

# The Limits of Complex Adaptation: An Analysis Based on a Simple Model of Structured Bacterial Populations

Douglas D. Axe\*

Biologic Institute, Redmond, Washington, USA

## Abstract

To explain life's current level of complexity, we must first explain genetic innovation. Recognition of this fact has generated interest in the evolutionary feasibility of complex adaptations—adaptations requiring multiple mutations, with all intermediates being non-adaptive. Intuitively, one expects the waiting time for arrival and fixation of these adaptations to have exponential dependence on  $d$ , the number of specific base changes they require. Counter to this expectation, Lynch and Abegg have recently concluded that in the case of selectively neutral intermediates, the waiting time becomes independent of  $d$  as  $d$  becomes large. Here, I confirm the intuitive expectation by showing where the analysis of Lynch and Abegg erred and by developing new treatments of the two cases of complex adaptation—the case where intermediates are selectively maladaptive and the case where they are selectively neutral. In particular, I use an explicit model of a structured bacterial population, similar to the island model of Maruyama and Kimura, to examine the limits on complex adaptations during the evolution of paralogous genes—genes related by duplication of an ancestral gene. Although substantial functional innovation is thought to be possible within paralogous families, the tight limits on the value of  $d$  found here ( $d \leq 2$  for the maladaptive case, and  $d \leq 6$  for the neutral case) mean that the mutational jumps in this process cannot have been very large. Whether the functional divergence commonly attributed to paralogs is feasible within such tight limits is far from certain, judging by various experimental attempts to interconvert the functions of supposed paralogs. This study provides a mathematical framework for interpreting experiments of that kind, more of which will be needed before the limits to functional divergence become clear.

**Cite as:** Axe DD (2010) The limits of complex adaptation: An analysis based on a simple model of structured bacterial populations. *BIO-Complexity* 2010(4):1-10. doi:10.5048/BIO-C.2010.4

**Editor:** James Keener

**Received:** May 13, 2010; **Accepted:** September 18, 2010; **Published:** December 21, 2010

**Copyright:** © 2010 Axe. This open-access article is published under the terms of the Creative Commons Attribution License, which permits free distribution and reuse in derivative works provided the original author(s) and source are credited.

**Note:** A *Critique* of this paper, when available, will be assigned doi:10.5048/BIO-C.2010.4.c.

\* Email: daxe@biologicinstitute.org

## INTRODUCTION

Although much of the traditional work in population genetics has focused on how existing alleles are propagated under various conditions, there has in recent years been a growing interest in the origin of the alleles themselves. Of particular interest are alleles that are not merely adaptive, in that they enhance fitness, but *innovative*, in that they endow their hosts with unprecedented capabilities. Since we now have a large and rapidly growing catalog of functional protein systems that seem to be fundamentally complex [1], there is a growing sense that innovations of this kind would require *complex adaptations*, meaning adaptations needing not just one specific new mutation but several, with all intermediates being non-adaptive. If so, this may present a probabilistic challenge to the standard evolutionary model because it would require fortuitous convergence of multiple rare events in order for a selective benefit to be realized.

Three potential routes to the fixation of complex adaptations have been recognized. The simplest is the *de novo* appearance in one organism of all necessary changes, which for large innovations is tantamount to *molecular saltation*. This route has the advantage of avoiding non-adaptive intermediates but the disadvantage of requiring a very rare convergence of mutations. The second potential route is *sequential fixation*, whereby point mutations become fixed successively, ultimately producing the full set needed for the innovation. By this route, the rate of appearance of

each successive intermediate en route to the complex adaptation is boosted by allowing the prior intermediate to become fixed. But because these fixation events have to occur without the assistance of natural selection (or, in the case of maladaptive intermediates, even *against* natural selection) they are in themselves improbable events. The third potential route is *stochastic tunneling*, which differs from sequential fixation only in that it depends on each successive point mutation appearing without the prior one having become fixed. Here fixation occurs only after all the mutations needed for the innovation are in place. This route therefore benefits from an absence of improbable fixation events, but it must instead rely on the necessary mutations appearing within much smaller subpopulations.

Molecular saltation seems incompatible with Darwinian evolution for the same reason all forms of saltation do—namely, the apparent inability of ordinary processes to accomplish extraordinary changes in one step. If specific nucleotide substitutions occur spontaneously at an average rate of  $u$  per nucleotide site per cell, and a particular innovation requires  $d$  specific substitutions, then the rate of appearance of the innovation by molecular saltation (i.e., construction *de novo* in a single cell) is simply  $u^d$  per cell. This means that the expected waiting time (in generations) for appearance and fixation of the innovation scales as  $u^{-d}$ . But since  $u$  has to be a very small fraction in order for a genome to be faithfully replicated (the upper bound being roughly the inverse of the working genome length in bases),  $u^{-d}$  becomes exceedingly large

even for modest values of  $d$ , resulting in exceedingly long waiting times.

Because of this, sequential fixation and stochastic tunneling are thought to be the primary ways that complex adaptations become fixed. However, in view of the fact that the underlying limitation is an unavoidable aspect of statistics—that independent rare events only very rarely occur in combination—it seems certain that all chance-based mechanisms must encounter it. Whether or not this poses a serious problem for Darwinism is, of course, another matter. If simple adaptations suffice to explain nearly everything that needs to be explained, with feasible complex adaptations explaining the rest, then there is no problem. Still, it is important to characterize the limits of feasibility for complex adaptation in order to draw conclusions about the limits of Darwinism.

Among the many treatments of complex adaptation by sequential fixation and/or stochastic tunneling (e.g., references 2–6), one recently offered by Lynch and Abegg [6] is of particular interest because it claims that the above limitation vanishes in situations where the genetic intermediates en route to a complex adaptation are selectively neutral. For this case, they report that “regardless of the complexity of the adaptation, the time to establishment is inversely proportional to the rate at which mutations arise at single sites.” In other words, they find the waiting time for appearance and fixation of a complex adaptation requiring  $d$  base changes to scale as  $u^{-1}$  rather than the commonly assumed  $u^{-d}$ . Because this should apply not only in the case of strict neutrality (which may seldom exist) but also in the more realistic case of approximate neutrality, and because it represents a striking departure from common probabilistic intuitions, it is important for this result to be examined carefully.

Indeed, a simple thought experiment seems to justify skepticism with respect to it. The number of ways to specify  $d$  base changes in a gene that is  $L$  bases in length is:

$$n_{d \text{ in } L} = 3^d \binom{L}{d} = \frac{3^d L!}{d!(L-d)!}, \quad (1)$$

which can easily exceed the number organisms that have existed on earth. There are, for example, some  $10^{72}$  ways to make thirty base changes to a gene of kilobase length. For any given gene of this size, then, we can be certain that the vast majority of variants that differ at thirty base positions have never existed. And since existence is a prerequisite for fixation, we can be equally certain that the vast majority of these variants cannot have become fixed in any real population. Moreover, we do not have to make any assumptions about selective neutrality in order to draw this conclusion.

In what follows, I offer an analysis of the mathematical reasoning behind Lynch and Abegg’s treatment of the neutral-intermediates case, highlighting what appear to be significant errors, and I develop what I believe to be accurate treatments of arbitrarily complex adaptations in bacterial populations, both in the case of mildly deleterious intermediates and in the case of neutral intermediates.

## ANALYSIS

### Assessing Lynch and Abegg’s treatment of the neutral case

*Sequential fixation.* Innovation by sequential fixation has to work against natural selection if any of the genetic intermediates are maladaptive, which is apt to be the case in many evolutionary scenarios. In such cases, sequential fixation is conceivable only

in small populations because the efficiency of natural selection in large populations makes maladaptive fixation nearly impossible [7]. Focusing therefore on small populations, and initially on the limiting case of selectively neutral intermediates, Lynch and Abegg calculate the mean waiting time for arrival of an allele carrying a particular complex adaptation that is destined to be fixed within a diploid population as<sup>1</sup>:

$$w_{seq} \approx [d \cdot u]^{-1} + [(d-1)u]^{-1} + [(d-2)u]^{-1} + \dots + [2u]^{-1} + [2uN_e \phi_2]^{-1}, \quad (2)$$

where  $d$  is the number of specific base substitutions needed to produce the complex adaptation,  $N_e$  is the effective population size<sup>2</sup>,  $u$  is the mean rate of specific base substitution (per site per gamete), and  $\phi_2$  is the probability that an instance of producing the adaptive allele will result in fixation (given by Equation 1 of reference 6).

When written in the above expanded form, we see that the waiting time is being equated with a sum of  $d$  terms, each of these terms being the inverse of a rate. As explained by Lynch and Abegg [6], the individual rates are simply the rates of appearance within the whole population (per generation) of instances of the successive genotypic stages that are destined to become fixed, given a population in which the prior stage is fixed. So, for example, in a case where the complex adaptation requires five specific base substitutions ( $d = 5$ ) and the predominant genotype in the initial population lacks all five, there are five possible ways for an allele of the initial type (call it *stage 0*) to progress by mutation to the next stage (*stage 1*). The first term enclosed in square brackets is  $5u$  in this case, which is the per-gamete rate of appearance of stage-1 alleles in a stage-0 population. Because each individual allele in a population of neutral variants is expected to become fixed with a probability equal to the inverse of the total allele count [8], the rate of appearance of stage-1 alleles that are destined to become fixed in a diploid population of  $N$  individuals is equal to the total rate of appearance ( $2N \times 5u$  per generation) divided by  $2N$ , which equals the per-gamete rate of mutation to stage 1, namely  $5u$  per generation.

When a stage-1 allele is fixed, there are now four possible ways for mutation to produce a stage-2 allele. The second term in square brackets,  $4u$ , is again the rate of appearance of stage-2 alleles that are destined to become fixed within a population where a stage-1 allele has become fixed, and so on for the third and fourth terms. The final term differs because the complete complex adaptation (stage 5) has an enhanced fixation probability resulting from its selective advantage.

Since the mean wait for a stochastic event to occur is just the inverse of its mean rate of occurrence, the right-hand side of Equation 2 is now seen to be the sum of waiting times, specifically the mean waiting times for appearance of destined-to-be-fixed alleles at each successive stage within populations in which the prior stage has become fixed. The actual process of fixation (after a destined-to-be-fixed allele appears) is relatively fast in small populations, taking an average of  $4N_e$  generations for neutral alleles [8]. Assuming this to be negligible, Equation 2 may seem at first glance to represent precisely the intended quantity—the overall waiting time for fixation of the complete complex adaptation from a stage-0 starting point. But if the implications of this equation are as implausible as has been argued, then there must be a mistake in the calculation.

<sup>1</sup> See Equation 5a of reference 6, which uses  $\bar{t}_{e,s}$  to represent the same quantity.

<sup>2</sup> Much of the genetic drift in real populations results from non-uniform population structure and dynamics. In essence, the effective size of a real population is the size of an ideal population lacking these non-uniformities that has the same level of genetic drift. A detailed discussion of  $N_e$  is found in the results and Discussion section.

This indeed appears to be the case. Specifically, of all the possible evolutionary paths a population can take, the analysis of Lynch and Abegg considers only those special paths that lead directly to the desired end—the complex adaptation. This is best illustrated with an example. Suppose a population carries an allele that confers no selective benefit in its current state (e.g., a pseudogene or a gene duplicate) but which would confer a benefit if it were to acquire five specific nucleotide changes relative to that initial state, which we will again refer to as stage 0. Lynch and Abegg assign a waiting time of  $(5u)^{-1}$  for a stage-1 allele to become fixed in this situation, which is valid *only* if we can safely assume that the population remains at stage 0 during this wait. But this cannot be assumed. A stage-0 allele of kilobase length, for example, would have about 200-fold more correct bases than incorrect ones (with respect to the complex adaptation), which means the rate of degradation (i.e., fixation of changes that make the complex adaptation more remote) would be about 600-fold higher<sup>3</sup> than the rate of progression to stage 1. It is therefore very unlikely in such a case that the population will wait at stage 0 long enough to reach stage 1, and the situation becomes progressively worse as we consider higher stages.

It is possible to adjust the problem to some extent in order to achieve a more favorable result. For example, if we suppose that all changes except the five desired ones are highly maladaptive, then fixation becomes restricted to changes at the five sites. But even under these artificial restrictions, Equation 2 is incorrect in that it ignores back mutations [6]. Since the aim is to acquire the correct bases at all  $d$  sites, and there are more incorrect possibilities than correct ones at each site, counterproductive changes must substantially outnumber productive changes as the number of correct bases increases. So even in this highly favorable case, the analysis suffers from neglect of counterproductive competing paths. Productive changes cannot be ‘banked’, whereas Equation 2 presupposes that they can.

*Stochastic tunneling.* Lynch and Abegg’s treatment of stochastic tunneling with neutral intermediates is also problematic. To derive an expression for the waiting time when the population is large enough to preclude fixation of intermediate stages (Equation 5b of reference 6), they approximate the frequency of stage- $d$  alleles at  $t$  generations as  $(ut)^d$ . While they note that this approximation is valid only if  $ut \ll 1$ , they overlook the fact that this restricts their analysis to exceedingly small values  $(ut)^d$ . Specifically, they equate this term with the substantial allele frequency at which fixation becomes likely<sup>4</sup>, and then proceed to solve for  $t$ , taking the result to be valid for arbitrarily large values of  $d$ . This leads them to the unexpected conclusion that “in very large populations with neutral intermediates, as  $d \rightarrow \infty$ , the time to establishment converges on the reciprocal of the per-site mutation rate, becoming independent of the number of mutations required for the adaptation” [6]. But since they have in this way neglected the effect of  $d$ , it should be no surprise that they find  $d$  to have little effect.

### Modeling complex adaptation in large structured populations

*Approach.* Turning now to the problem of deriving a satisfactory expression for the overall time for appearance and fixation of an arbitrarily complex adaptation, we begin by considering appropriate ways to define the problem. In the first place, since our primary interest is to place reliable limits on what is evolu-

tionarily feasible, we will focus on the kinds of populations that are most apt to produce complex adaptations, by which we mean those providing the most opportunities for chance events to accomplish something of significance. In this regard, bigger populations are definitely better. Lynch and Abegg describe an advantage of very small populations being that maladaptive alleles can become fixed by genetic drift [6]. But the ease of fixation is one thing, and the consequence is another. In view of the fact that maladaptive mutations are much more common than adaptive ones [9, 10], populations that are too small for purifying selection to work will inevitably undergo genetic degradation, which makes extinction the most likely outcome. We will therefore consider a global bacterial population where the effective population size may be on the order of  $10^9$ , which is much higher than estimated sizes for more complex forms of life [11].

Analysis of bacterial genomes shows that many genes are significantly similar to other genes in the same genome. These similar genes are presumed to be *paralogs*, meaning that they originated from a common ancestral gene by means of gene duplication events. Although the functional redundancy of duplicate genes implies that most of them will be lost, their lack of functional constraint also frees them to acquire mutations that may produce new functions [12,13]. We will therefore frame our description of complex adaptation around the assumption that it is initiated by a duplication event that provides a suitable starting gene. The alternative assumption, that a population carries a preexisting unconstrained allele, will be treated as a special case.

Also, considering the interest in the role of small populations in the origin of complex adaptations, it will be helpful to base the present analysis on an explicitly structured whole population instead of relying on the concept of effective population size for simplification. By treating small subpopulations directly, we will get a picture of population dynamics on multiple scales—both the vast scale of global bacterial populations and the much smaller scale of local colonies. This will enable us to examine events akin to fixation in small populations (i.e., establishment at the colony level) in order to see how likely these are to trigger global fixation.

*Case 1: Complex adaptations with maladaptive intermediates.* Seeking a mathematical description of the waiting time for appearance and fixation of arbitrarily complex adaptations in bacterial populations, we begin with the following basic assumptions:

- A1– Cells reproduce asexually by binary fission;
- A2– The global population is of steady size, with local proliferation balanced by local extinction or near extinction;
- A3– All genetic intermediates en route to the complex adaptation, beginning with a new duplicate gene, confer a non-negligible selective disadvantage (relative to the wild type);
- A4– The global population is large enough to prevent fixation of these maladaptive intermediates; and
- A5– Selective benefit appears when the complex adaptation is complete.

Maruyama and Kimura described a mathematical model that captures the feast-or-famine nature of bacterial populations referred to in the second of the above assumptions, where rapid proliferation is balanced by frequent local extinctions [14]. The model to be developed here is similar in many respects to their *island* model. Specifically, we conceive of a global bacterial population consisting of numerous lines, each line starting with a single colonizing cell and proceeding through multiple rounds of

<sup>3</sup> Of the three possible changes to an incorrect base in this example, only one corrects it.

<sup>4</sup> This being  $(4N_e s_2)^{-1}$  where  $s_2$  is the fractional advantage conferred by the stage- $d$  allele.

binary fission until nutrient exhaustion causes line extinction. As each line expires, a new line is colonized with a cell chosen randomly from the whole population. Lines are taken to persist for a fixed period of  $\tau$  wild-type generations but to run their courses asynchronously, so that the size of the whole population remains nearly constant.

Because gene expression carries a metabolic cost, we expect duplicate genes to be maladaptive [15-18] unless their expression is substantially curtailed. Consistent with this, a recent study has found that a wide variety of mutations that reduce expression of a non-functional gene are rapidly fixed during extended serial culture [19]. Accordingly, in the present case we assume (A3 above) that all genetic intermediates en route to the complex adaptation are sufficiently maladaptive that natural selection prevents their accumulation. In the next section we consider the alternative case of selectively neutral intermediates.

In terms of the stage notation introduced above, the initial population is presumed to be at stage 0, which is taken as the wild-type state. Duplication events that provide starting points for the complex adaptation of interest convert stage-0 cells into stage-1 cells (carrying stage-1 alleles) which are now at a disadvantage relative to the wild type. A stage-1 cell must undergo  $d$  specific base changes (in any order) to be converted into a stage- $(d+1)$  cell, which then benefits from the complex adaptation. Because each line within the population begins with a single cell, we may similarly describe lines in terms of the stage of their colonizing cell. Note that this stage notation refers not to specific alleles or genotypes but rather to classes of alleles and genotypes defined with respect to their proximity to the complex adaptation.

A synopsis of the whole process by which a complex adaptation originates and becomes fixed, according to this model, would therefore run as follows. The initial population, consisting entirely of stage-0 cells (and therefore stage-0 lines), immediately gives rise to mutant cells, some of these being stage-1 cells and some potentially being of a higher stage. But because the rate of production of cells at any stage is proportional to the number of cells at the prior stage (with a very small proportionality constant), far fewer mutants are produced at each successive stage. The actual number of stage-3 cells, for example, is apt to be zero at any given time. Even so, there is some nonzero probability that a stage-3 cell will be produced within the next generation. That being the case, if we wait long enough a stage-3 cell is *bound* to appear, and indeed any number of appearances can be had by extending the wait sufficiently. Eventually, enough cells will have been produced at each stage for lines to be colonized by them. However, because lines colonized by cells at stages 1 through  $d$  grow slightly less well than wild-type lines, they will be unable to sustain their presence in the population through re-colonization, depending instead on continual regeneration by mutations. Therefore, prior to the first colonization by a stage- $(d+1)$  cell, the composition of the whole population may be described in terms of a time-averaged steady state where lines nominally of stages 0 through  $d$  make up certain proportions of the population and cells at stages 0 through  $d+1$  make up certain proportions of those lines. Because these proportions correspond to probabilistic expectations, they have meaning even if they are infinitesimally small (e.g., a genotype expected to make up one part in  $10^{30}$  of a global population numbering  $10^{20}$  individuals should be represented, on average, by one cell every  $10^{10}$  generations). Finally, when stage- $(d+1)$  cells begin to colonize lines, the steady state is broken by their ability to mount an increasing rate of re-colonization, which leads to rapid fixation of the complex adaptation. The aim is to calculate when this transition is expected to occur.

During the steady-state period, the expected proportion of cells in the population at any incomplete stage  $j$ ,  $F_j$ , may be expressed as the sum of contributions from lines colonized by cells up to stage  $j$ .<sup>5</sup> Because stage- $i$  lines (i.e., lines colonized by stage- $i$  cells) are present in the same overall proportion as stage- $i$  cells (namely  $F_i$ ) the steady state equations are of the form:

$$F_j = \sum_{i=0}^j \sigma_i \cdot F_i \cdot f_{i,j} \quad \text{for } 0 \leq j \leq d, \quad (3)$$

where  $f_{i,j}$  is the mean proportion of stage- $j$  cells in a stage- $i$  line ( $i \leq j$ ), and  $\sigma_i$  is the mean size of a stage- $i$  line divided by the mean size of a stage-0 line<sup>6</sup>, all of these means being the expected time-averaged values over the line duration.

Assuming mutations are rare, the steady-state population will consist predominantly of stage-0 cells. Consequently,  $F_0$  in the first term of the above summation is very nearly 1. Similarly,  $f_{i,j}$  is well approximated by 1 in the final term. With these simplifying approximations, the steady state relations represented by Equation 3 have the following recursive solution (see Supplement [20] for derivation) describing the prevalence of cells at stages 1 through  $d$ :

$$\begin{aligned} \alpha_1 &\approx \left( \frac{1}{1-\sigma} \right) \\ \alpha_2 &\approx \alpha_1 \left\{ 1 + \sigma \left[ \binom{2}{1} \cdot \alpha_1 \right] \right\} \\ \alpha_3 &\approx \alpha_1 \left\{ 1 + \sigma \left[ \binom{3}{1} \cdot \alpha_1 + \binom{3}{2} \cdot \alpha_2 \right] \right\} \\ &\quad \vdots \\ \alpha_d &\approx \alpha_1 \left\{ 1 + \sigma \cdot \sum_{k=1}^{d-1} \left[ \binom{d}{k} \cdot \alpha_k \right] \right\}, \end{aligned} \quad (4)$$

where  $\alpha_i \equiv F_i / f_{0,i}$ , and  $\sigma$  is taken to have the same value ( $< 1$ ) for all maladaptive stages.

To calculate values for  $f_{i,j}$  ( $i < j$ ), we first develop equations describing the expected number of stage- $j$  cells in a stage- $i$  line,  $n_{i,j}$ , as functions of the line age,  $a$ . The expected number of stage- $i$  cells in a stage- $i$  line (i.e., the number of cells unaffected by mutation since the line was colonized) is nearly equal to the size of the line (mutations being rare), which grows according to a simple doubling equation:

$$n_{i,i}(a) = 2^{\alpha(1+s_i)}, \quad (5)$$

where  $a$  is measured in wild-type generations (the standard time unit throughout these calculations), and  $s_i$  is the coefficient of selection (i.e., fractional advantage) of stage- $i$  cells relative to the wild type ( $s_i < 0$  for  $1 \leq i \leq d$ , indicating a selective disadvantage).

Now, consider a stage- $i$  line currently of age  $a$ . Suppose a mutation that happened when the line was of age  $t$  ( $\leq a$ ) converted a stage- $i$  cell into a stage- $(i+1)$  cell. That converted cell would have produced  $2^{(a-t)(1+s_{i+1})}$  descendants by the time the line age is  $a$ . Consequently, the total expected number of stage- $(i+1)$  cells in this stage- $i$  line has the following dependence on line age:

<sup>5</sup> Higher stage lines may be ignored because they are much more rare than stage- $j$  lines and, on top of this, would require reverse mutations in order to produce stage- $j$  cells.

<sup>6</sup> Because lines persist for a fixed time, those colonized by cells that divide more slowly undergo fewer divisions, resulting in fewer cells. Note that  $\sigma_0 = 1$ .

$$n_{i,i+1}(a) = r_{i+1} \int_0^a n_{i,i}(t) \cdot 2^{(a-t)(1+s_{i+1})} dt, \quad (6)$$

where  $r_{i+1}$  is the rate (per cell) at which spontaneous mutation converts stage- $i$  cells into stage- $(i+1)$  cells. Similar equations apply for cells at each subsequent stage within the same stage- $i$  line, the general form of these equations being:

$$n_{i,i+k}(a) = r_{i+k} \int_0^a n_{i,i+k-1}(t) \cdot 2^{(a-t)(1+s_{i+k})} dt \quad (7)$$

for  $0 \leq i \leq d$ ;  $1 \leq k \leq d+1-i$ ,

where again  $r_j$  is the per cell rate at which stage- $(j-1)$  cells are mutated to stage- $j$  cells.

Values for each  $f_{i,j}$  are calculated from Equation 7 as ratios of time-averaged means:

$$f_{i,j} = \int_0^\tau n_{i,j}(a) da / \int_0^\tau n_{i,i}(a) da \quad (8)$$

for  $0 \leq i \leq d$ ;  $i+1 \leq j \leq d+1$ ,

where again the constant  $\tau$  represents the duration of a line. Equations 7 and 8, in conjunction with Equations 4, enable calculation of the proportion of lines in the population colonized by cells up to stage  $d$ , these proportions being the succession of  $F_i$  values calculated as  $F_i = \alpha_i \cdot f_{0,i}$ , with  $1 \leq i \leq d$ . From these  $F_i$  values, the expected number of lines per wild-type generation colonized by stage- $(d+1)$  cells (which carry the complete complex adaptation) may be calculated as:

$$r_{col} = \frac{\ell}{\tau} \left[ f_{0,d+1} + \sum_{i=1}^d (\sigma \cdot F_i \cdot f_{i,d+1}) \right] \quad (9)$$

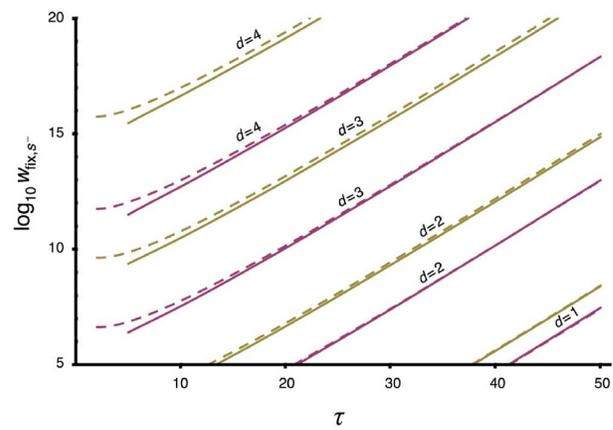
where  $\ell$  is the number of lines in the population,  $\ell/\tau$  therefore being the total number of lines colonized per wild-type generation. Because this equation accounts for colonizing cells originating from stage-0 lines through stage- $d$  lines, it covers all possible sources of stage- $(d+1)$  colonizers when no stage- $(d+1)$  lines yet exist. Since this is the case of interest, Equation 9 will be used to calculate the expected waiting time for stage- $(d+1)$  colonization events.

But because fixation of the complex adaptation is actually the event of interest, we first consider the probability that a stage- $(d+1)$  colonization event will cause fixation of the stage- $(d+1)$  genotype. This will be less than 1 because a single instance of colonization does not guarantee fixation. Specifically, stochastic simulation shows that the likelihood of fixation per stage- $(d+1)$  colonization event is approximately equal to the product of the line duration,  $\tau$ , and the fractional growth advantage conferred by the complex adaptation,  $s^+$ , provided that  $\tau s^+ < 1$  (see Supplement [20]).

So, multiplying Equation 9 by  $\tau s^+$  gives  $r_{fix}$ , the expected number of successful stage- $(d+1)$  colonization events (i.e., ones that will lead to fixation) per wild-type generation. Treating this as a Poisson process, the median waiting time for fixation (i.e., the time point at which fixation becomes the likely outcome<sup>7</sup>) is given by  $w_{fix} = \ln 2 / r_{fix}$ . When the full calculation is done (see Supplement [20]), the waiting time for the case of maladaptive intermediates,  $w_{fix,s^-}$ , is found to be well approximated by:

$$w_{fix,s^-} \approx \frac{2^\tau}{d! N \tilde{u} \tau^2 s^+} \left( \frac{1 + s^- - 2^{\tau s^-}}{\tau u} \right)^d, \quad (10)$$

<sup>7</sup> For the rare adaptations of interest here, the timescale for the actual selective sweep (following the colonization event destined to succeed in this) is negligible compared to the waiting time for colonization.



**Figure 1. Dependence of  $w_{fix,s^-}$  on  $\tau$  and  $d$  as represented by Equation 10 and full analytical solution.** Both  $w_{fix,s^-}$  and  $\tau$  are measured in wild-type generations (the standard unit of time). Dashed curves were obtained by numerical solution of Equations 3 through 9. Solid curves show the approximation provided by Equation 10. Curves are shown for  $s^- = -0.01$  (olive) and for  $s^- = -0.001$  (purple). Other parameter values are as listed in Table 1 for Equation 10. doi:10.5048/BIO-C.2010.4.f1

where time is measured in wild-type generations,  $s^-$  is the coefficient of selection (negative) conferred by incomplete stages (i.e., stages 1 to  $d$ ),  $N$  is the total population size,  $\tilde{u}$  is the rate of relevant gene duplication events (per cell), and  $u$  is the rate of specific base substitution events (per nucleotide site per cell; taken to have a single fixed value equivalent to  $r_{d+1}$  of Equation 7). Comparing the dependence of  $w_{fix,s^-}$  on  $d$  and  $\tau$  as represented by Equation 10 to the dependence observed by direct numerical solution of the above analytical equations (Equations 3–9) confirms that Equation 10 represents the overall behavior well, provided that the line duration,  $\tau$ , is about five or more generations (Figure 1). The fact that the numerator of the term enclosed in brackets becomes zero when  $s^-$  is zero is a limitation of the approximation, which breaks down as  $s^- \rightarrow 0$ . Equation 10 provides reasonable accuracy<sup>8</sup> only if the magnitude of  $s^-$  is at least  $10^{-4}$ . We next consider the alternative case where  $s^-$  is sufficiently close to zero to be of negligible effect, after which we will compare the expected behavior in the two cases.

*Case 2: Complex adaptations with neutral intermediates.* As in the prior case, duplicate alleles are here produced at a constant average rate in every cell. However, because they cannot in the present case be eliminated by purifying selection, they must be eliminated by deletion mutations instead. Two of the five basic assumptions therefore differ from the prior case, these being:

*A3'*– All genetic intermediates en route to the complex adaptation, beginning with a new duplicate gene, are selectively neutral (relative to the wild type);

*A4'*– Genome size is stabilized by a balance between neutral duplications and neutral deletions.

The prior case required a mathematical treatment of the propagation of mutant alleles in lines at various stages because selection was presumed to be relevant at all stages. Since negative selection is absent in the current case, selection only becomes relevant at

<sup>8</sup> Duplicate alleles with more mutations should be progressively more rare if purifying selection is the main way that they are eliminated. The approximations used to arrive at Equations 4 provide  $F_i$  values that are consistent with this if  $s^- = -10^{-4}$ , but may not be if  $s^- = -10^{-5}$ .

**Table 1: Nomenclature and standard parameter values**

Parameter	Symbol	Value	Used by Eqns:
Total population size	$N$	$10^{20}$ (ref. 21)	10, 16
Effective population size	$N_e$	$10^9$ (ref. 11)	17, 18, 20
Line duration	$\tau$	43 generations*	10, 16
Specific base mutation rate	$u$	$10^{-9}$ per site per cell†	10, 16, 17, 18, 20
Gene duplication rate	$\tilde{u}$	$10^{-8}$ per gene per cell [3]	10, 16, 20
Indel rate	$u_\Delta$	$10^{-6}$ per allele per cell‡	16, 20
Complexity of adaptation	$d$	Integer variable	10, 16, 17, 18, 20
Adaptive selection coefficient	$s^+$	+0.01	10, 16, 17, 18, 20
Maladaptive selection coefficient	$s^-$	-0.001 §	10, 17
Gene length	$L$	1000 bases	16, 20
Random substitution limit	$\lambda$	15	16, 20

\* Based on Equation 19 and standard values of  $N$  and  $N_e$ .

† Based on the expected total mutation rate for bacteria with a genome size of  $10^6$  base pairs [22].

‡ Effective removal of neutral duplicates requires that  $u_\Delta \gg \tilde{u}$ .

§ Assuming that bacteria typically express about  $10^3$  genes concurrently, and that the duplicate is expressed constitutively at a typical level.

|| With 75% of random base substitutions being non-silent (unpublished simulation) and 1.5 kcal/mol reduction in folding stability being typical for single amino acid substitutions (see Figure 5 of reference 23), 15 random base changes would be expected to destabilize most proteins.

the final stage. Notice, however, that the selective advantage of the complex adaptation,  $s^+$ , only entered Equation 10 at the point where the probability of fixation per colonization ( $\tau s^+$ ) was applied, indicating that positive selection has no significant effect on the pre-fixation steady state. We may therefore proceed in the current case by temporarily ignoring  $s^+$  in order to calculate the probability that a cell chosen randomly from the whole population in its initial steady state carries the complex adaptation. That probability, once calculated, replaces the contents of the square brackets in Equation 9 in order for the effect of  $s^+$  on  $w_{\text{fix}}$  to be accounted for as before.

Taking  $L$  to be the length in bases of the gene of interest and  $u_\Delta$  to be the mean rate at which insertion or deletion mutations (*indels*) appear in a single copy of that gene (per cell), a duplicate allele is expected to acquire  $3uL + u_\Delta$  mutations per generation<sup>9</sup>, including both indels and base substitutions. Applying a Poisson distribution, we then have the following expression for the expected number of duplicate alleles in any single cell that have not acquired any mutations:

$$\bar{n}_0 = \tilde{u} \int_0^\infty e^{-(3uL + u_\Delta)a} da = \frac{\tilde{u}}{3uL + u_\Delta}, \quad (11)$$

where the integration is over the cell's ancestral history (generations ago) from the present to the infinite past. The result is simply the ratio of the duplication rate per cell to the total rate of mutation per duplicate allele per cell. Because the expected rates of gene duplication events and mutations in duplicate genes are constant over time<sup>10</sup>, pristine duplicate alleles acquire their first mutation at precisely the same average rate that singly mutated alleles acquire their second mutation, and so on. This means that  $\bar{n}_i = \bar{n}_0$  for all positive integral values of  $i$ , making the right hand side of Equation 11 a general expression for the expected number of duplicate alleles per cell carrying any particular total number of mutations.

<sup>9</sup> The mean rate of mutation to a specific base,  $u$ , is here multiplied by three to obtain the total rate of base substitution per nucleotide site per cell.

<sup>10</sup> More precisely, mutations in duplicate alleles are constant until  $L$  is reduced by deletions. But since alleles cease to be of interest once deletions have occurred, this will not affect our calculation.

Making use of that, and assuming that indels preclude progression to the complex adaptation, we calculate that a single cell is expected to carry the following number of duplicate alleles with exactly  $k$  base substitutions and no indels:

$$\bar{n}_{k,0} = \left( \frac{\tilde{u}}{3uL + u_\Delta} \right) \left( \frac{3uL}{3uL + u_\Delta} \right)^k. \quad (12)$$

Given such an allele, and assuming  $k \geq d$ , the probability that it has the  $d$  base changes required for the complex adaptation is calculated as the number of ways to make  $k - d$  changes without disturbing the  $d$  key sites, divided by the total number of ways to make  $k$  changes, which for  $k \ll L$  is well approximated by:

$$P_{(d|k)} \approx \frac{3^{k-d} \binom{L-d}{k-d}}{3^k \binom{L}{k}} = 3^{-d} \binom{k}{d} \binom{L}{d}^{-1}. \quad (13)$$

Multiplying Equations 12 and 13 then gives the following as the expected number of duplicate alleles in any single cell that carry exactly  $k$  mutations,  $d$  of which are those required for the complex adaptation and none of which are indels:

$$\bar{n}_{k,d,0} \approx \frac{3^{k-d} \tilde{u} (uL)^k}{(3uL + u_\Delta)^{k+1}} \binom{k}{d} \binom{L}{d}^{-1}. \quad (14)$$

Having taken into account that indels prevent progression of the affected allele to the complex adaptation, we now need to account for the fact that random base changes have the same effect once they exceed the capacity of the encoded protein to sustain destabilizing amino acid substitutions (the *buffering effect* described in reference 24). Taking  $\lambda$  to represent the limit on random base substitutions per duplicate allele (beyond which alleles are considered nonviable), we have the following for the expected number of duplicate alleles in any single cell that carry the complex adaptation in working form:

$$\bar{n}_{c.a.} \approx \sum_{k=d}^{d+\lambda} \left[ \frac{3^{k-d} \tilde{u}(uL)^k}{(3uL + u_{\Delta})^{k+1}} \binom{k}{d} \binom{L}{d}^{-1} \right]. \quad (15)$$

Because this will be much less than one, it may be interpreted as the expected mean fraction of the population that carries the complex adaptation prior to the onset of its fixation. We therefore replace the expression in square brackets in Equation 9 with the above expression for the purpose of calculating the waiting time for fixation in the neutral case,  $w_{\text{fix},0}$ , resulting in the following final equation:

$$w_{\text{fix},0} \approx \frac{2^{\tau}}{\tau s^+ N} \left\{ \sum_{k=d}^{d+\lambda} \left[ \frac{3^{k-d} \tilde{u}(uL)^k}{(3uL + u_{\Delta})^{k+1}} \binom{k}{d} \binom{L}{d}^{-1} \right] \right\}^{-1}. \quad (16)$$

## RESULTS AND DISCUSSION

### Comparing predicted dependence of $w_{\text{fix}}$ on $d$

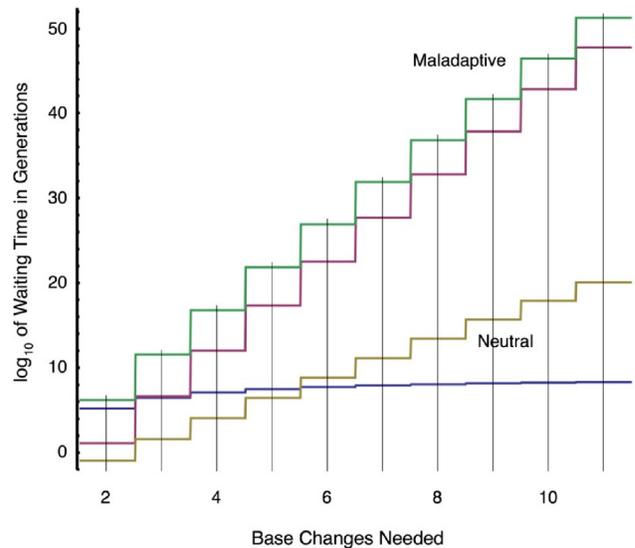
Of primary interest is the question of how the timescale for fixation of complex adaptations scales with the number of specific base changes required. Having argued that larger populations are more conducive to complex adaptation than smaller ones, we have restricted our analyses of the two cases (maladaptive and neutral intermediates) to large bacterial populations. It will therefore be informative to compare the results with the corresponding results reported recently by Lynch and Abegg [6]. For the case of maladaptive intermediates, and assuming a preexisting stage-1 allele (i.e., no duplication event required), their analysis gave the following result<sup>11</sup> for  $w_{\text{fix},s^-}$ :

$$w_{\text{fix},s^-} \approx \frac{|s^-|}{4d! N_e s^+} \left( \frac{|s^-|}{u} \right)^d, \quad (17)$$

where the effective population size,  $N_e$ , may be interpreted either as the actual size of an ideal population or as the effective size of a real population (see discussion below). In the case of neutral intermediates, and again assuming a preexisting stage-1 allele, they found  $w_{\text{fix},0}$  to be approximated as (see Equation 5b of reference 6):

$$w_{\text{fix},0} \approx \left[ u(4N_e s^+)^{1/d} \right]^{-1}. \quad (18)$$

Figure 2 compares the dependence of  $w_{\text{fix}}$  on the number of base changes needed for the complex adaptation, both for the case of maladaptive intermediates (as predicted by Equations 10 and 17) and for the case of neutral intermediates (as predicted by Equations 16 and 18). When four or more base changes are required for the complex adaptation, there is general agreement that neutral intermediates greatly reduce the waiting times, as expected. Contrary to expectation, though, Equations 17 and 18 predict that maladaptive intermediates lead to more rapid fixation than neutral intermediates when two changes are required, and that the two cases are nearly equivalent when three changes are required (compare blue and purple lines in Figure 2). Equations 10 and 16 show the expected behavior, with neutral intermediates consistently causing significant reduction of waiting times (green and olive lines in Figure 2).

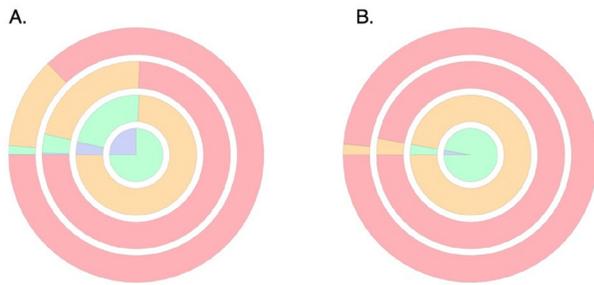


**Figure 2. Comparison of equations describing the dependence of  $w_{\text{fix}}$  on the number of base changes needed for a complex adaptation.** Shown are waiting times calculated according to the equations derived in this work (Equation 10 [green]; Equation 16 [olive]) and according to the equations from the work of Lynch and Abegg (Equation 17 [purple]; Equation 18 [blue]). Parameter values are as listed in Table 1, except for the rate of gene duplication. Because Equations 17 and 18 assume a preexisting allele as a starting point for the complex adaptation, Equations 10 and 16 have been adapted to conform to the same assumption. This was done by reducing the value of  $d$  by one (for Equations 10 and 16 only) and assigning a value of  $u(d+1)$  to  $\tilde{u}$ . For equation 10, the result is precisely equivalent to assuming that the population carries a preexisting allele within which the required base changes are to be made. For equation 16, the equivalence is approximate. doi:10.5048/BIO-C.2010.4.f2

Although Equations 10 and 17 give qualitatively similar results for the maladaptive case, the numerical predictions differ substantially—by five orders of magnitude for adaptations requiring two mutation events. Interestingly, Equation 10 predicts less steep increases in waiting time with increasing complexity of the adaptation than Equation 17 does. Successive increments in  $d$  cause the waiting time predicted by Equation 17 to increase by a factor of  $|s^-|/(ud)$ , which decreases gradually as  $d$  increases. The more rapid decline in steepness predicted by Equation 10 is attributable to the effects of population structure, as indicated by the presence of  $\tau$  within the bracketed expression raised to the power  $d$ . An intuitive explanation for this behavior is that the structured model accounts not only for mutations as a source of rare alleles but also for colonization events, from which these alleles receive a significant numerical boost.

In the neutral case, the solution of Lynch and Abegg (Equation 18) differs qualitatively from the other solutions in that it predicts increasing complexity to have a rapidly diminishing effect on the waiting time. As discussed above, this deviates significantly from the expected behavior. The staircase behavior produced by plotting the logarithm of  $w_{\text{fix}}$  as predicted by Equations 10, 16 and 17 (Figure 2), is consistent with the expectation that each additional base change needed to construct the complex adaptation multiplies the improbability of success and therefore the wait for success to be achieved. Because higher values of  $d$  provide more constructive mutation possibilities in the early stages en route to the complex adaptation, the staircases gradually become less steep as  $d$  increases. But Equation 18 predicts

<sup>11</sup> This is the inverse of their rate equation (Equation 6b of reference 6) rewritten in terms of the current nomenclature.



**Figure 3. Relative contribution of lines up to stage  $d = 3$  to the pool of cells up to stage 4 in Case 1.** Concentric pie charts show the relative contributions of lines at stage 0 (blue), stage 1 (green), stage 2 (orange), and stage 3 (pink) to the whole pool of cells in the population at stage 1 (center circle) through stage 4 (outer ring), the pool size being  $<1$  at higher stages. A) Assuming  $s^- = -0.01$ . B) Assuming  $s^- = -0.001$ . Other parameter values are as listed in Table 1 for Equation 10. doi:10.5048/BIO-C.2010.4.f3

a much more dramatic (and implausible, for the reasons stated above) flattening, with  $w_{\text{fix}}$  asymptotically approaching  $u^{-1}$  as  $d$  approaches infinity.

Behe and Snoko's earlier treatment of the neutral case shows the expected staircase behavior, but because they modeled the transient approach to the pre-fixation steady state, that behavior only becomes evident for fixation times in excess of about  $10^8$  generations.<sup>12</sup> Their assumptions about the effects of mutations differ substantially from the assumptions used in the current treatment. Specifically, they assumed that most amino acid substitutions caused by a single base change to a wild-type gene eliminate the function of the encoded protein [2]. Considerable care is needed in evaluating the experimental evidence on this issue, because what is judged to be non-functional depends on how the experiment was performed. Using a calibrated activity threshold where 'non-functional' means having activity comparable to or lower than that of a non-enzyme catalyst, I found only 5% of single substitutions to eliminate the activity of a ribonuclease [25]. Half of these cases of drastic impairment are explained by the fact that important active-site side chains were replaced, the remaining half apparently being attributable to drastic structural disruption [25,26]. Arguably, only mutations in the latter half, amounting to 2–3% of single substitutions, preclude evolutionary acquisition of new functions, since new functions typically require changes to the active site. But despite the different ways of accounting for the elimination of duplicate alleles by mutation, the fact that mutations must have this effect in the neutral case yields a broadly consistent picture.

### The role of subpopulations in fixation via maladaptive intermediates

A significant advantage of the structured population model used here is that it enables us to see how small subpopulations (lines) influence the process of fixing complex adaptations in large populations when the intermediates are maladaptive. Considering a complex adaptation requiring three specific changes to a duplicate allele ( $d = 3$ ), Figure 3 shows the relative contributions of lines at all relevant stages to the pool of cells at all stages ('pool' here being construed probabilistically, such that the pool size may be  $<1$ ). Stage- $i$  lines are seen to be the primary source

of stage- $i$  cells up to the last maladaptive stage (stage 3). Stage-4 lines do not yet exist in the pre-fixation steady state, but again lines of the highest relevant stage, stage 3, are the primary source of stage-4 cells. This pattern of lines of the highest relevant stage being the dominant source of cells at each stage becomes more pronounced as the selective disadvantage of intermediate stages is reduced (compare panels A and B of Figure 3).

Something like sequential fixation is therefore occurring on a small scale even in populations large enough to preclude actual fixation of maladaptive intermediates. That is, fixation of a complex adaptation only becomes likely on a timescale where all intermediate stages have had an opportunity to colonize lines. Once this happens, the primary route to the complex adaptation is through lines that were colonized by cells at the successive intermediate stages. Sequential *colonization* is therefore the dominant route to complex adaptations when selection precludes sequential fixation. Both routes depend on the proliferation of cells at each stage in order for the next stage to be reached, the difference being that colonization is local proliferation, whereas fixation is global.

### Limitations of $N_e$ for describing real populations

The complexity of real global populations means that mathematical descriptions of them must always rely on simplification. Although the island model used in this work is itself highly simplified, it does explicitly represent an important aspect of real bacterial populations that most analyses only represent implicitly—namely, the highly non-uniform descendant distribution that comes with a "boom and bust" mode of procreation. The fact that bacteria flourish wherever they can means that local proliferation is a common occurrence, which means that local extinctions or near-extinctions must likewise be common occurrences.<sup>13</sup> An ideal population, in which half of the cells in each generation are randomly chosen to undergo binary fission, would also produce a wide distribution of progeny, but not the extreme non-uniformity of real populations.

Still, because the ideal model is a simpler basis for calculation, a common compromise between simplicity and realism is to perform calculations for an ideal population with the aim of correcting the discrepancy by estimating an *effective* population size, meaning the size of an ideal population that should behave (with respect to the parameter of interest) like the real one. Equation 16 illustrates how this can work. Because  $\tau$  and  $N$  appear together outside the bracketed expression, it is possible to define  $N_e$  as a function of  $\tau$  and  $N$  such that  $w_{\text{fix},0}$  may be expressed in terms of  $N_e$  without explicit dependence on  $\tau$ . In fact, when the method of Maruyama and Kimura [14] is used to calculate  $N_e$  directly from the population model used here, we find it to be well approximated by (see Supplement [20]):

$$N_e \approx \tau N / 2^{\tau-1}, \quad (19)$$

which means Equation 16 can be rewritten as:

$$w_{\text{fix},0} \approx \frac{2}{s^+ N_e} \left\{ \sum_{k=d}^{d+k} \frac{3^{k-d} \tilde{u}(uL)^k}{(3uL + u_\Delta)^{k+1}} \binom{k}{d} \binom{L}{d}^{-1} \right\}^{-1}. \quad (20)$$

Since Equation 20 shows no explicit dependence on  $\tau$ , it gives the same result for all combinations of  $\tau$  and  $N$  that correspond to the same value of  $N_e$ . Notice that this includes the ideal population, where  $\tau = 1$  and  $N = N_e$ .

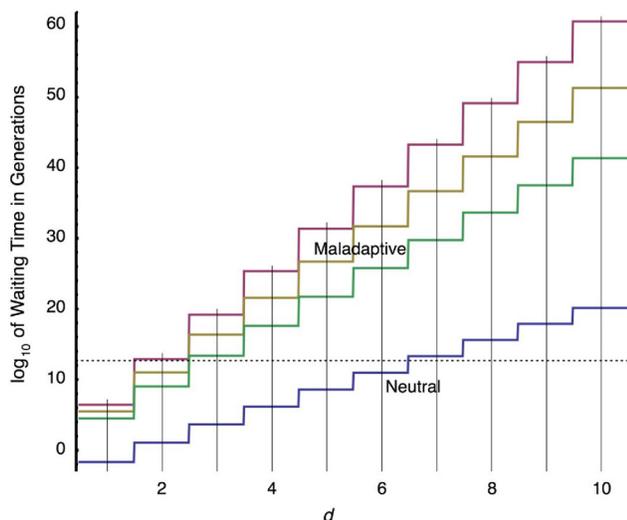
<sup>12</sup> See Figure 6 of reference 2 and the accompanying discussion. The staircases appear as slides in that plot because it represents the number of required changes as a continuous variable, whereas it is actually discrete.

<sup>13</sup> For every cell from some past generation that has billions of surviving descendants, billions from that generation have no descendants.

But it is not always possible to obtain a general expression in this way. Notice that replacing  $N$  in Equation 10 with  $N_e$  (according to Equation 19) does not eliminate  $\tau$ , indicating that the island model is not reducible to the ideal model when it comes to calculating waiting times in the case of maladaptive intermediates. In fact, the bracketed expression in Equation 10 combines  $s^-$  and  $\tau$  in a way that altogether prevents factoring  $\tau$  out. The implication of this is that the effect of purifying selection on complex adaptation is intrinsically dependent on the structure of the population, which means that real populations cannot be reduced to ideal ones in this case.

### Limitations on complex adaptation

Returning now to the subject of evolutionary innovation, we consider the implications of the analyses presented here for the emergence of new protein functions. In particular, to what extent might complex adaptations have contributed to functional innovations within paralogous protein families? As shown here, the answer depends in part on the metabolic cost of duplicate genes. Figure 4 provides a direct comparison of the two cases examined. In the case where the cost of the duplicate gene was high enough to have a selective effect, Equation 10 shows that complex adaptations within families would have been restricted to about two specific base changes. In the alternative case, where the metabolic cost was insignificant, Equation 16 (or 20) shows that complex adaptations requiring up to six base changes would have been feasible.



**Figure 4. Dependence of waiting time on  $d$  and  $s^-$ .** Shown are waiting times calculated by Equation 10 with  $s^- = -0.01$  (purple),  $s^- = -0.001$  (olive), and  $s^- = -0.0001$  (green), and by Equation 16 or 20 (blue) which assume  $s^- \approx 0$ . Other parameter values are as listed in Table 1 for the respective equations. The dashed line marks the boundary between feasible waiting times (below line) and waiting times that exceed the age of life on earth (above line), assuming  $10^3$  generations per year [27]. Notice that the predicted behavior for the neutral case (blue) is fully consistent with the trend seen in the non-neutral case, even though the two cases called for very different mathematical treatments.

[doi:10.5048/BIO-C.2010.4.f4](https://doi.org/10.5048/BIO-C.2010.4.f4)

The difference between these restrictions would have been important in the history of life whenever six changes could have produced an innovation that two changes could not have. Presumably some of these innovations happened and some did not, depending on the costs involved. Be that as it may, the most significant implication comes not from how the two cases contrast but rather how they cohere—both showing severe limitations to complex adaptation. To appreciate this, consider the tremendous number of cells needed to achieve adaptations of such limited complexity. As a basis for calculation, we have assumed a bacterial population that maintained an effective size of  $10^9$  individuals through  $10^3$  generations each year for billions of years. This amounts to well over a billion trillion opportunities (in the form of individuals whose lines were not destined to expire imminently) for evolutionary experimentation. Yet what these enormous resources are expected to have accomplished, in terms of combined base changes, can be counted on the fingers.

This striking disparity between the scale of the accomplishment and the scale of the resources needed to achieve it suggests an important corollary having to do with reasonable adjustments to the model or to parameter values. Specifically, since the predicted scale of complex adaptation is so small even when vast resources are allowed, it follows that this scale will be quite insensitive to adjustments. A thousand-fold increase in  $N$  or  $N_e$ , for example, would merely shift the staircase plots of Figure 4 down three units on the logarithmic scale, adding at most one base change to the limit. Increasing  $u$  helps by decreasing the steepness of the staircases, but even an implausibly generous adjustment leaves us with only a modest boost in the complexity of an attainable adaptation. Similarly, while it is easy to think of ways that the model could be made more realistic (e.g., by letting  $\tau$ ,  $s^+$ , or  $s^-$  vary to represent environmental variation, or by letting any of the mutation rates vary to represent the stress response, or by including homologous recombination or horizontal gene transfer), it is not apparent that any of these added complications would remove what appears to be a fundamental probabilistic limitation. In the end, the conclusion that complex adaptations cannot be very complex without running into feasibility problems appears to be robust.

Finally, this raises the question of whether these limits to complex adaptation present a challenge to the Darwinian explanation of protein origins. The problem of explaining completely new protein structures—new *folds*—is so acute that it can be framed with a very simple mathematical analysis [1]. Greater mathematical precision is needed when we consider the small-scale problem of functional diversification among proteins sharing a common fold. All such proteins are thought to have diverged through speciation and/or gene duplication events. In many cases, however, attempts to demonstrate the corresponding functional transitions in the laboratory require more than six base changes to achieve even weak conversions (see, for example, references 28–30). Although studies of this kind tend to be interpreted as supporting the Darwinian paradigm, the present study indicates otherwise, underscoring the importance of combining careful measurements with the appropriate population models.

1. Axe DD (2010) The case against a Darwinian origin of protein folds. *BIO-Complexity* 2010(1):1-12. [doi:10.5048/BIO-C.2010.1](https://doi.org/10.5048/BIO-C.2010.1)
2. Behe MJ, Snoke DW (2004) Simulating evolution by gene duplication of protein features that require multiple amino acid residues. *Protein Sci* 13: 2651-2664. [doi:10.1110/ps.04802904](https://doi.org/10.1110/ps.04802904)
3. Lynch M (2005) Simple evolutionary pathways to complex proteins. *Protein Sci* 14: 2217-2225. [doi:10.1110/ps.041171805](https://doi.org/10.1110/ps.041171805)

4. Durrett R, Schmidt D (2008) Waiting for two mutations: with applications to regulatory sequence evolution and the limits of Darwinian evolution. *Genetics* 180: 1501-1509. [doi:10.1534/genetics.107.082610](https://doi.org/10.1534/genetics.107.082610)
5. Orr HA (2002) The population genetics of adaptation: The adaptation of DNA sequences. *Evolution* 56: 1317-1330. [doi:10.1111/j.0014-3820.2002.tb01446.x](https://doi.org/10.1111/j.0014-3820.2002.tb01446.x)
6. Lynch M, Abegg A (2010) The rate of establishment of complex adaptations. *Mol Biol Evol* 27: 1404-1414. [doi:10.1093/molbev/msq020](https://doi.org/10.1093/molbev/msq020)

7. Kimura M (1980) Average time until fixation of a mutant allele in a finite population under continued mutation pressure: Studies by analytical, numerical, and pseudo-sampling methods. *P Natl Acad Sci USA* 77: 522-526. [doi:10.1073/pnas.77.1.522](https://doi.org/10.1073/pnas.77.1.522)
8. Kimura M, Ohta T (1969) The average number of generations until fixation of a mutant gene in a finite population. *Genetics* 61: 763-771.
9. Axe DD (2000) Extreme functional sensitivity to conservative amino acid changes on enzyme exteriors. *J Mol Biol* 301: 585-595. [doi:10.1006/jmbi.2000.3997](https://doi.org/10.1006/jmbi.2000.3997)
10. Povolotskaya IS, Kondrashov FA (2010) Sequence space and the ongoing expansion of the protein universe. *Nature* 465: 922-927. [doi:10.1038/nature09105](https://doi.org/10.1038/nature09105)
11. Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302: 1401-1404. [doi:10.1126/science.1089370](https://doi.org/10.1126/science.1089370)
12. Hughes AL (2002) Adaptive evolution after gene duplication. *Trends Genet* 18: 433-434. [doi:10.1016/S0168-9525\(02\)02755-5](https://doi.org/10.1016/S0168-9525(02)02755-5)
13. Chothia C, Gough G, Vogel C, Teichman SA (2003) Evolution of the protein repertoire. *Science* 300: 1701-1703. [doi:10.1126/science.1085371](https://doi.org/10.1126/science.1085371)
14. Maruyama T, Kimura M (1980) Genetic variability and effective population size when local extinction and recolonization of subpopulations are frequent. *P Natl Acad Sci USA* 77: 6710-6714. [doi:10.1073/pnas.77.11.6710](https://doi.org/10.1073/pnas.77.11.6710)
15. Dykhuizen D (1978) Selection for tryptophan auxotrophs of *Escherichia coli* in glucose-limited chemostats as a test of the energy conservation hypothesis of evolution. *Evolution* 32: 125-150. [doi:10.2307/2407415](https://doi.org/10.2307/2407415)
16. Wagner A (2005) Energy constraints on the evolution of gene expression. *Mol Biol Evol* 22: 1365-1374. [doi:10.1093/molbev/msi126](https://doi.org/10.1093/molbev/msi126)
17. Stoebel DM, Dean AM, Dykhuizen DF (2008) The cost of expression of *Escherichia coli* lac operon proteins is in the process, not in the products. *Genetics* 178: 1653-1660. [doi:10.1534/genetics.107.085399](https://doi.org/10.1534/genetics.107.085399)
18. Kuo C-H, Ochman H (2010) The extinction dynamics of bacterial pseudogenes. *PLoS Genetics* 6: e1001050. [doi:10.1371/journal.pgen.1001050](https://doi.org/10.1371/journal.pgen.1001050)
19. Gauger AK, Ebnert S, Fahey PF, Seelke R (2010) Reductive evolution can prevent populations from taking simple adaptive paths to high fitness. *BIO-Complexity* 2010(2):1-9. [doi:10.5048/BIO-C.2010.2](https://doi.org/10.5048/BIO-C.2010.2)
20. Supplement to this paper. [doi:10.5048/BIO-C.2010.4.s](https://doi.org/10.5048/BIO-C.2010.4.s)
21. Milkman R, Stoltzfus A (1988) Molecular evolution of the *Escherichia coli* chromosome. II. Clonal segments. *Genetics* 120: 359-366.
22. Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. *Genetics* 148: 1667-1686.
23. Guerois R, Nielsen JE, Serrano (2002) Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. *J Mol Biol* 320: 369-387. [doi:10.1016/S0022-2836\(02\)00442-4](https://doi.org/10.1016/S0022-2836(02)00442-4)
24. Axe DD (2004) Estimating the prevalence of protein sequences adopting functional enzyme folds. *J Mol Biol* 341: 1295-1315. [doi:10.1016/j.jmb.2004.06.058](https://doi.org/10.1016/j.jmb.2004.06.058)
25. Axe DD, Foster NW, Fersht AR (1998) A search for single substitutions that eliminate enzymatic function in a bacterial ribonuclease. *Biochemistry-US* 37: 7157-7166. [doi:10.1021/bi9804028](https://doi.org/10.1021/bi9804028)
26. Axe DD, Foster NW, Fersht AR (1999) An irregular beta-bulge common to a group of bacterial RNases is an important determinant of stability and function in barnase. *J Mol Biol* 286: 1471-1485. [doi:10.1006/jmbi.1999.2569](https://doi.org/10.1006/jmbi.1999.2569)
27. Levin BR (1981) Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. *Genetics* 99: 1-23.
28. Graber R, Kasper P, Malashkevich VN, Strop P, Gehring H *et al.* (1999) Conversion of aspartate aminotransferase into an L-aspartate  $\beta$ -decarboxylase by a triple active-site mutation. *J Biol Chem* 274: 31203-31208. [doi:10.1074/jbc.274.44.31203](https://doi.org/10.1074/jbc.274.44.31203)
29. Xiang H, Luo L, Taylor KL, Dunaway-Mariano D (1999) Interchange of catalytic activity within the 2-enoyl-coenzyme A hydratase/isomerase superfamily based on a common active site template. *Biochemistry-US* 38: 7638-7652. [doi:10.1021/bi9901432](https://doi.org/10.1021/bi9901432)
30. Ma H, Penning TM (1999) Conversion of mammalian 3 $\alpha$ -hydroxysteroid dehydrogenase to 20 $\alpha$ -hydroxysteroid dehydrogenase using loop chimeras: Changing specificity from androgens to progestins. *P Natl Acad Sci USA* 96: 11161-11166. [doi:10.1073/pnas.96.20.11161](https://doi.org/10.1073/pnas.96.20.11161)