steps of the process conditional on starting at rank $X_0 = i$. The number of steps taken is given by

$$T_i = \sum_{n=0}^{\infty} I_{\{X_n > 1\}}, \tag{11}$$

where $I_{\{A\}}$ denotes the indicator function for the event A. Therefore, the distribution of T_i and, consequently the mean walk length, $E(T_i)$, depends on knowing $P(X_n = 1)$ for all n, which we were unable to derive explicitly. However, if $P(X_n = 1)$ converges to 1 at a geometric rate, then in practice we only need to approximate $P(X_n = 1)$ for a handful of values of n. A geometric rate of convergence is guaranteed if there is a positive probability that X_1 is equal to 1, i.e. if $P(X_1 = 1) > 0$, which holds for the model under consideration. This implies a geometric rate of convergence since, $P(X_n > 1) \leqslant (P(X_1 > 1))^n$. This inequality holds since X_n is stochastically decreasing, i.e. $P(X_n < x) \leqslant P(X_{n+1} < x)$. This can be understood intuitively as follows. Consider two processes, where the first behaves according to the rules we have established, and the second moves one step according to our rules, but if rank 1 is not reached, then it restarts at rank i. Because the second process always starts over it takes, on average, more steps to reach rank 1 than for the first process. The second process reaches rank 1 according to the geometric distribution, i.e. the probability of not reaching rank 1 in n steps is equal to $P(X_1 > 1)^n$. Therefore, the above equation provides a bound on the mean walk length as follows:

$$E(T_i) = \sum_{n=0}^{\infty} P(X_n > 1) \leqslant \sum_{n=0}^{\infty} P(X_1 > 1)^n = 1/P(X_1 = 1), \tag{12}$$

where $P(X_1 = 1) \equiv P(X_1 = 1 | X_0 = i)$ is given by Eq. (8). While this typically provides a very crude bound on $E(T_i)$, we note that it follows from geometric convergence that

$$\sum_{n=k}^{\infty} P(X_n > 1) \leqslant P(X_1 > 1)^{k+1}/P(X_1 = 1). \tag{13}$$

Thus, $E(T_i)$ can be approximated as accurately as desired by starting with an $\varepsilon > 0$ and choosing k so that $P(X_1 > 1)^{k+1}/P(X_1 = 1) < \varepsilon$ and

$$E(T_i) \approx \sum_{n=1}^{k-1} P(X_n > 1). \tag{14}$$

As before, we have no explicit formula for $P(X_n > 1)$.

We can calculate $\mathrm{Var}(T_i)$ by noting that $\mathrm{Var}(T_i) = E(T_i^2) - E(T_i)^2$ and that $I_{\{X_n > 1\}} I_{\{X_k > 1\}} = I_{\{X_n > 1\}}$ whenever $n > k$. Thus

$$\begin{aligned} E(T_i^2) &= E\left(\sum_{k=0}^{\infty} \sum_{n=0}^{\infty} I_{\{X_n > 1\}} I_{\{X_k > 1\}} \right) \\ &= \sum_{n=0}^{\infty} P(X_n > 1) + 2 \sum_{n=1}^{\infty} \sum_{k=0}^{n-1} P(X_n > 1) \\ &= E(T_i) + 2 \sum_{n=1}^{\infty} n P(X_n > 1). \end{aligned} \tag{15}$$

Therefore,

$$\mathrm{Var}(T_i) = E(T_i) - (E(T_i))^2 + 2 \sum_{n=1}^{\infty} n P(X_n > 1). \tag{16}$$

4. Discussion

We have examined some properties of adaptation under the mutational landscape model. Our results build on work by Orr (2002) and explore the process of adaptation beyond the fixation of a single-beneficial mutation. Orr's work showed that the movement of a population for a single step due to natural selection follows some surprisingly simple rules; we have found similar results for the process that follows this fixation event. Although intuition might suggest that adaptation on a random fitness landscape would be highly unpredictable, we have shown that it has a simple structure. Orr (2002) found that natural selection tends to move the population from rank i to $(i + 2)/4$, but this new rank is relative to a set of sequences which no longer represent potential single-step adaptive substitutions. After a single substitution, natural selection has a new suite of sequences from which to choose, yet the fitness rank of the current population tends to change little; it increases by 1 on average. Thus, overall, the rank in a full step decreases on average from i to $(i + 6)/4$.

We also explored some properties of entire adaptive walks. We found that walks consisting of only a single step should be common for moderate initial ranks and that the fitness rank of an evolving population converges geometrically fast to 1. We also provided a bound on the expected number of steps in an adaptive walk. Taken together, these results indicate that adaptive walks under the scenario under consideration involve a small number of substitutions. This was also shown through simulations by both Gillespie (1991) and Orr (2002). Our work, however, provides an explanation for why this should be the case and provides a more thorough description of the properties of adaptation under the mutational landscape model.

Acknowledgments

This work was supported by grants from the National Institutes of Health (P20 RR16448, NIH-R01 GM076040-01, and NIH NCRR 1P20RR016448-01) and the National

Science Foundation (NSF-DEB-0515738). Additional support for DRR was provided by the Idaho INBRE Program (NIH P20 RR16454). We would like to thank H.A. Wichman for many useful discussions and comments on this work.

Appendix A. Derivation of the probability that an adaptive walk will be complete after a single step

$$P(X_1 = 1|X_0 = i) = \frac{1}{i-1}\sum_{j=1}^{i-1}\frac{1}{j} - \frac{\ln 2}{i-1} + \frac{1}{i-1}\int_0^{1/2}\frac{t^{i-1}}{1-t}\,\mathrm{d}t.$$

Recall the distribution of Y_1 given by (2). Define the probability generating function for Y_1 by $G_i(s) = E(s^{Y_1}|X_0 = i)$, then

$$\begin{aligned} G_i(s) &= \sum_{j=1}^{i-1}\frac{1}{i-1}\sum_{k=j}^{i-1}\frac{s^j}{k} \\ &= \frac{1}{i-1}\sum_{k=1}^{i-1}\sum_{j=1}^{k}\frac{s^j}{k} \\ &= \frac{1}{i-1}\sum_{k=1}^{i-1}\frac{s}{k}\left(\frac{1-s^k}{1-s}\right) \\ &= \frac{s}{1-s}\left[\frac{1}{i-1}\sum_{k=1}^{i-1}\frac{1}{k} - \frac{1}{i-1}\sum_{k=1}^{i-1}\frac{s^k}{k}\right]. \end{aligned}$$

Recall from Eq. (2) that

$$P(Y_1 = 1|X_0 = i) = \frac{1}{i-1}\sum_{k=1}^{i-1}\frac{1}{k}$$

and note that it follows from standard calculus that

$$\sum_{k=1}^{i-1}\frac{s^k}{k} = \sum_{k=1}^{i-1}\int_0^s t^{k-1}\,\mathrm{d}t = -\ln(1-s) - \int_0^s\frac{t^{i-1}}{1-t}\,\mathrm{d}t.$$

Therefore,

$$G_i(s) = \frac{s}{1-s}\left(P(Y_1 = 1|X_0 = i) + \frac{\ln(1-s)}{i-1} + \frac{\int_0^s (t^{i-1})/(1-t)\,\mathrm{d}t}{i-1}\right). \quad (17)$$

It follows from Eq. (7) that $P(X_1 = 1|X_0 = i) = E((1/2)^{Y_1}|X_0 = i) = G_i(1/2)$. Therefore,

$$P(X_1 = 1|X_0 = i) = G_i(1/2) = P(Y_1 = 1|X_0 = i) - \frac{\ln 2}{i-1} + \frac{1}{i-1}\int_0^{1/2}\frac{t^{i-1}}{1-t}\,\mathrm{d}t.$$

Appendix B. Derivation of the expected fitness rank after *n* steps

$$E(X_n) = 1 + \frac{i-1}{4^n} + \frac{3}{4}\sum_{j=0}^{n-1}\left(\frac{1}{4}\right)^j P(X_{n-j} > 1).$$

Note that the one step transition probabilities from X_{n-1} to X_n given by Eq. (7) only apply if $X_{n-1} > 1$, otherwise $P(X_n = 1|X_{n-1} = 1) = 1$. Therefore, we need to consider $X_{n-1} > 1$ and $X_{n-1} = 1$ separately when calculating $E(X_n)$. That is,

$$\begin{aligned} E(X_n) &= E(E(X_n|X_{n-1})) \\ &= E(E(X_n I_{\{X_{n-1}>1\}}|X_{n-1})) \\ &\quad + E(E(X_n I_{\{X_{n-1}=1\}}|X_{n-1})). \end{aligned} \quad (18)$$

It follows from Eq. (5) that the first expression on the right side of (18) can be written as

$$\begin{aligned} &E(E(X_n I_{\{X_{n-1}>1\}}|X_{n-1})) \\ &\quad = E\left(\left(\frac{X_{n-1}+6}{4}\right)I_{\{X_{n-1}>1\}}\right) \\ &\quad = \frac{1}{4}E(X_{n-1}I_{\{X_{n-1}>1\}}) + \frac{3}{2}P(X_{n-1} > 1) \end{aligned} \quad (19)$$

and second expression can be written as

$$\begin{aligned} E(E(X_n I_{\{X_{n-1}=1\}}|X_{n-1})) &= E(X_{n-1}I_{\{X_{n-1}=1\}}) \\ &= P(X_{n-1} = 1) \\ &= 1 - P(X_{n-1} > 1). \end{aligned} \quad (20)$$

Similarly, using (20) we rewrite (19) as

$$\begin{aligned} &E(E(X_n I_{\{X_{n-1}>1\}}|X_{n-1})) \\ &\quad = \tfrac{1}{4}E(X_{n-1}(1 - I_{\{X_{n-1}=1\}})) + \tfrac{3}{2}P(X_{n-1} > 1) \\ &\quad = \tfrac{1}{4}E(X_{n-1}) - \tfrac{1}{4} + \tfrac{3}{4}P(X_{n-1} > 1). \end{aligned} \quad (21)$$

Substituting Eqs. (21) and (20) into (18) gives

$$E(X_n) = \tfrac{1}{4}E(X_{n-1}) + \tfrac{3}{4} + \tfrac{3}{4}P(X_{n-1} > 1).$$

Starting with $X_0 = i$ and proceeding inductively gives

$$\begin{aligned} E(X_n) &= \frac{i}{4^n} + \frac{3}{4}\sum_{j=0}^{n-1}\left(\frac{1}{4}\right)^j + \frac{3}{4}\sum_{j=0}^{n-1}\left(\frac{1}{4}\right)^j P(X_{n-j} > 1) \\ &= 1 + \frac{i-1}{4^n} + \frac{3}{4}\sum_{j=0}^{n-1}\left(\frac{1}{4}\right)^j P(X_{n-j} > 1). \end{aligned}$$

Choose $\alpha = \max\{P(X_1 > 1), 1/3\}$ and note that it follows from discussion following Eq. (10) that $P(X_n > 1) \leqslant P(X_1 > 1)^n \leqslant \alpha^n$. Therefore,

$$\begin{aligned} \frac{3}{4}\sum_{j=0}^{n-1}\left(\frac{1}{4}\right)^j P(X_{n-j} > 1) &\leqslant \frac{3}{4}\sum_{j=0}^{n-1}\left(\frac{1}{4}\right)^j \alpha^{n-j} \\ &= \frac{3}{4}\sum_{j=0}^{n-1}\alpha^n\left(\frac{1}{4\alpha}\right)^j \leqslant \alpha^{n+1}\frac{4}{4\alpha-1}. \end{aligned}$$

Remark. Note that an explicit formula for $E(X_n)$ requires an explicit expression for $P(X_j = 1)$ for $j = 1, \ldots, n$. To calculate this probability we must consider every set of single step mutations requiring j steps to reach rank 1. To get a feel for why this is a rather intractable calculation, the following example might be helpful. For example, starting at rank $i = 20$ and $j = 5$, then one possible five step path from rank 20 to rank 1 would be to first move from rank 20 to 15, then 15 to 9, 9 to 10, 10 to 3, and finally 3 to 1. The probability of this path would be the product of 5 terms, and this path is only one of an infinite number of paths. To get the full probability, it would be necessary to sum over all possible pathways. Below is the general formula for calculating $P(X_j = 1)$ starting at rank i, where each one step probability $P_{k_{j-1}k_{j-2}} = P(X_{j-1} = k_{j-1} | X_{j-2} = k_{j-2})$ is calculated using Eq. (7)

$$P(X_j = 1) = \sum_{k_{j-1}=1}^{\infty} \sum_{k_{j-2}=1}^{\infty} \cdots \sum_{k_1=1}^{\infty} P_{ik_1} P_{k_1k_2} \cdots P_{k_{j-2}k_{j-1}} P_{k_{j-1}1}.$$

References

Flyvbjerg, H., Lautrup, B., 1992. Evolution in a rugged fitness landscape. Phys. Rev. A 46, 6714–6723.

Gillespie, J.H., 1983. A simple stochastic gene substitution model. Theor. Popul. Biol. 23, 202–215.

Gillespie, J.H., 1984. Molecular evolution over the mutational landscape. Evolution 38, 1116–1129.

Gillespie, J.H., 1991. The Causes of Molecular Evolution. Oxford University Press, New York.

Gumbel, E.J., 1958. Statistics of Extremes. Columbia University Press, New York.

Gumbel, E.J., von Schelling, H., 1950. The distribution of the number of exceedances. Ann. Math. Stat. 21, 247–262.

Haldane, J.B.S., 1927. A mathematical theory of natural and artificial selection. V. Selection and Mutation. Proc. Cambridge Philos. Soc. 23, 838–844.

Karlin, S., Taylor, H.M., 1975. A First Course in Stochastic Processes. Academic Press, New York.

Kassen, R., Bataillon, T., 2006. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nat. Gen. 38, 484–488.

Kauffman, S., Levin, S., 1987. Towards a general theory of adaptive walks on rugged landscapes. J. Theor. Biol. 128, 11–45.

Macken, C.A., Perelson, A.S., 1989. Protein evolution on rugged landscapes. Proc. Natl Acad. Sci. USA 86, 6191–6195.

Macken, C.A., Hagan, P.S., Perelson, A.S., 1991. Evolutionary walks on rugged landscapes. SIAM J. Appl. Math. 51, 799–827.

Maynard Smith, J., 1962. In: Good, I.J. (Ed.), The Scientist Speculates: An Anthology of Partly-baked Ideas. Basic Books, Inc., New York, pp. 252–256.

Maynard Smith, J., 1970. Natural selection and the concept of a protein space. Nature 225, 563–564.

Orr, H.A., 2002. The population genetics of adaptation: the adaptation of DNA sequences. Evolution 56, 1317–1330.

Orr, H.A., 2003a. The distribution of fitness effects among beneficial mutations. Genetics 163, 1519–1526.

Orr, H.A., 2003b. A minimum on the mean number of steps taken in adaptive walks. J. Theor. Biol. 220, 241–247.

Orr, H.A., 2005. The probability of parallel evolution. Evolution 59, 216–220.

Rokyta, D.R., Joyce, P., Caudle, S.B., Wichman, H.A., 2005. An empirical test of the mutational landscape model of adaptation using a single-stranded DNA virus. Nat. Gen. 37, 441–444.

Rosenberg, N.A., 2005. A sharp minimum on the mean number of steps taken in adaptive walks. J. Theor. Biol. 237, 17–22.