DOI: 10.1534/genetics.109.106757

The Population Genetics of Adaptation: Multiple Substitutions on a Smooth Fitness Landscape

Robert L. Unckless[1] and H. Allen Orr

Department of Biology, University of Rochester, Rochester, New York 14627

Manuscript received June 29, 2009
Accepted for publication September 3, 2009

ABSTRACT

Much recent work in the theoretical study of adaptation has focused on the so-called strong selection–weak mutation (SSWM) limit, wherein adaptation is due to new mutations of definite selective advantage. This work, in turn, has focused on the first step (substitution) during adaptive evolution. Here we extend this theory to allow multiple steps during adaptation. We find analytic solutions to the probability that adaptation follows a certain path during evolution as well as the probability that adaptation arrives at a given genotype regardless of the path taken. We also consider the probability of parallel adaptation and the proportion of the total increase in fitness caused by the first substitution. Our key assumption is that there is no epistasis among beneficial mutations.

RECENTLY, there has been a great deal of interest in the genetics of adaptation. While much of this interest has focused on experimental studies of adaptive evolution (*e.g.*, Lenski and Travisano 1994; Holder and Bull 2001; Rokyta *et al.* 2005; Weinreich *et al.* 2006; Betancourt 2009), there has also been considerable interest in theoretical analyses of adaptation (reviewed in Gillespie 1991; Orr 2005; Joyce *et al.* 2008), a topic that was, until recently, surprisingly neglected.

These theoretical studies have considered a number of questions: What does the distribution of fitness effects among new beneficial mutations look like? What does the distribution of fitness effects among *fixed* beneficial mutations look like? How does clonal interference among competing mutations distort this distribution? Do early substitutions generally have larger effects on fitness than later ones? And how likely is it that independently evolving populations substitute the same beneficial mutations?

A number of approaches have been taken to these and similar questions, including analysis of phenotypic and DNA sequence-based models. Among sequence-based studies, a good deal of attention has focused on the so-called strong selection–weak mutation (SSWM) scenario, introduced by Gillespie (1983, 1984, 1991). Under this scenario, selection is strong enough that mutations are either definitely beneficial or definitely deleterious and neutral mutations are not allowed. Also, mutation is weak enough that the population is, at any point in time, essentially composed of a single wild-type DNA sequence and mutations are rare enough that double mutants and the complications of clonal interference (Gerrish and Lenski 1998) can be ignored. Adaptation in the SSWM domain thus involves the appearance and substitution of new mutations. It is usually assumed that adaptation occurs in response to a sudden change in the environment. Recurrent mutation from the current (and now somewhat maladapted) wild-type allele produces different beneficial mutations. While most of these new mutations are lost accidentally by genetic drift each time they appear, one mutation will ultimately escape stochastic loss and be substituted. At this point, the population arrives at a new wild-type sequence and the process begins anew. Adaptation thus features the stepwise substitution of single mutations.

Even in this simple SSWM scenario, mathematical analysis of adaptation has proved difficult. The most important results were derived by Gillespie (1983, 1984, 1991) himself. Given recurrent mutation from a wild-type DNA sequence to m different beneficial mutations, Gillespie calculated the probability that natural selection would, at the next substitution event, fix any particular one of these mutations. His result, as explained in more detail below, revealed that the probability that a particular mutation is fixed at the next event depends on the magnitude of its selective advantage relative to those of all available beneficial mutations. Gillespie was thus able to write down transition probabilities for the next substitution event in adaptation. As empirical studies of adaptation accumulate, it is of interest to examine the predictability of those patterns observed when populations adapt via *multiple* substitutions.

Here we extend Gillespie's analysis to allow multiple substitution events. Specifically, we show how one can calculate the probability that a population will, after the substitution of several beneficial mutations, arrive at

[1]*Corresponding author:* Department of Biology, University of Rochester, Rochester, NY 14627. E-mail: runckles@mail.rochester.edu

a particular genotype. We also calculate the probability that evolution takes a particular path to this genotype, as well as the probability that two independently evolving populations adapt in parallel, arriving at the same genotype after multiple substitutions. One of our results was derived previously by Weinreich *et al.* (2006). For completeness and clarity we briefly rederive his result here. As expected, we find that Gillespie's one-substitution results become special cases of our multiple-substitution results.

As emphasized below, our key assumption is that the beneficial effects of mutations are independent; *i.e.*, no epistasis occurs among mutations. As we also emphasize, our analysis proves more difficult mathematically than Gillespie's because the study of multiple substitutions given independent fitness effects necessarily involves a complex dependence on history: the identity of mutations fixed late during adaptation depends on the identity of mutations fixed earlier.

THE MODEL

We consider a single bout of adaptation to a sudden change in the environment but this bout may involve multiple substitution events.

Following Gillespie (1983, 1984, 1991), we assume that $Ns \gg 1$, where N is population size (our population is haploid) and s is a selection coefficient. Although selection is strong in the sense that $Ns \gg 1$, the absolute magnitude of selection coefficients might well be small. In fact, SSWM theory generally assumes—and we assume—that s is modest enough that the probability of fixation of a new unique mutation is $\sim 2s$ (Haldane 1927). We also assume that $N\mu \ll 1$, where μ is the per site rate of mutation. Because $N\mu$ is small, double mutations occur at a rate proportional to μ^2 and can be ignored. (We assume throughout most of our analysis that mutation rates are equal at all sites; we relax this assumption later.)

Perhaps most important, we also assume that beneficial mutations have independent fitness effects; *i.e.*, there is no epistasis. Thus if a beneficial mutation has fitness effect s on an original wild-type genetic background, it will also have effect s after other beneficial mutations have fixed (Kauffman 1993; Kim and Orr 2005). Independent fitness effects among mutations represent one extreme in a range of models that allow any degree of epistasis from complete to none (Kauffman 1993; Macken and Stadler 1995). The case of complete epistasis among mutations is often referred to as adaptation on a rugged or random fitness landscape, while the case of no epistasis among mutations is often referred to as adaptation on a smooth or additive fitness landscape (Kauffman 1993; Macken and Stadler 1995). (Strictly speaking, our case of independent fitness effects among mutations involves multiplicative fitness effects, which are additive on a log scale and are approximately additive when selection is weak.)

Because mutations have independent fitness effects, adaptation will feature the stepwise substitution of each of the m beneficial mutations. The order in which these substitutions occur, however, is far from obvious. In particular, at the beginning of an "adaptive walk," the wild-type allele recurrently mutates to many mutations, only m of which are beneficial (see Figure 1 for a simple example). After one of these m mutations is fixed (the first substitution), the new wild-type allele recurrently mutates to many mutations, only $m-1$ of which are beneficial. These are the same beneficial mutations as before—and have the same selection coefficients as before—except that one mutation is no longer available as it has already been substituted. This cycle of recurrent mutation followed by substitution continues until all m beneficial mutations have been substituted; the population has now arrived at an optimum and adaptation is, for the moment, complete.

Throughout our analysis, the selective advantages of the m beneficial mutations are considered given, *i.e.*, we are not concerned with draws from a distribution of selection coefficients, but with the fate of m mutations of known selective advantage.

RESULTS

Preliminary comments: Each of the mutations available to evolution is labeled mutation 1, 2, and so on. As m mutations are available, evolution might involve $K = 1, 2, \ldots, m$ substitutions. Each of the m mutations is assumed to reside at a different site in a gene or a small genome. The relevant field of genotypes available to evolution therefore includes $\sum_{K=1}^{m}\binom{m}{K} = 2^m - 1$ genotypes, where $\binom{m}{K} = m!/K!(m-K)!$. As we generally assume that mutations are beneficial, all of these genotypes represent an improvement over the original genotype. Given enough time, and the assumption of independent fitness effects, adaptation will ultimately arrive at a genotype that includes all m mutations.

Although our interest is adaptive evolution, we also, for reasons of comparison, derive various analytical results under neutral evolution.

The probability that evolution takes a particular path: Because it is simple and provides an important baseline, we first consider neutral evolution involving m mutations. We then consider evolution by positive natural selection.

Neutral case: This case can be dealt with by a simple combinatoric argument. Given m neutral mutations with K substituted, evolution can take $m!/(m-K)!$ different paths. As each of these paths has the same chance of being taken, the probability that evolution follows a particular path is just $(m-K)!/m!$.

Selection case: Now consider the case in which substitutions are driven by positive natural selection. Each of the m mutations has a known selective advantage $s_1, s_2, \ldots, s_m$.

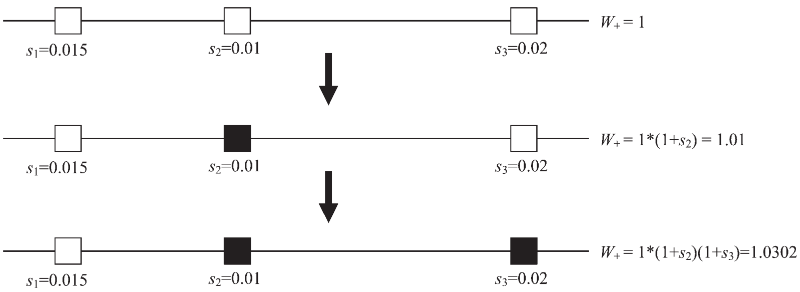

FIGURE 1.—Adaptation when beneficial mutations have independent fitness effects. In the example shown, $m = 3$ beneficial mutations are available to evolution (open boxes), each having a different selective advantage. During the time period shown, two beneficial mutations are substituted ($K = 2$; represented by solid boxes). The fitness of the wild type was initially set to one. After each substitution, fitness increases as shown.

This case is more complex than the neutral one as it features a dependence on history: the probability that an allele will be fixed at the second substitution, for example, depends on the identity of the allele fixed at the first substitution. To solve this problem, we, following WEINREICH *et al.* (2006), generalize GILLESPIE's (1984) approach. From Gillespie, we know that the probability that a beneficial mutation having advantage s_j is fixed at the first substitution is $s_j/\sum_{i=1}^{m} s_i$, the selective advantage of mutation j normalized by the sum of all selective advantages. (Gillespie's derivation of this result involved calculation of the minimum of several exponentially distributed waiting times, *i.e.*, the waiting times to fixation of each of the m beneficial mutations.) So if three beneficial mutations are available (see Figure 1), the probability that the first substitution event involves mutation 2 is $s_2/(s_1 + s_2 + s_3)$. Conditional on this event, we then consider the second substitution. The probability that mutation 3, say, is fixed at the second event is $s_3/(s_1 + s_3)$. It follows that the total probability that mutation 2 was fixed at the first event and mutation 3 at the second event is $[s_2/(s_1 + s_2 + s_3)][s_3/(s_1 + s_3)]$.

This approach is generalized easily, although some notation is required. Let T_i represent the identity of the ith mutation substituted (in the example above, $T_1 = 2$ and $T_2 = 3$). It is easy to see that, after K substitutions, the probability, $P_{\text{path}} = P(T_1 = t_1, T_2 = t_2, \ldots, T_K = t_K)$, of taking a particular path is

$$P_{\text{path}} = \frac{\prod_{i=1}^{K} s_{t_i}}{\prod_{j=1}^{K}(S - \sum_{i=1}^{j-1} s_{t_i})}, \qquad (1)$$

where $S = \sum_{i=1}^{m} s_i$ is the sum of selection coefficients among all m beneficial mutations. Equation 1, which was essentially derived by WEINREICH *et al.* (2006), collapses to Gillespie's solution when $K = 1$.

Equation 1 also lets us recover our earlier neutral result, given one modification. Equation 1 is undefined when all selection coefficients equal zero. But, for the present purposes, the salient aspect of the neutral case is that all alleles have equal probabilities of fixation. As expected, then, we recover the neutral solution from Equation 1 when all selection coefficients are equal (the s_i can be arbitrarily small but nonzero).

Comparing the neutral and the selection cases, we see that, when natural selection acts on a set of mutations with different-sized beneficial effects, the mean probability of taking a path, averaged over all paths, equals the neutral probability (as it must). But some paths now have a greater-than-neutral probability of being taken while others have a smaller-than-neutral probability.

The probability of arriving at a particular genotype: We now turn to our main problem and one that is subtler than the above: calculating the probability that natural selection arrives at a given genotype after K substitutions regardless of the order of substitutions that led to that genotype. We again begin with the neutral case and then turn to selection.

Neutral case: The neutral case can again be dispensed with quickly. Given m beneficial mutations, there are $\binom{m}{K}$ different sets of K substitutions, *i.e.*, possible resulting genotypes, where we ignore order of substitution. Under neutrality, each genotype has the same probability of being arrived at. Thus the probability of arriving at a particular genotype is $1/\binom{m}{K}$.

Selection case: Now consider the case in which substitutions are driven by positive natural selection. This case is far more difficult than the neutral one for several reasons: (i) the probability that any particular path is taken depends on history; (ii) different paths have different probabilities of being taken; and (iii) we must sum probabilities over all relevant paths.

We begin by noting that when $K = m$, *i.e.*, all available mutations have been fixed, the probability that evolution arrives at the genotype that includes all mutations is obviously one. Only the $K < m$ case is nontrivial. We can again try to solve this problem by brute force. We saw above, for example, that, with $m = 3$ and $K = 2$, the probability that selection substitutes mutation 2 followed by mutation 3 is $[s_2/(s_1 + s_2 + s_3)][s_3/(s_1 + s_3)]$. Consequently, the probability that evolution arrives at the genotype that carries mutations 2 and 3 but not 1, regardless of the order of substitutions, is $[s_2/(s_1 + s_2 + s_3)][s_3/(s_1 + s_3)] + [s_3/(s_1 + s_2 + s_3)][s_2/(s_1 + s_2)]$.

This brute-force approach becomes unwieldy as m and K grow. Fortunately, an important simplification is possible: our problem is analogous to a type of urn problem. Each ball in an urn represents a beneficial mutation and each draw of a ball (without replacement) represents a substitution. What distinguishes our problem from more familiar urn problems is that different balls can have different probabilities of being drawn,

TABLE 1

The probability that evolution arrives at particular genotypes

	Selection coefficients	Mean P_G	$P_G(1, 2, 3, 4)$ (most likely genotype)	$P_G(6, 7, 8, 9)$ (least likely genotype)	$\frac{P_G(1,2,3,4)}{P_G(6,7,8,9)}$	$E[s_{\text{fixed}} \mid K=1]$	$E[\text{prop}_{K=1}]$
Neutral or equal s	All 0 or all equal	0.0079	0.0079	0.0079	1	0.0200[a]	0.1111[a]
Selection (small variation in s)	0.024, 0.023, 0.022, 0.021, 0.020, 0.019, 0.018, 0.017, 0.016	0.0079	0.0142	0.0041	3.443	0.0203	0.1130
Selection (large variation in s)	0.039, 0.035, 0.030, 0.025, 0.020, 0.015, 0.010, 0.005, 0.001	0.0079	0.0972	2.226×10^{-5}	4296	0.0279	0.1550

All numerical values derive from analytical solutions with $m = 9$ mutations and $K = 4$ substitutions. In all cases, the mean selection coefficient among the available beneficial mutations equals 0.02. Mean P_G is the mean probability of arriving at a genotype, averaging over all possible genotypes. Calculations assume $2s$ approximations to the probability of fixation. Essentially exact solutions using $1\text{-}e^{-2s}$ for the probability of fixation differ slightly from those shown: *e.g.*, in the large variation case (which represents a worst case for SSWM approximations), the probabilities of the most and the least likely genotype, expected fitness gain due to the first substitution, and proportion of total fitness gain due to the first substitution are 0.0945, 2.463×10^{-5}, 0.0278, and 0.154, respectively.

[a] Results assume all $s_i = 0.02$.

just as different mutations can have different selective advantages and thus probabilities of fixation. Our question is: What is the probability of drawing a particular set of K balls from an urn containing m balls when each ball has a different probability of being drawn? (We sum over all relevant orders of draws.) The solution to this problem is provided by a special case of the multivariate Wallenius noncentral hypergeometric (MWNH) distribution (Chesson 1976; Fog 2008). The appendix provides a brief introduction to this bit of probability theory.

To use the MWNH distribution, we introduce the indicator variables, X_1, X_2, etc. $X_1 = 1$ means that mutation 1 was substituted, while $X_1 = 0$ means it was not, and so on. The number of substitutions is just

$$K = \sum_{i=1}^{m} x_i. \tag{2}$$

Given m, K, and the selective advantages of mutations, we want to find quantities like $P(X_1 = 0, X_2 = 1, X_3 = 1)$: *i.e.*, the probability that evolution arrives at a genotype carrying mutations 2 and 3 but not 1. More generally, we want to find $P_G = P(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m)$.

The MWNH approach shows that

$$P_G = \int_0^1 \prod_{i=1}^{m} (1 - t^{s_i/D})^{x_i}\, dt, \tag{3}$$

where $D = \sum_{i=1}^{m} s_i(1 - x_i)$ is the sum of selection coefficients among mutations that are *not* substituted. Equation 3 is one of our main results.

As expected, Equation 3 collapses to Gillespie's probabilities when $K = 1$. For instance, with $m = 3$, Equation 3 shows that $P(X_1 = 0, X_2 = 1, X_3 = 0) = s_2/(s_1 + s_2 + s_3)$, as it must. Similarly, when $m = 3$ and $K = 2$, Equation 3 shows that $P(X_1 = 0, X_2 = 1, X_3 = 1) = [s_2/(s_1 + s_2 + s_3)][s_3/(s_1 + s_3)] + [s_3/(s_1 + s_2 + s_3)][s_2/(s_1 + s_2)]$, as we found by brute force. The important point is that Equation 3 allows us to find the probability of arriving at any arbitrary genotype by natural selection, no matter how many beneficial mutations are available or how many substitutions occur. To take an example, consider the situation described in Table 1: $m = 9$, $K = 4$, and the selection coefficients among beneficial mutations are as shown in the table. Table 1 provides the probability of arriving at the "best" genotype after four substitutions (the genotype that includes the four mutations with the largest selection coefficients) as well as the probability of arriving at the "worst" (but still adaptive) genotype after four substitutions (the genotype that includes the four mutations with the smallest selection coefficients). These probabilities, which were all calculated from Equation 3, can be quite different: natural selection is more likely to push an adapting population to certain genotypes than others.

Again, we can also recover our earlier neutral results, given the same modification as before: because Equation 3 requires that $D > 0$, we cannot use the MWNH machinery when all selection coefficients equal zero. However, if all alleles have equal selection coefficients and thus fixation probabilities, we recover the neutral solution from Equation 3.

By analogy with our earlier results, it is also easy to see that, when natural selection acts on a set of mutations with different-sized beneficial effects, the mean probability of arriving at a genotype after K substitutions (averaged over all relevant genotypes) equals the neutral probability (as it must). The difference is that, given K substitutions, some genotypes now have a greater-than-neutral probability of being arrived at and others have a smaller-than-neutral probability of being arrived at (again, see Table 1 for examples). The MWNH approach allows us to quantity the effect of this distortion of neutral probabilities due to positive natural selection (see Table 1).

Expected fitness and proportion of fitness increase due to first substitution: We can also write down the expected fitness at each substitution. Matters are especially simple, and interesting, at the first substitution. In particular, the expected s for the mutation fixed at $K = 1$ is

$$\begin{aligned} E[s_{\text{fixed}} \mid K = 1] &= s_1\left(\frac{s_1}{S}\right) + s_2\left(\frac{s_2}{S}\right) + \ \ldots \ + s_m\left(\frac{s_m}{S}\right) \\ &= \frac{\sum_{j=1}^{m} s_j^2}{S} = \bar{s} + \frac{\text{Var}[s]}{\bar{s}}, \end{aligned} \tag{4}$$

where s_{fixed} is a selection coefficient among fixed (not merely new) mutations, S is again the sum of selection coefficients among new beneficial mutations, $\bar{s} = S/m$ is the mean selection coefficient among new beneficial mutations, and $\text{Var}[s] = \sum_{i=1}^{m}(s_i - \bar{s})^2/m$ is their variance. Equation 4, which is implicit in Orr (2002), captures the obvious, but important, fact that natural selection outperforms random choice among beneficial mutations. If, at the first substitution, evolution were to randomly choose a mutation for fixation, we would have $E[s_{\text{fixed}}|K = 1] = \bar{s}$. Instead, evolution by natural selection does better than this. How much better depends on the variance among selection coefficients. Table 1 confirms that, when variation in selection coefficients is appreciable, the expected fitness gain due to the first substitution can be substantially larger than the mean selection coefficient among new mutations. (Because it would violate our SSWM assumptions—and in particular our $2s$ approximation to the probability of fixation—Var[s] in Equation 4 cannot grow too large. Equation 4 is, therefore, a SSWM approximation.) Note that because Equation 4 concerns only the first substitution, it assumes nothing about epistasis among beneficial mutations.

We can also calculate the expected proportion, $E[\text{prop}_{K=1}]$, of the total increase in fitness occurring over a complete bout of adaptation that is due to the first substitution. Once all m beneficial mutations are substituted, fitness increases by approximately S (where the approximation neglects higher-order terms involving products of the s_i). From Equation 4, then, we have

$$\begin{aligned} &E[\text{prop}_{K=1}] \\ &\quad \approx E[s_{\text{fixed}}/S \mid K = 1] = \frac{\sum_{j=1}^{m} s_j^2}{S^2} = \frac{1}{m} + \frac{\text{Var}[s]}{m\bar{s}^2}. \end{aligned} \tag{5}$$

Equation 5 shows that a disproportionately large share of the total increase in fitness is caused by the first substitution: $E[\text{prop}_{K=1}] > 1/m$, assuming only weak selection, independent fitness effects, that multiple mutations are available, and that not all have the same selective advantage (for numerical examples, see Table 1). Put differently, adaptation is characterized by a curve of diminishing returns through time, with earlier substitutions having larger effects on fitness than later ones. While this pattern has been noted in many previous studies of adaptation (Orr 1998, 2002; Joyce *et al.* 2008), Equation 5 provides a simple demonstration of the point, at least when mutations have independent effects. Two special cases are also worth noting. First, if only $m = 1$ beneficial mutation is available, then $\text{Var}[s] = 0$ and, from Equation 5, $E[\text{prop}_{K=1}] = 1$, as it must. Second, if multiple beneficial mutations are available but all have the same selective advantage, then $\text{Var}[s] = 0$ and $E[\text{prop}_{K=1}] = 1/m$; *i.e.*, the first substitution contributes $1/m$ to the total increase in fitness, and each subsequent substitution makes the same contribution.

It is far more difficult to derive the expected fitness at later substitutions or the proportion of the total fitness increase due to later substitutions, although the MWNH approach allows some progress. After K substitutions, the probability that adaptation has arrived at a particular genotype, P_G, is given by Equation 3. The increase in fitness over the fitness of the starting allele is $\sim \sum_{i=1}^{m} s_i x_i$, where it is understood that only K of the x_i terms are nonzero (as only K substitutions occurred) and we again neglect higher-order terms involving products of the s_i. Summing over the $\binom{m}{K}$ possible combinations of mutations that involve K substitutions, we get

$$\begin{aligned} &E\left[\sum s_{\text{fixed}} \mid K \text{ substitutions}\right] \\ &\quad \approx \sum_{j=1}^{\binom{m}{K}\ \textit{combinations}} \left[\left(\sum_{i=1}^{m} s_i x_{i,j}\right) P_{G,j}\right], \end{aligned} \tag{6}$$

where j is an index that represents each of the combinations of K substitutions given m mutations. Equation 6 is obviously only notation that represents the desired solution and it must be evaluated numerically after listing the $\binom{m}{K}$ possible combinations of substitutions. In a few simple cases, *e.g.*, $m = 3$, $K = 2$, closed-form solutions to Equation 6 are possible but even these are complicated (not shown).

Probability of parallel evolution: So far we have considered the evolution of a single population. But we can use the approaches taken above to consider the evolution of two or more populations. In particular, we can consider the probability that two or more populations evolve in parallel, substituting the same mutations.

For simplicity, we consider a scenario in which two strictly allopatric populations begin with the same wild-type sequence and experience the same environmental change. Both populations are thus presented with the same set of m mutations and—because the populations share the same new environment—these mutations have the same selective advantages, $s_1, s_2, \ldots, s_m$.

We now consider the probability that, after K substitutions, both populations arrive at the same genotype, whether or not they take the same path to it. In the neutral case, we saw that the probability that a population arrives at a particular genotype is $1/\binom{m}{K}$. The probability that two independently evolving populations arrive at this genotype is thus just $1/\binom{m}{K}^2$. Because there are $\binom{m}{K}$ different genotypes that include K substitutions, the total probability that two populations arrive at the same genotype after K substitutions is $K!(m-K)!/m!$.

In the positive selection case, similar logic shows that

$$P(\text{same genotype}) = \sum_{j=1}^{\binom{m}{K}\ \text{combinations}} P_{\mathrm{G},j}^2, \qquad (7)$$

where P_{G} is given by Equation 3. Equation 7 is again obviously only notation that represents the desired solution; it must be evaluated numerically after listing all $\binom{m}{K}$ combinations of genotypes that include K substitutions. It is worth noting that Equation 7 collapses to the neutral probability of parallel evolution when all s_i are equal, as it should. Similarly, when $K = m$, *i.e.*, all beneficial mutations have been substituted in both populations, $P(\text{same genotype}) = 1$, as expected.

Analogous calculations of the probability that two independent populations not only arrive at the same genotype but also take the same path to it are also possible but are complicated; we suppress them here. Again, though, they can be calculated numerically from Equation 1, squaring the resulting probabilities and summing over all paths to the same genotype.

Extensions: Finally, we note that the above analysis can be generalized in several ways. First, we have assumed, following most SSWM theory, that the rate of mutation to each mutation is equal. This need not be true. The above results can be extended to allow for mutational bias by replacing selection coefficients, s_i, in the right-hand side of our equations by $\mu_i s_i$, where μ_i is the rate of mutation to the ith allele (see also Rokyta *et al.* 2005). The biological point is that the probability of substituting a particular mutation at the next step in evolution depends on the product of its advantage and the rate at which it appears by mutation. Modifying Gillespie's (1983, 1984) result, the probability that the jth mutation is substituted at the next step in evolution now takes the form $\mu_j s_j / \sum_{i=1}^{m} \mu_i s_i$.

Second, we have assumed, following most SSWM theory, that the selective advantage of beneficial mutations is modest enough that the probability of fixation is $\Pi \approx 2s$. But we can allow arbitrarily large selective advantages by using the more exact probability $\Pi \approx 1 - \exp(-2s)$ throughout. Analytic results obviously become cumbersome but numerical calculation is straightforward.

Third, we have assumed that adaptation involves new mutations. Our results can be extended, albeit approximately, to alleles from the standing genetic variation, so long as alleles start at very low frequencies. Then, all copies of a mutation enjoy nearly independent probabilities of fixation and initial frequency is easily incorporated into our analysis. In particular, if alleles have low mutation–selection balance frequencies, we can replace s_i in the right-hand side of our solutions by $\mu_i s_{\mathrm{b},i}/s_{\mathrm{d},i}$, where $s_{\mathrm{d},i}$ is a mutation's disadvantage in the old environment and $s_{\mathrm{b},i}$ is its advantage in the new one (similarly, see Orr and Betancourt 2001). The probability that the jth mutation is substituted at the next step in evolution now takes the form $(\mu_j s_{\mathrm{b},j}/s_{\mathrm{d},j}) / \sum_{i=1}^{m} \mu_i s_i / s_{\mathrm{d},i}$. While this approach assumes that mutations start at deterministic mutation–selection equilibrium frequency (*i.e.*, it ignores the stationary distribution of starting frequency), computer simulations confirm that it provides a reasonably accurate approximation when the absolute numbers of mutations segregating are very small (not shown).

DISCUSSION

We have extended Gillespie's (1983, 1984, 1991) analysis of adaptation in the SSWM domain to allow multiple substitutions. In particular, while Gillespie showed how one can calculate the probability that, at the next substitution, evolution arrives at a genotype that includes any one of m available beneficial mutations, here we calculate the probability that, after K substitutions, evolution arrives at a genotype that includes a particular set of the m mutations. This calculation builds on earlier work, by both Gillespie and Weinreich *et al.* (2006), that allowed calculation of the probability that evolution takes a particular path to this genotype. We also calculate the expected fitness effect of the first substitution, the (approximate) proportion of the total increase in fitness caused by the first substitution, and the probability of parallel adaptation between two independently evolving populations. For several of these statistics, we compare our results with those expected under neutral evolution. Our key biological conclusion is that positive natural selection distorts the probability that evolution arrives at a given genotype or takes a given path to this genotype away from the analogous neutral probabilities. While this effect is obvious intuitively, our analysis lets us quantify it (for numerical examples, see Table 1).

The calculations presented here are generally more difficult than those in Gillespie's classic SSWM work. The reason is simple. When considering evolution one step into the future, evolution features no history dependence. The identity of the mutation fixed at the next substitution in evolution depends only on the present state of the population. But when considering evolution several steps into the future, evolution *does* feature history dependence. The identity of the mutation that is most likely to get fixed at substitution two, for instance, depends on the identity of the mutation fixed at substitution one. Statistically, then, our problem involves sampling without replacement: given a set of m beneficial mutations, one mutation is fixed at each substitution event, progressively shrinking the pool of mutations available to evolution. Also, as Gillespie emphasized, the sampling of mutations at each substitution involves a kind of competition: each of the beneficial mutations available at any point in time can be thought of as competing with all others for fixation at the next substitution event. The probability that a particular beneficial mutation "wins" this competition depends on its selective advantage relative to those of all available beneficial mutations. Combining these properties of sampling without replacement and sampling with competition, we find that the solutions to several of our problems—for instance, calculation of the probability that adaptation arrives at a given genotype after K substitutions—involve fairly obscure probability theory, *e.g.*, the multivariate Wallenius noncentral hypergeometric distribution.

Given that our solution to the probability that adaptation arrives at a particular genotype given multiple substitutions includes an integral, it may be asked if it represents an improvement over, say, a solution derived from Monte Carlo simulation of adaptation (given known selection coefficients). We believe that the answer is yes, for at least two reasons. First, and technically, Equation 3 allows accurate numerical calculation of probabilities even when some relevant paths involve extremely low probabilities; these extreme probabilities might not be accurately found by Monte Carlo simulation of adaptation. (Given that Equation 3 allows accurate determination of the probabilities of *a priori* rare outcomes—but outcomes that might nonetheless occur in actual experiments—it may serve as the basis for statistical tests that ask, *e.g.*, whether beneficial fitness effects are actually independent.) Second, and more generally, although Equation 3 does not allow ready biological insight, it does represent the solution to the simplest scenario for adaptation: adaptation from new mutations in the SSWM limit and with no epistasis. It may well be that more complex scenarios, *e.g.*, those involving epistasis, will yield analytic solutions that are variations on Equation 3. If so, the pattern of these variations might well provide insight into the biology of these different scenarios.

Our conclusions rest on several assumptions. Most obviously, we make standard SSWM assumptions: selection is strong in relative terms ($Ns \gg 1$) but modest in absolute terms (s is small enough that $\Pi \approx 2s$ approximations are appropriate); also, mutation is rare enough that the population is, at any point in time, composed of a single wild-type sequence. Perhaps more important, most of our results rest on the assumption that mutations have independent fitness effects; *i.e.*, the beneficial effect of a mutation does not depend on genetic background. This is a significant assumption and our results would almost certainly change under epistasis. (Indeed this explains why we have not attempted to compare the present results with empirical ones; in several of the best experiments reported to date, *e.g.*, Weinreich *et al.* 2006 and Betancourt 2009, epistasis for fitness among beneficial mutations is evident.)

It is important to emphasize therefore that we analyze the independent fitness case not because we believe it represents biological reality; it almost certainly does not. Instead we analyze this case because it represents the natural starting point—and a plausible null model—for more complex and realistic analyses.

For several of our results, the effects of epistasis could, in principle, be incorporated. For instance, after the fixation of the first beneficial mutation the selection coefficients for the remaining $m - 1$ mutations could be reassigned (allowing epistasis), and similarly after the second substitution, and so on. One could then calculate the probability that a given path is taken during evolution using these conditional selection coefficients. This is essentially the approach taken by Weinreich *et al.* (2006). In principle, one could sum over all relevant paths, finding the total probability of arriving at a given genotype. Unfortunately, however, these calculations are essentially brute force and numerical. We see no way of deriving general analytic solutions to the probability of arriving at a given genotype, *e.g.*, an analog of our Wallenius distribution result, when arbitrary forms of epistasis are allowed.

In any case, we do not pursue this complex problem here. Our hope is that the independent fitness case, while simple—and likely simplistic—can at least serve as a baseline against which more complex scenarios can be compared.

We thank two anonymous reviewers and N. Takahata for very helpful comments and suggestions. We also thank members of the ecology and evolution group at the University of Rochester for critical comments. This research was supported by funds from the National Institutes of Health (GM51932) to H.A.O. and from the Robert and Mary Sproull Fellowship of the University of Rochester to R.L.U.

LITERATURE CITED

Betancourt, A., 2009 Genomewide patterns of substitution in adaptively evolving populations of the RNA bacteriophage MS2. Genetics **181:** 1535–1544.

Chesson, J., 1976 A non-central multivariate hypergeometric distribution arising from biased sampling with application to selective predation. J. Appl. Probab. **13:** 795–797.

Fog, A., 2008 Calculation methods for Wallenius' noncentral hypergeometric distribution. Commun. Stat. Simul. Comput. **37:** 258–273.

Gerrish, P. J., and R. E. Lenski, 1998 The fate of competing beneficial mutations in an asexual population. Genetica **102/103:** 127–144.

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

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

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

Haldane, J. B. S., 1927 A mathematical theory of natural and artificial selection, part V: selection and mutation. Proc. Camb. Philos. Soc. **23:** 838–844.

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

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.

Kauffman, S. A., 1993 *The Origins of Order.* Oxford University Press, New York.

Kim, Y., and H. A. Orr, 2005 Adaptation in sexuals *vs.* asexuals: clonal interference and the Fisher–Muller model. Genetics **171:** 1377–1386.

Lenski, R. E., and M. Travisano, 1994 Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. Proc. Natl. Acad. Sci. USA **91:** 6808–6814.

Macken, C. A., and P. F. Stadler, 1995 Evolution on fitness landscapes, pp. 43–86 in *1993 Lectures in Complex Systems*, edited by L. Nadel and D. L. Stein. Addison–Wesley, Reading, MA.

Orr, H. A., 1998 The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. Evolution **52:** 935–949.

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

Orr, H. A., 2005 Genetic theories of adaptation: a brief history. Nat. Rev. Genet. **6:** 119–127.

Orr, H. A., and A. Betancourt, 2001 Haldane's sieve and adaptation from the standing genetic variation. Genetics **157:** 875–884.

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

Ross, S., 1994 *A First Course in Probability Theory.* MacMillan College Publishing, New York.

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

Communicating editor: N. Takahata

APPENDIX: THE WALLENIUS DISTRIBUTION

The hypergeometric distribution describes the number of balls of a given color drawn from an urn when sampling without replacement (Ross 1994). When more than two colors are present in the urn, the appropriate distribution is the multivariate hypergeometric. Our problem is more complex than this, however, as we sample not only without replacement but with bias: some balls (*i.e.*, mutations with larger selective advantages) are more likely to be drawn (substituted) than others (those with smaller selective advantages). Sampling, in other words, features competition among objects. When sampling is without replacement *and* with bias, the distribution of balls drawn from an urn is given by the Wallenius noncentral hypergeometric distribution (Chesson 1976; Fog 2008; see also http://www.agner.org/random/theory/nchyp1.pdf and http://en.wikipedia.org/wiki/Wallenius%27_noncentral_hypergeometric_distribution).

The multivariate form of this distribution is

$$P(\text{draw } x_i \text{ balls of the } i\text{th color}) = \prod_{i=1}^{c}\binom{b_i}{x_i}\int_0^1\prod_{i=1}^{c}(1-t^{s_i/D})^{x_i}\,dt,$$

where b_i is the number of balls in the urn of the ith color, c is the total number of colors present, s_i measures the strength of the bias for drawing a ball of the ith color (these s_i can be scaled arbitrarily), and $D=\sum_{i=1}^{c}s_i(b_i-x_i)$.

In our biological problem, there is one mutant of each type ($b_i=1$ for all i) and the total number of mutant types is m. The above distribution thus becomes

$$P(\text{draw } x_i \text{ balls of the } i\text{th color}) = \int_0^1\prod_{i=1}^{m}(1-t^{s_i/D})^{x_i}\,dt,$$

as in the text.