Error and repair catastrophes: A two-dimensional phase diagram in the quasispecies model

Emmanuel Tannenbaum* and Eugene I. Shakhnovich
Harvard University, Cambridge, Massachusetts 02138, USA

(Received 9 June 2003; revised manuscript received 4 September 2003; published 15 January 2004)

This paper develops a two-gene, single fitness peak model for determining the equilibrium distribution of genotypes in a unicellular population which is capable of genetic damage repair. The first gene, denoted by σ_{via}, yields a viable organism with first-order growth rate constant $k>1$ if it is equal to some target ''master'' sequence $\sigma_{via,0}$. The second gene, denoted by σ_{rep}, yields an organism capable of genetic repair if it is equal to some target ''master'' sequence $\sigma_{rep,0}$. This model is analytically solvable in the limit of infinite sequence length, and gives an equilibrium distribution which depends on $\mu \equiv L\epsilon$, the product of sequence length and per base pair replication error probability, and ϵ_r, the probability of repair failure per base pair. The equilibrium distribution is shown to exist in one of the three possible ''phases.'' In the first phase, the population is localized about the viability and repairing master sequences. As ϵ_r exceeds the fraction of deleterious mutations, the population undergoes a ''repair'' catastrophe, in which the equilibrium distribution is still localized about the viability master sequence, but is spread ergodically over the sequence subspace defined by the repair gene. Below the repair catastrophe, the distribution undergoes the error catastrophe when μ exceeds $\ln k/\epsilon_r$, while above the repair catastrophe, the distribution undergoes the error catastrophe when μ exceeds $\ln k/f_{del}$, where f_{del} denotes the fraction of deleterious mutations.

DOI: 10.1103/PhysRevE.69.011902 PACS number(s): 87.23.Kg, 87.16.Ac, 64.90.+b

I. INTRODUCTION

To cope with genetic damage to their genomes, cellular organisms have developed a host of mechanisms to repair and, if necessary, replace damaged DNA. Environmental damage due to mutagens, metabolic free radicals, and radiation is repaired by enzymes which continuously scan the DNA molecule and repair the damaged portions. Replication errors are also repaired by several methods. In *Escherichia coli*, the DNA replicase Pol III has a built-in proofreading mechanism which results in a replication error probability of 10^{-7}–10^{-6} per base pair. Furthermore, immediately following replication, a second proofreading mechanism, known as mismatch repair, identifies and corrects mismatched base pairs. In *E. coli*, the mismatch repair system reduces the error probability to 10^{-10}–10^{-8} per base pair [1].

The DNA mismatch repair system is of considerable interest because it is believed that mismatch repair deficient strains, or mutators, play an important role in the emergence of antibiotic drug resistance and cancer [2–8]. Because mutators have mutation rates which are 10–10 000 times higher than wild-type strains, they can more rapidly adapt to hostile environments, thereby explaining their potential importance in understanding drug resistance. However, mutators can accumulate genetic damage much more rapidly than nonmutators, and hence can serve as an intermediate for the appearance of cancerous cells in multicellular organisms.

In an earlier work, we developed a simple, analytically solvable model to determine the equilibrium population of mutators in an asexual, unicellular population of replicating organisms [9] (we should point out a related study, Ref. [10], that was published almost simultaneously to us). The main result of this model was that at equilibrium, the population can exist in one of two ''phases.'' For sufficiently efficient repair, the population was shown to exist in a ''repairer'' phase, in which the fraction of repairers is a finite, positive quantity which depends only on the efficiency of repair and the fraction of the genome coding for repair. The equilibrium genotype of the population is localized about the ''master'' subsequence for which repair is functioning. When the repair efficiency drops below a critical value, the population delocalizes over the repair sequence subspace, and the fraction of repairers becomes zero in the limit of infinite genome length. This phase is naturally termed the ''mutator'' phase. In Ref. [9] the transition from the repairer to the mutator phases was called the repair catastrophe.

The solution of the model presented in Ref. [9] is incomplete, in that it describes the equilibrium behavior of the system in the low-mutation rate regime. This allowed one to assume that only point mutations were important, considerably simplifying the calculations. However, another phase transition has also been shown to occur when the mutation rate becomes too large. Above a critical mutation rate, replicative selection can no longer recover the loss of information due to genetic damage. This phenomenon is known as the error catastrophe, and was first predicted to occur by Eigen and co-workers [11,12]. It has since been studied in a number of theoretical papers [13–29] (and references therein), and has also been observed experimentally [30,31].

Because the model presented in our paper was only solved in the point-mutation regime, it did not incorporate the effect of the error catastrophe. The assumption underlying our initial approach was that mutators, despite their higher than wild-type mutation rates, are still viable organisms, and so live well below the error catastrophe.

The method used in our paper has since been generalized, so that it is no longer necessary to assume only point mutations. Thus, the interplay between the error and repair catastrophes can be studied, making this paper a natural extension

*Electronic address: etannenb@fas.harvard.edu

of the work presented in Ref. [9].

This paper is organized as follows. In the following section, we present a brief review of the quasispecies equations developed by Eigen, which are often the starting point for studies in evolutionary dynamics. We continue in Sec. III by developing the form of the quasispecies equations for our mutator model. We solve this model in Sec. IV. Specifically, we solve for the equilibrium fraction of viable organisms and viable repairers, which will allow the construction of a two-dimensional phase diagram incorporating both the error and repair catastrophes. We study the localization of the distribution and its limiting forms, comparing the results in certain cases with the corresponding results obtained in Ref. [9]. Finally, in Sec. V we present our conclusions.

II. THE QUASISPECIES EQUATIONS

The quasispecies equations are possibly the simplest for modeling the evolutionary dynamics of a unicellular, asexual population of replicating organisms. We let n_σ denote the number of organisms with genome σ, and κ_σ denote the first-order growth rate constant of an organism with genome σ. If $\kappa_m(\sigma,\sigma')$ is taken to be the first-order mutation rate constant from σ to σ', then the time evolution of n_σ is given by

$$\frac{dn_\sigma}{dt}=\kappa_\sigma n_\sigma+\sum_{\sigma'\neq\sigma}[\kappa_m(\sigma',\sigma)n_{\sigma'}-\kappa_m(\sigma,\sigma')n_\sigma]. \quad (1)$$

The mapping $K:\{\sigma\}\rightarrow\{\kappa_\sigma\}$ defines what is called the fitness landscape. In general, the fitness landscape will be time dependent, since organisms usually live in dynamic environments [8,16]. However, since in this paper we wish to study equilibrium behaviors, we take the fitness landscape to be static.

The conversion to Eigen's quasispecies equations is accomplished by converting from absolute populations to population fractions. Thus, we define $n=\Sigma_\sigma n_\sigma$ and $x_\sigma=n_\sigma/n$. When reexpressed in terms of the x_σ, the dynamical equations become

$$\frac{dx_\sigma}{dt}=\sum_{\sigma'}\kappa_m(\sigma',\sigma)x_{\sigma'}-\bar{\kappa}(t)x_\sigma, \quad (2)$$

where $\bar{\kappa}(t)\equiv\Sigma_\sigma\kappa_\sigma x_\sigma$ and $\kappa_m(\sigma,\sigma)\equiv\kappa_\sigma-\Sigma_{\sigma'\neq\sigma}\kappa_m(\sigma,\sigma')$. Note that $\bar{\kappa}(t)$ is simply the mean fitness of the population, and arises as a normalization term which ensures that $\Sigma_\sigma x_\sigma=1$ at all times.

We may simplify the notation further by defining $\vec{x}=(x_\sigma)$ to be the vector of population fractions, and $A\equiv[A_{\sigma\sigma'}=\kappa_m(\sigma',\sigma)]$ to be the matrix of mutation rate constants. We may also define $\vec{\kappa}$ to be the vector of growth rate constants, so that $\bar{\kappa}(t)=\vec{\kappa}\cdot\vec{x}$. Then we obtain

$$\frac{d\vec{x}}{dt}=A\vec{x}-(\vec{\kappa}\cdot\vec{x})\vec{x}. \quad (3)$$

Eigen showed that the system evolves to an equilibrium distribution given by the eigenvector corresponding to the largest eigenvalue of A [11,12]. If the equilibrium distribution is denoted by $\vec{x}_{equil}$, and the largest eigenvalue is denoted by λ, then it is clear that $\lambda=\vec{\kappa}\cdot\vec{x}_{equil}$.

To obtain an expression for $\kappa_m(\sigma',\sigma)$, let us assume that mutations occur due to replication errors. We take a per base pair replication error probability of $\epsilon_{\sigma'}$ for σ'. Let $l\equiv D_H(\sigma,\sigma')$ denote the Hamming distance between σ and σ'. Then it is a standard result (see, e.g., Ref. [12]) that

$$\kappa_m(\sigma',\sigma)=\kappa_{\sigma'}\left(\frac{\epsilon_{\sigma'}}{S-1}\right)^l(1-\epsilon_{\sigma'})^{L-l}. \quad (4)$$

III. A TWO-GENE MODEL INCORPORATING ERROR REPAIR

A simple model to study quasispecies dynamics with genetic repair is a two-gene, single fitness peak (SFP) model. We take our genome to have an alphabet size S, composed of "bases" $0,1,\ldots,S-1$. The first gene, denoted by σ_{via}, has length L_{via}, and controls the viability of the organism. We assume that there is a unique, "fit" sequence $\sigma_{via,0}$ such that $\kappa_\sigma=k>1$ if $\sigma_{via}=\sigma_{via,0}$. Otherwise, $\kappa_\sigma=1$. There is no loss of generality in assuming $\kappa_\sigma=1$ for the unfit sequences, since time may always be rescaled so that the unfit κ_σ becomes 1.

The second gene, denoted by σ_{rep}, has length L_{rep}, and is responsible for the enzymatic machinery involved in repair. As with viability, there is a unique sequence, $\sigma_{rep,0}$, for which repair is functioning, and has a per base pair failure probability of ϵ_r. For all other σ_{rep} repair is inactivated and the organism is a mutator.

For the mutators, the per base pair replication error probability is taken to be ϵ. Thus, for $\sigma_{rep,0}$, the per base pair replication error probability is $\epsilon_r\epsilon$. If ϵ_σ denotes the per base pair replication error probability of genome σ, then $\epsilon_\sigma=\epsilon_r\epsilon$ if $\sigma_{rep}=\sigma_{rep,0}$, and ϵ otherwise.

This model is clearly an oversimplification of the actual genome and replication dynamics of an organism. Nevertheless, a two-gene, SFP model is probably the simplest for studying evolutionary dynamics with genetic damage repair, and it is therefore a natural starting point before considering more complicated systems. Despite its simplicity, this model still yields sufficiently rich behavior to be of interest.

To determine the equilibrium distribution of genotypes in this model, note that, by symmetry, we may assume that x_σ depends only on the Hamming distance $l\equiv D_H(\sigma_{via},\sigma_{via,0})$ and $l'\equiv D_H(\sigma_{rep},\sigma_{rep,0})$ [22]. We define the Hamming class $C_H(l,l')=\{\sigma=\sigma_{via}\sigma_{rep}|D_H(\sigma_{via},\sigma_{via,0})=l,D_H(\sigma_{rep},\sigma_{rep,0})=l'\}$. It is readily shown that $C_H(l,l')$ contains

$$C_{ll'}\equiv\binom{L_{via}}{l}\binom{L_{rep}}{l'}(S-1)^{l+l'}$$

elements. Since x_σ is assumed to depend only on the Hamming class of σ, we may define $x_{ll'}=x_\sigma$ for $\sigma\in C_H(l,l')$.

We may also note that $\kappa_\sigma = k$ if $l=0$ and 1 otherwise, so that κ_σ depends only on l. Therefore, we redenote κ_σ by κ_l. Similarly, we redenote ϵ_σ by $\epsilon_{l'}$. Finally, define $z_{ll'} = C_{ll'} x_{ll'}$ to be the total fraction of the population in $C_H(l,l')$.

We wish to express the quasispecies equations in terms of the $z_{ll'}$. The final result, derived in Appendix A, is given by

$$\frac{dz_{ll'}}{dt} = \sum_{l_1=0}^{L_{via}-l} \sum_{l_2=0}^{l} \sum_{l_1'=0}^{L_{rep}-l'} \sum_{l_2'=0}^{l'} \binom{L_{via}-l-l_1+l_2}{l_2} \binom{l_1+l-l_2}{l_1} \binom{L_{rep}-l'-l_1'+l_2'}{l_2'} \binom{l_1'+l'-l_2'}{l_1'} \kappa_{l_1+l-l_2} \epsilon_{l_1'+l'-l_2'}^{l_2+l_2'} (1 - \epsilon_{l_1'+l'-l_2'})^{L_{via}+L_{rep}-l-l'-l_1-l_1'} \left(\frac{\epsilon_{l_1'+l'-l_2'}}{S-1}\right)^{l_1+l_1'} \left(1-\frac{\epsilon_{l_1'+l'-l_2'}}{S-1}\right)^{l+l'-l_2-l_2'} z_{l_1+l-l_2,l_1'+l'-l_2'} - [(k-1)z_0+1] z_{ll'}, \tag{5}$$

where $z_0 \equiv \sum_{l'=0}^{L_{rep}} z_{0l'}$ is simply the total fraction of viable organisms.

We now let the viability and repair sequence lengths L_{via}, L_{rep} approach ∞, while keeping $\alpha \equiv L_{via}/L_{rep}$, $\mu \equiv L\epsilon$, and ϵ_r fixed, where $L \equiv L_{via} + L_{rep}$ is the total sequence length. Since $\epsilon_{l_1'+l'-l_2'} = \epsilon$ or $\epsilon_r \epsilon$, it is clear that $\mu_{l_1'+l'-l_2'} \equiv L\epsilon_{l_1'+l'-l_2'}$ remains fixed in the limit $L \to \infty$.

We claim that, for a given l, l', the only terms which survive the limiting process are the $l_1 = l_1' = 0$ terms. We then note that, as $L_{via}, L_{rep} \to \infty$,

$$\binom{L_{via}-l+l_2}{l_2} \epsilon_{l'-l_2'}^{l_2} \to \frac{1}{l_2!} (L_{via}\epsilon_{l'-l_2'})^{l_2} = \frac{1}{l_2!} \left(\frac{\alpha}{\alpha+1} \mu_{l'-l_2'}\right)^{l_2} \tag{6}$$

and

$$(1-\epsilon_{l'-l_2'})^{L_{via}-l} \to e^{-(\alpha/\alpha+1)\mu_{l'-l_2'}}. \tag{7}$$

Taking similar limits for the L_{rep} terms, we obtain the infinite sequence length equations

$$\frac{dz_{ll'}}{dt} = \sum_{l_2=0}^{l} \sum_{l_2'=0}^{l'} \frac{\kappa_{l-l_2}}{l_2! l_2'!} \alpha^{l_2} \left(\frac{\mu_{l'-l_2'}}{\alpha+1}\right)^{l_2+l_2'} e^{-\mu_{l'-l_2'}} z_{l-l_2,l'-l_2'} - \overline{\kappa}(t) z_{ll'}. \tag{8}$$

To understand why only the $l_1 = l_1' = 0$ terms survive, let us consider the mutational contribution from those $z_{l_1+l-l_2,l_1'+l'-l_2'}$ for which at least one of $l_1, l_1' > 0$. A $\sigma' \in C_H(l_1+l-l_2, l_1'+l'-l_2')$ was obtained from a $\sigma \in C_H(l,l')$ by changing l_1 of the $L_{via}-l$ bases in σ_{via} which were equal to the corresponding bases in $\sigma_{via,0}$, and similarly for l_1' and σ_{rep}. Therefore, for σ_{via}' to mutate to σ_{via}, l_1 of the changed bases must back mutate to the corresponding bases in $\sigma_{via,0}$. However, in the limit of infinite sequence length, the number of unchanged bases in σ_{via}', given by $L_{via}-l_1-l+l_2$, becomes infinite, and so the probability of a mutation occurring at one of those bases approaches 1, so that the probability of back mutation goes to 0.

This heuristic argument is given a more rigorous justification in Appendix A. As a simple check, we also ensure that total population is conserved in the limiting process.

IV. SOLUTION OF THE MODEL

A. The phase diagram

We begin our solution of the model by computing the equilibrium values of z_0 and z_{00}. We begin with the dynamical equations for $z_{0l'}$,

$$\frac{dz_{0l'}}{dt} = \frac{k}{l'!} \left(\frac{\epsilon_r \mu}{\alpha+1}\right)^{l'} e^{-\epsilon_r \mu} z_{00} + k e^{-\mu} \sum_{l_1'=0}^{l'-1} \frac{1}{l_1'!} \left(\frac{\mu}{\alpha+1}\right)^{l_1'} z_{0,l'-l_1'} - [(k-1)z_0+1] z_{0l'}. \tag{9}$$

We may sum from $l' = 0 - \infty$ to obtain the dynamical equation for z_0. Together with the dynamical equation for z_{00}, we have the pair of equations

$$\frac{dz_0}{dt} = k(e^{-(\alpha/\alpha+1)\epsilon_r \mu} - e^{-(\alpha/\alpha+1)\mu}) z_{00} + [k e^{-(\alpha/\alpha+1)\mu} - (k-1)z_0 - 1] z_0, \tag{10}$$

TABLE I. The possible equilibrium values of (z_0, z_{00}) as a function of μ and ϵ_r.

z_0	z_{00}
$\frac{ke^{-\epsilon_r\mu}-1}{k-1}$	$\frac{e^{-\epsilon_r\mu}-e^{-(\alpha/\alpha+1)\mu}}{e^{-(\alpha/\alpha+1)\epsilon_r\mu}-e^{-(\alpha/\alpha+1)\mu}} z_0$
$\frac{ke^{-(\alpha/\alpha+1)\mu}-1}{k-1}$	0
0	0

$$\frac{dz_{00}}{dt} = [ke^{-\epsilon_r\mu} - (k-1)z_0 - 1]z_{00}. \tag{11}$$

We may obtain the equilibrium solution of these equations by setting the left-hand sides to zero. A summary of the possible solutions is given in Table I. We should note that with z_0 and z_{00} calculated, it is a relatively simple matter to recursively compute the equilibrium values of all the $z_{ll'}$. This is detailed in Appendix B.

We need to map out the regions in the (μ, ϵ_r) parameter space for which the various solutions are valid. First of all, note that we must have $z_0 \in [0,1]$ and $z_{00} \in [0, z_0]$. For the first solution set to hold, we must therefore have $0 \leqslant ke^{-\epsilon_r\mu} - 1 \leqslant k-1$. The second inequality is automatically satisfied. For the first inequality to hold, we must have $\epsilon_r\mu \leqslant \ln k$. In order for $z_{00} \in [0, z_0]$, we must then have $0 \leqslant (e^{-\epsilon_r\mu} - e^{-(\alpha/\alpha+1)\mu}) / (e^{-(\alpha/\alpha+1)\epsilon_r\mu} - e^{-(\alpha/\alpha+1)\mu}) \leqslant 1$. Again, the second inequality is automatically satisfied, but the first only holds when $\epsilon_r \leqslant \alpha/\alpha+1$. Therefore, the first solution pair is only valid when $\epsilon_r\mu \leqslant \ln k$, and $\epsilon_r \leqslant \alpha/(\alpha+1)$. However, the other two solution pairs may still yield physical values for (z_0, z_{00}) in the domain of validity of the first solution pair. To resolve this issue, we note that we want a solution which gives $z_{00} \to 1$ as $\epsilon_r \to 0$. That is, if repair is perfect, then at equilibrium the population should only consist of viable repairers. Therefore, as $\epsilon_r \to 0$, we expect the first solution pair to hold, since it gives the correct limiting behavior. By continuity, the first solution pair holds over the set $\Omega_1 \equiv \{(\mu, \epsilon_r) \in [0,\infty) \times [0,1] \,|\, \epsilon_r\mu \leqslant \ln k, \epsilon_r \leqslant \alpha/\alpha+1\}$.

As ϵ_r is increased beyond $\alpha/\alpha+1$, the first solution is no longer valid, but the second solution may still be valid if $0 \leqslant ke^{-(\alpha/\alpha+1)\mu} - 1 \leqslant k-1$. Again, the second inequality is automatically satisfied, while the first only holds when $(\alpha/\alpha+1)\mu \leqslant \ln k$. The third solution pair may still be physical in the domain of validity of the second solution pair. To resolve this issue, we may note that we want a solution which gives $z_0 \to 1$ as $\mu \to 0$. That is, in the limit of no replication errors, all of the population is viable. Therefore, as $\mu \to 0$ with $\epsilon_r > \alpha/\alpha+1$, we expect the second solution pair to hold, since it gives the correct limiting behavior. By continuity, the second solution pair holds over the set $\Omega_2 \equiv \{(\mu, \epsilon_r) \in [0,\infty) \times [0,1] \,|\, (\alpha/\alpha+1)\mu \leqslant \ln k, \epsilon_r > (\alpha/\alpha+1)\}$. The third solution is then the solution over the domain $\Omega_3 \equiv ([0,\infty) \times [0,1])/(\Omega_1 \cup \Omega_2)$.

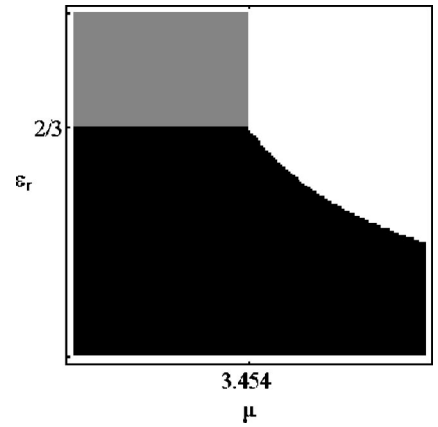

FIG. 1. Diagram illustrating the solution domains Ω_1 (black), Ω_2 (gray), and Ω_3 (white). The μ axis is labeled only at $\ln k/\epsilon_{r,crit} \approx 3.454$, and the ϵ_r axis is labeled only at $\epsilon_{r,crit} = 2/3$.

Figure 1 illustrates the three solution domains $\Omega_1, \Omega_2, \Omega_3$ for $\alpha = 2$, $k = 10$. In Ω_1, the population is clustered within finite Hamming distances about the viable and repair "master" sequences. A finite positive fraction of the population is viable, and of the viable organisms, a finite positive fraction of the population is capable of repair. As ϵ_r is increased beyond $\epsilon_{r,crit} \equiv \alpha/\alpha+1$, so that $(\mu, \epsilon_r) \in \Omega_2$, the population becomes delocalized over the repair gene subspace, and the fraction of repairers becomes zero. This phenomenon is known as the repair catastrophe, and was first predicted in Ref. [9]. Nevertheless, if μ is still sufficiently small so that $(\alpha/\alpha+1)\mu \leqslant \ln k$, then the population is still localized about the viable "master" sequence, and the fraction of viable organisms is positive. In Ω_3, μ is sufficiently large that the population completely delocalizes over the gene sequence space, a phenomenon known as the error catastrophe.

We may use our three solution pairs to compute $\lambda = \overline{\kappa}$ for the three solution domains, or phases. We have $\lambda = (k-1)z_0 + 1$, so that

$$\lambda(\mu, \epsilon_r) = \begin{cases} ke^{-\epsilon_r\mu} & \text{for } (\mu, \epsilon_r) \in \Omega_1 \\ ke^{-(\alpha/\alpha+1)\mu} & \text{for } (\mu, \epsilon_r) \in \Omega_2 \\ 1 & \text{for } (\mu, \epsilon_r) \in \Omega_3. \end{cases} \tag{12}$$

Figure 2 shows a plot of λ versus (μ, ϵ_r) for $\alpha = 2$, $k = 10$. Figure 3 shows the corresponding plot for z_{00}.

The error and repair catastrophes both arise as a result of the interplay between two competing effects: (1) The selective advantage for being viable and for being a repairer and (2) The entropic tendency to be unviable and a mutator. For a sufficiently low mutation rate, the selective advantage for being viable is strong enough to localize the population about $\sigma_{via,0}$. However, when the mutation rate exceeds a critical value, the selective advantage for being viable is no longer sufficiently strong to localize the population about the viable master sequence, and the population delocalizes over the entire viability subspace. Below the repair catastrophe, this occurs when the effective growth rate of the viable, repairing sequence $\sigma_{via,0}\sigma_{rep,0}$ becomes comparable to the growth rates of the nonviable sequences, i.e., when $ke^{-\epsilon_r\mu} = 1$. Above the repair catastrophe, there is no longer any

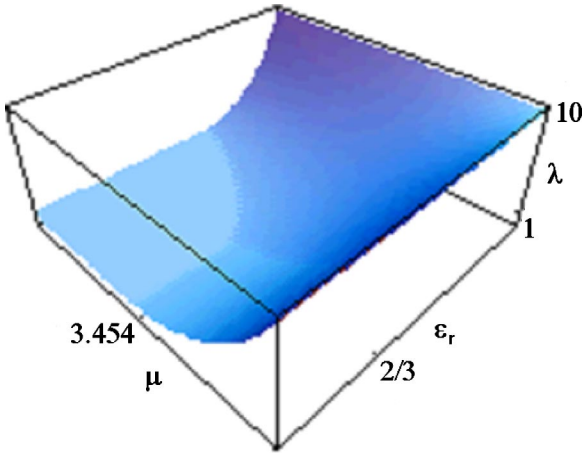

FIG. 2. Plot of λ for $\alpha=2$, $k=10$.

preference for being a repairer. The effective growth rate of the viable sequences due to mutation off of the viability peak is $ke^{-\alpha\mu/\alpha+1}$, hence, above the repair catastrophe, the error catastrophe occurs when $ke^{-\alpha\mu/\alpha+1}=1$.

Below the error catastrophe, viable repairers have a slower rate of mutation off of the viability peak than viable mutators, and hence have a higher effective growth rate. For sufficiently efficient repair, this discrepancy causes localization about $\sigma_{rep,0}$. However, when the repair error probability exceeds $\epsilon_{r,crit}=\alpha/\alpha+1=L_{via}/L$, the selective advantage for being a repairer is no longer sufficient to localize the population, and the distribution undergoes the repair catastrophe, in which the distribution delocalizes over the repair subspace. Note that $\epsilon_{r,crit}$ is simply the fraction of deleterious mutations and increases with increasing α. This makes sense, since, the greater the fraction of deleterious mutations, the greater the relative advantage for being a repairer. Thus, for large α, repair has to be highly inefficient before the repair catastrophe occurs. Conversely, for low α, repair has to be highly efficient to give the repairers a sufficiently large advantage for the distribution to localize about the repair master sequence.

It should be noted that the error and repair catastrophes are similar to thermodynamic phase transitions, in that they both arise from a competition between maximum fitness (minimum energy) and maximal entropy. When the replication and repair error probabilities are sufficiently low, maximal fitness (minimum energy) wins out, leading to localization on the sequence space. When the replication or repair error probabilities are sufficiently high, maximal entropy wins out, leading to delocalization on the sequence space. While not exact, this analogy nevertheless conceptually describes the origin of the phases observed in this study.

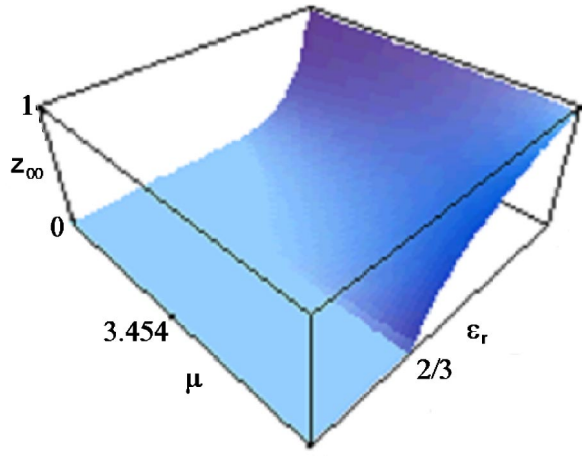

FIG. 3. Plot of z_{00} for $\alpha=2$, $k=10$.

In this vein, we should also note that the repair catastrophe is related to the evolution of mutational robustness [23–29]. In a series of studies initiated by Schuster and Swetina [23], and continued by Wilke and others [24–29], it was shown that at sufficiently high mutation rates, a quasispecies will evolve to a region of sequence space with maximal mutational support, and not necessarily to a region of maximal fitness. Essentially, at high mutation rates, the entropic cost of remaining localized to a highly fit but relatively nonrobust fitness peak becomes sufficiently large that the population favors residing on a less fit but more robust fitness "plateau." The repair catastrophe is similar to this phenomenon, in that, when repair becomes sufficiently inefficient, the entropic cost of being localized about the repairing sequence is no longer outweighed by the advantage of being a repairer, resulting in delocalization over the repair subspace.

B. Localization lengths

The final set of quantities we wish to compute are the following localization lengths of the equilibrium distribution:

$$\langle l' \rangle_{via} \equiv \sum_{l'=1}^{\infty} l' z_{0l'}, \tag{13}$$

$$\langle l \rangle_{rep} \equiv \sum_{l=1}^{\infty} l z_{l0}, \tag{14}$$

$$\langle l \rangle \equiv \sum_{l=1}^{\infty} \sum_{l'=0}^{\infty} l z_{ll'}, \tag{15}$$

$$\langle l' \rangle \equiv \sum_{l'=1}^{\infty} \sum_{l=0}^{\infty} l' z_{ll'}. \tag{16}$$

Using the dynamical equations for the $z_{ll'}$ we may compute the various localization lengths at equilibrium. The basic idea is to obtain an expression for the time derivatives of the localization lengths in terms of the localization lengths themselves, and then solve for the equilibrium value. We illustrate the technique for $\langle l' \rangle_{via}$. We have

$$\frac{d\langle l'\rangle_{via}}{dt}=\sum_{l'=1}^{\infty} l'\frac{dz_{0l'}}{dt}=k\frac{\epsilon_r\mu}{\alpha+1}e^{-(\alpha/\alpha+1)\epsilon_r\mu}z_{00}+ke^{-\mu}\sum_{l_1'=0}^{\infty}\sum_{l'=l_1'+1}^{\infty}\frac{l'-l_1'+l_1'}{l_1'!}\left(\frac{\mu}{\alpha+1}\right)^{l_1'}z_{0,l'-l_1'}-[(k-1)z_0+1]\langle l'\rangle_{via}$$

$$=k\frac{\epsilon_r\mu}{\alpha+1}e^{-(\alpha/\alpha+1)\epsilon_r\mu}z_{00}+ke^{-\mu}\sum_{l_1'=0}^{\infty}\frac{1}{l_1'!}\left(\frac{\mu}{\alpha+1}\right)^{l_1'}[\langle l'\rangle_{via}+l_1'(z_0-z_{00})]-[(k-1)z_0+1]\langle l'\rangle_{via}$$

$$=k\frac{\epsilon_r\mu}{\alpha+1}e^{-(\alpha/\alpha+1)\epsilon_r\mu}z_{00}+ke^{-(\alpha/\alpha+1)\mu}\left(\langle l'\rangle_{via}+\frac{\mu}{\alpha+1}(z_0-z_{00})\right)-[(k-1)z_0+1]\langle l'\rangle_{via}, \tag{17}$$

so at equilibrium we obtain

$$\langle l'\rangle_{via}=\frac{k\mu}{\alpha+1}\frac{(\epsilon_r e^{-(\alpha/\alpha+1)\epsilon_r\mu}-e^{-(\alpha/\alpha+1)\mu})z_{00}+e^{-(\alpha/\alpha+1)\mu}z_0}{(k-1)z_0+1-ke^{-(\alpha/\alpha+1)\mu}}. \tag{18}$$

To compute the remaining localization lengths using the above approach, we first need to compute $z_0'\equiv\Sigma_{l=0}^{\infty}z_{l0}$. Note that z_0' is simply the total fraction of repairers. We compute z_0' by evaluating $dz_0'/dt=\Sigma_{l=0}^{\infty}dz_{l0}/dt$. The result is an expression in terms of z_0', z_0, and z_{00}, which may be solved at equilibrium to obtain

$$z_0'=\frac{(k-1)e^{-\epsilon_r\mu/\alpha+1}}{(k-1)z_0+1-e^{-\epsilon_r\mu/\alpha+1}}z_{00}. \tag{19}$$

We then obtain

$$\langle l\rangle_{rep}=\frac{\alpha\epsilon_r\mu}{\alpha+1}e^{-\epsilon_r\mu/\alpha+1}\frac{(k-1)z_{00}+z_0'}{(k-1)z_0+1-e^{-\epsilon_r\mu/\alpha+1}} \tag{20}$$

$$(\langle l\rangle,\langle l'\rangle)=\frac{1}{z_0}(0,\langle l'\rangle_{via})+\frac{\mu}{(k-1)(\alpha+1)z_0}[(k-1)z_0+1-(1-\epsilon_r)z_0'-(k-1)(1-\epsilon_r)z_{00}](\alpha,1). \tag{21}$$

C. Limiting forms of the distribution

It is instructive to study the behavior of the distribution in the following limiting cases: (1) $\mu\to 0$. (2) $\alpha\to\infty$. (3) $\alpha\to 0$. We handle each of these cases in turn.

1. Behavior in the limit $\mu\to 0$

For $\mu\to 0$, note that $z_0\to 1$, and $z_{00}\to(\alpha/\alpha+1-\epsilon_r)/[\alpha/\alpha+1(1-\epsilon_r)]=1-\epsilon_r/[\alpha(1-\epsilon_r)]$, below the repair catastrophe. We may also note that $z_0'\to z_{00}$, and $(\langle l\rangle,\langle l'\rangle)\to(0,\langle l'\rangle_{via})$. This makes sense since in the limit $\mu\to 0$, we expect that the entire population becomes viable. For the same reason, $\langle l\rangle_{rep}\to 0$ as $\mu\to 0$. Finally, as $\mu\to 0$ for $\epsilon_r<\epsilon_{r,crit}$,

$$\langle l'\rangle_{via}\to\frac{1}{(\alpha+1)\left(\frac{\alpha}{\alpha+1}-\epsilon_r\right)}\left[1-(1-\epsilon_r)\left(1-\frac{\epsilon_r}{\alpha(1-\epsilon_r)}\right)\right]$$

$$=\frac{1}{(\alpha+1)(\epsilon_{r,crit}-\epsilon_r)}\epsilon_r\frac{\alpha+1}{\alpha}$$

$$=(1-\epsilon_{r,crit})/\epsilon_{r,crit}\times\epsilon_r/(\epsilon_{r,crit}-\epsilon_r).$$

As expected, these results agree with the point-mutation limit expressions obtained in Ref. [9].

2. Behavior in the limit $\alpha\to\infty$

For $\alpha\to\infty$ we obtain $\epsilon_{r,crit}=1$. Hence, we are always below the repair catastrophe. As long as $\mu<\ln k/\epsilon_r$, then $z_{00}\to z_0=(ke^{-\epsilon_r\mu}-1)/(k-1)$. Thus, the solution pairs presented in Table I reduce to two possible solutions. Either $z_0=z_{00}=(ke^{-\epsilon_r\mu}-1)/(k-1)$ if we are below the error catastrophe ($\epsilon_r\mu<\ln k$) or $z_0=z_{00}=0$ if we are above the error catastrophe. This means that the fraction of mutators is always zero. To understand this behavior, note that the probability of mutating off of the repairer sequence is $1-e^{-\epsilon_r\mu/\alpha+1}$, while the probability of mutating off of a mutator sequence is $1-e^{-(\mu/\alpha+1)}$. Both go to 0 as $\alpha\to\infty$. However, since for $\epsilon_r<1$ the repairer sequence has a greater selective advantage than the mutator sequence, the repair strain comes to dominate the population. Only at $\epsilon_r=1$ is there an ambiguity, since $(e^{-\epsilon_r\mu}-e^{-\alpha\mu/\alpha+1})/(e^{-\alpha\epsilon_r\mu/\alpha+1}-e^{-\alpha\mu/\alpha+1})\to 0/0$, which is undefined. Physically, since at $\epsilon_r=1$ there is no difference between what we call a repairer and a mutator, we expect delocalization over the repair subspace, so that $z_{00}\to 0$.

We may also note that $z_0'\to z_{00}/z_0=1$ for $\epsilon_r<1$, and 0 for $\epsilon_r=1$. We also have $\langle l\rangle_{rep}\to k\epsilon_r\mu e^{-\epsilon_r\mu}/(ke^{-\epsilon_r\mu}-1)$. Also, $\langle l'\rangle_{via}\to 0$, for $\epsilon_r<1$, and ∞ for $\epsilon_r=1$.

3. Behavior in the limit $\alpha \to 0$

For $\alpha \to 0$, we have $\epsilon_{r,crit} \to 0$. Therefore, for all $\epsilon_r > 0$ we are beyond the repair catastrophe, and since $(\alpha/\alpha+1)\mu = 0 < \ln k$, we are below the error catastrophe as well, so that $z_0 = 1$ with $z_{00} = 0$. This makes sense, since, for $\alpha = 0$, the probability of mutating off of the viability peak is $1 - e^{-(\alpha/\alpha+1)\mu} \to 0$. Thus, the entire population is viable at equilibrium. As for z_{00}, we note that $z_{00} = 0$ for $\epsilon_r > 0$, but for $\epsilon_r = 0$ we obtain the expression, $(e^0 - e^0)/(e^0 - e^0) \times z_0 = 0/0$. Physically, we must have $z_{00} = 1$ at $\epsilon_r = 0$. This ambiguity is therefore resolved by letting $\alpha \to 0$ for $\epsilon_r = 0$. That is, we evaluate z_{00} for $\alpha = 0, \epsilon_r = 0$ by setting $z_{00}|_{\alpha=0,\epsilon_r=0} = \lim_{\alpha \to 0} z_{00}|_{\alpha,\epsilon_r=0}$.

As expected, for $\epsilon_r > 0$ we have $z_0' = 0$, $\langle l' \rangle_{via} = \infty$, $\langle l \rangle_{rep} = 0$, $\langle l \rangle = 0$, and $\langle l' \rangle = \infty$. Again, for $\epsilon_r = 0$ we resolve any ambiguities by letting $\alpha \to 0$, giving, as expected, $z_0' = 1$, $\langle l' \rangle_{via} = \langle l \rangle_{rep} = \langle l \rangle = \langle l' \rangle = 0$.

V. CONCLUSIONS

This paper presented a two-gene, single fitness peak model to determine the equilibrium distribution of genotypes in a unicellular population capable of replication error repair. The work presented here was a continuation of Ref. [9], in which the equilibrium distribution of mutators was studied for mutation rates well below the error catastrophe. This paper obtained the equilibrium behavior of the two-gene model for arbitrary mutation rates, thereby incorporating both the error and repair catastrophes into a single, two-dimensional phase diagram. While our model is probably the simplest one could use for studying evolutionary dynamics in the presence of genetic repair, it does nevertheless make experimentally testable predictions. As mentioned in the Introduction, the error catastrophe has already been observed [30,31]. The repair catastrophe would be more difficult to observe experimentally, since it would be necessary to selectively interfere with the DNA mismatch repair system. If possible, however, it would be interesting to try to experimentally map out the phase diagram shown in Fig. 1 for an actual organism, such as *E. coli*.

In Ref. [9] it was noted that the equilibrium distribution of mutators did not depend on μ, but only on ϵ_r and α. This was interesting since the larger the value of μ the greater the difference in mutation rates off of the viability peak between repairers and mutators. One might also naively expect the repair catastrophe to disappear entirely as $\mu \to 0$, since the difference in viability between repairers and mutators disappears in the limit of no mutations. In Ref. [9] it was assumed that mutations were sufficiently slow so that only point mutations needed to be considered. In the complete model, when we allow for mutations between any two genomes, we do indeed obtain a μ dependence on the equilibrium distribution of mutators. Interestingly, however, the repair catastrophe still occurs at $\epsilon_{r,crit} = \alpha/\alpha + 1$, unchanged from the point-mutation result in Ref. [9].

ACKNOWLEDGMENT

This research was supported by the National Institutes of Health.

APPENDIX A: DERIVATION OF THE FINITE AND INFINITE SEQUENCE LENGTH FORMS OF THE DYNAMICAL EQUATIONS

1. The finite sequence length equations

To obtain the finite sequence length equations given by Eq. (5), we first obtain the quasispecies equations in terms of the $x_{ll'}$. To do this, we need to sum the mutational contributions of all σ to the time evolution of $x_{ll'}$. Let $\sigma_{ll'} \in C_H(l,l')$. Any σ may be obtained from $\sigma_{ll'}$ by changing the appropriate bases. Let us write $\sigma_{ll'} = \sigma_{via,l}\sigma_{rep,l'}$ and $\sigma = \sigma_{via}\sigma_{rep}$. By definition of the Hamming class, $\sigma_{via,l}$ differs from $\sigma_{via,0}$ in exactly l places. Therefore, $\sigma_{via,l}$ is identical to $\sigma_{via,0}$ in $L_{via} - l$ places. Of these $L_{via} - l$ bases, let l_1 denote the number of bases which are changed in σ. Of the l bases in $\sigma_{via,l}$ which are distinct from the corresponding bases in $\sigma_{via,0}$, let l_2 denote the number which are changed back to the corresponding bases in $\sigma_{via,0}$ when creating σ, and let l_3 denote the number which are changed to bases which are still distinct from the corresponding ones in $\sigma_{via,0}$. The base changes determined by l_1, l_2, and l_3 yields a σ_{via} which is a Hamming distance of $l_1 + l - l_2$ from $\sigma_{via,0}$, and a Hamming distance of $l_1 + l_2 + l_3$ from $\sigma_{via,l}$.

For the repair gene, we may define l_1', l_2', and l_3' similarly. Thus, given some $\sigma_{ll'} \in C_H(l,l')$, the vector $(l_1, l_2, l_3, l_1', l_2', l_3')$ defines a set of base changes to a $\sigma_{l_1+l-l_2, l_1'+l'-l_2'} \in C_H(l_1 + l - l_2, l_1' + l' - l_2')$, such that $D_H(\sigma_{ll'}, \sigma_{l_1+l-l_2, l_1'+l'-l_2'}) = l_1 + l_2 + l_3 + l_1' + l_2' + l_3'$. We then obtain that

$$\kappa_m(\sigma_{l_1+l-l_2, l_1'+l'-l_2'}, \sigma_{ll'}) = \kappa_{l_1+l-l_2} \left(\frac{\epsilon_{l_1'+l'-l_2'}}{S-1} \right)^{l_1+l_2+l_3+l_1'+l_2'+l_3'} (1 - \epsilon_{l_1'+l'-l_2'})^{L_{via}+L_{rep}-l_1-l_2-l_3-l_1'-l_2'-l_3'}. \quad \text{(A1)}$$

The total mutational flow rate into a given $\sigma_{ll'}$ may be obtained by summing over the mutational flow rates from all possible $(l_1, l_2, l_3, l_1', l_2', l_3')$. To put together a final expression, we still need to account for degeneracy, since, in general, for a given $\sigma_{ll'}$ and vector $(l_1, l_2, l_3, l_1', l_2', l_3')$, there are multiple ways for generating a new gene sequence. For a given l_1, we need to choose l_1 elements out of $L_{via} - l$. Since each selected base can be changed to $S - 1$ other bases,

the total number of possibilities for l_1 is $\binom{L_{via}-l}{l_1}(S-1)^{l_1}$. For a given l_2, we need to choose l_2 elements out of l. Since each selected base is restored to the corresponding base in $\sigma_{via,0}$, the total number of possibilities for l_2 is $\binom{l}{l_2}$. Finally, for a given l_3, we need to choose l_3 elements out of the remaining $l-l_2$. Since each selected base is changed, but is not changed back to the corresponding element in $\sigma_{via,0}$, there are $S-2$ possibilities per changed base, hence the total number of possibilities for l_3 is $\binom{l-l_2}{l_3}(S-2)^{l_3}$. Performing a similar analysis for the repair gene, and putting everything together, we obtain a total sequence degeneracy of

$$\binom{L_{via}-l}{l_1}\binom{l}{l_2}\binom{l-l_2}{l_3}\binom{L_{rep}-l'}{l_1'}\binom{l'}{l_2'}\binom{l'-l_2'}{l_3'}(S-1)^{l_1+l_1'}(S-2)^{l_3+l_3'}.$$

Putting everything together, we obtain

$$\frac{dx_{ll'}}{dt}=\sum_{l_1=0}^{L_{via}-l}\sum_{l_2=0}^{l}\sum_{l_3=0}^{l-l_2}\sum_{l_1'=0}^{L_{rep}-l'}\sum_{l_2'=0}^{l'}\sum_{l_3'=0}^{l'-l_2'}\binom{L_{via}-l}{l_1}\binom{l}{l_2}\binom{l-l_2}{l_3}\binom{L_{rep}-l'}{l_1'}\binom{l'}{l_2'}\binom{l'-l_2'}{l_3'}(S-1)^{l_1+l_1'}(S-2)^{l_3+l_3'}\kappa_{l_1+l-l_2}$$
$$\times\left(\frac{\epsilon_{l_1'+l'-l_2'}}{S-1}\right)^{l_1+l_2+l_3+l_1'+l_2'+l_3'}(1-\epsilon_{l_1'+l'-l_2'})^{L_{via}+L_{rep}-l_1-l_2-l_3-l_1'-l_2'-l_3'}x_{l_1+l-l_2,l_1'+l'-l_2'}-\overline{\kappa}(t)x_{ll'}. \tag{A2}$$

We may sum over l_3 and l_3' to obtain,

$$\frac{dx_{ll'}}{dt}=\sum_{l_1=0}^{L_{via}-l}\sum_{l_2=0}^{l}\sum_{l_1'=0}^{L_{rep}-l'}\sum_{l_2'=0}^{l'}\binom{L_{via}-l}{l_1}\binom{l}{l_2}\binom{L_{rep}-l'}{l_1'}\binom{l'}{l_2'}\kappa_{l_1+l-l_2}\epsilon_{l_1'+l'-l_2'}^{l_1+l_1'}(1$$
$$-\epsilon_{l_1'+l'-l_2'})^{L_{via}+L_{rep}-l-l'-l_1-l_1'}\left(\frac{\epsilon_{l_1'+l'-l_2'}}{S-1}\right)^{l_2+l_2'}\left(1-\frac{\epsilon_{l_1'+l'-l_2'}}{S-1}\right)^{l+l'-l_2-l_2'}x_{l_1+l-l_2,l_1'+l'-l_2'}-\overline{\kappa}(t)x_{ll'}. \tag{A3}$$

We may then use $z_{ll'}=C_{ll'}x_{ll'}$ to obtain Eq. (5) after some manipulation.

2. The infinite sequence length equations

To establish the infinite sequence length form of Eq. (5) in Sec. III, we need to first establish some basic inequalities to facilitate the computation of upper bounds. We begin with the following inequality, for $l_1>0$:

$$\binom{l_1+l-l_2}{l_1}\left(\frac{\epsilon_{l_1'+l'-l_2'}}{S-1}\right)^{l_1}\left(1-\frac{\epsilon_{l_1'+l'-l_2'}}{S-1}\right)^{l-l_2}\leqslant\left(\frac{\epsilon}{S-1}\right)^{l_1}\prod_{k=1}^{l_1}\frac{k+l-l_2}{k}=\left(\frac{\epsilon}{S-1}\right)^{l_1}\prod_{k=1}^{l_1}\left(1+\frac{l-l_2}{k}\right)\leqslant\left(\frac{l+1}{S-1}\epsilon\right)^{l_1}. \tag{A4}$$

Note that this inequality also holds for $l_1=0$. A similar inequality holds for the primed indices. Our next inequality is simply

$$\binom{L_{via}-l-l_1+l_2}{l_2}\epsilon_{l_1'+l'-l_2'}^{l_2}(1-\epsilon_{l_1'+l'-l_2'})^{L_{via}-l-l_1}\leqslant 1, \tag{A5}$$

and similarly for the primed indices. Finally, we may note that $z_{ll'}\leqslant 1$ for all l,l'. Now, to simplify the calculation, denote the summand in Eq. (5) of Sec. III by $S_{ll_1l_2l'l_1'l_2'}$. Then putting together our inequalities, we obtain

$$\sum_{l_2=0}^{l}\sum_{l_2'=0}^{l'}S_{l0l_2l'0l_2'}\leqslant\sum_{l_1=0}^{L_{via}-l}\sum_{l_2=0}^{l}\sum_{l_1'=0}^{L_{rep}-l'}\sum_{l_2'=0}^{l'}S_{ll_1l_2l'l_1'l_2'}\leqslant\sum_{l_2=0}^{l}\sum_{l_2'=0}^{l'}S_{l0l_2l'0l_2'}+\sum_{l_1=1}^{L_{via}-l}\sum_{l_2=0}^{l}\sum_{l_2'=0}^{l'}k\left(\frac{l+1}{S-1}\epsilon\right)^{l_1}$$
$$+\sum_{l_1'=1}^{L_{rep}-l'}\sum_{l_2'=0}^{l'}\sum_{l_2=0}^{l}k\left(\frac{l'+1}{S-1}\epsilon\right)^{l_1'}+\sum_{l_1=1}^{L_{via}-l}\sum_{l_2=0}^{l}\sum_{l_1'=1}^{L_{rep}-l'}\sum_{l_2'=0}^{l'}k\left(\frac{l+1}{S-1}\epsilon\right)^{l_1}\left(\frac{l'+1}{S-1}\epsilon\right)^{l_1'}$$
$$=\sum_{l_2=0}^{l}\sum_{l_2'=0}^{l'}S_{l0l_2l'0l_2'}+k(l+1)^2(l'+1)\frac{\epsilon}{S-1}\frac{1-\left(\frac{l+1}{S-1}\epsilon\right)^{L_{via}-l}}{1-\frac{l+1}{S-1}\epsilon}+k(l+1)(l'+1)^2\frac{\epsilon}{S-1}$$

$$\frac{1-\left(\frac{l'+1}{S-1}\epsilon\right)^{L_{rep}-l'}}{1-\frac{l'+1}{S-1}\epsilon}+k(l+1)^2(l'+1)^2\left(\frac{\epsilon}{S-1}\right)^2\frac{1-\left(\frac{l+1}{S-1}\epsilon\right)^{L_{via}-l}}{1-\frac{l+1}{S-1}\epsilon}\frac{1-\left(\frac{l'+1}{S-1}\epsilon\right)^{L_{rep}-l'}}{1-\frac{l'+1}{S-1}\epsilon}. \tag{A6}$$

Now, following the argument from Sec. III, we have that, as $L_{via},L_{rep}\to\infty$ at fixed α,μ,ϵ_r, we get that

$$\sum_{l_2=0}^{l}\sum_{l_2'=0}^{l'}S_{l0l_2l'0l_2'}\to\sum_{l_2=0}^{l}\sum_{l_2'=0}^{l'}\frac{\kappa_{l-l_2}}{l_2!l_2'!}\alpha^{l_2}\left(\frac{\mu_{l'-l_2'}}{\alpha+1}\right)^{l_2+l_2'}e^{-\mu_{l'-l_2'}}z_{l-l_2,l'-l_2'}, \tag{A7}$$

hence, since $\epsilon\to 0$ at fixed μ when $L_{via},L_{rep}\to\infty$, we see from the inequalities given in Eq. (A6) that

$$\sum_{l_1=0}^{L_{via}-l}\sum_{l_2=0}^{l}\sum_{l_1'=0}^{L_{rep}-l'}\sum_{l_2'=0}^{l'}S_{ll_1l_2l'l_1'l_2'}\to\sum_{l_2=0}^{l}\sum_{l_2'=0}^{l'}\frac{\kappa_{l-l_2}}{l_2!l_2'!}\alpha^{l_2}\left(\frac{\mu_{l'-l_2'}}{\alpha+1}\right)^{l_2+l_2'}e^{-\mu_{l'-l_2'}}z_{l-l_2,l'-l_2'}. \tag{A8}$$

The convergence is not uniform, however, since our upper bound depends on l,l'.

This establishes the infinite sequence length form of our dynamical equations, as given in Eq. (8) of Sec. III. We may verify that total probability is conserved in our limiting process. Defining $z=\Sigma_{l,l'}z_{ll'}$, we obtain

$$\begin{aligned}\frac{dz}{dt}&=\sum_{l=0}^{\infty}\sum_{l'=0}^{\infty}\sum_{l_1=0}^{l}\sum_{l_1'=0}^{l'}\frac{\kappa_{l-l_1}}{l_1!l_1'!}\alpha^{l_1}\left(\frac{\mu_{l'-l_1'}}{\alpha+1}\right)^{l_1+l_1'}e^{-\mu_{l'-l_1'}}z_{l-l_1,l'-l_1'}-\overline{\kappa}(t)z\\&=\sum_{l_1=0}^{\infty}\sum_{l_1'=0}^{\infty}\sum_{l=l_1}^{\infty}\sum_{l'=l_1'}^{\infty}\frac{\kappa_{l-l_1}}{l_1!l_1'!}\alpha^{l_1}\left(\frac{\mu_{l'-l_1'}}{\alpha+1}\right)^{l_1+l_1'}e^{-\mu_{l'-l_1'}}z_{l-l_1,l'-l_1'}-\overline{\kappa}(t)z\\&=\sum_{k_1=0}^{\infty}\sum_{k_1'=0}^{\infty}\kappa_{k_1}e^{-\mu_{k_1'}}z_{k_1,k_1'}\sum_{l_1=0}^{\infty}\sum_{l_1'=0}^{\infty}\frac{1}{l_1!l_1'!}\alpha^{l_1}\left(\frac{\mu_{k_1'}}{\alpha+1}\right)^{l_1+l_1'}-\overline{\kappa}(t)z\\&=\sum_{k_1=0}^{\infty}\sum_{k_1'=0}^{\infty}\kappa_{k_1}z_{k_1,k_1'}-\overline{\kappa}(t)z\\&=\overline{\kappa}(t)z-\overline{\kappa}(t)z=0.\end{aligned} \tag{A9}$$

Thus, since z starts off at 1, it remains 1 at all times, hence total probability is conserved in the infinite sequence limit.

APPENDIX B: A RECURSIVE FORMULA FOR THE POPULATION DISTRIBUTION

Given z_0,z_{00}, the equilibrium equations may be solved recursively to obtain any $z_{ll'}$ for a given (μ,ϵ_r) pair. For $l'>0$, we have

$$\frac{dz_{0l'}}{dt}=\frac{k}{l'!}\left(\frac{\epsilon_r\mu}{\alpha+1}\right)^{l'}e^{-\epsilon_r\mu}z_{00}+ke^{-\mu}\sum_{l_1'=1}^{l'-1}\frac{1}{l_1'!}\left(\frac{\mu}{\alpha+1}\right)^{l_1'}z_{0,l'-l_1'}+ke^{-\mu}z_{0l'}-[(k-1)z_0+1]z_{0l'}, \tag{B1}$$

so at equilibrium we have

$$z_{0l'}=\frac{1}{(k-1)z_0+1-ke^{-\mu}}\left[\frac{k}{l'!}\left(\frac{\epsilon_r\mu}{\alpha+1}\right)^{l'}e^{-\epsilon_r\mu}z_{00}+ke^{-\mu}\sum_{l_1'=1}^{l'-1}\frac{1}{l_1'!}\left(\frac{\mu}{\alpha+1}\right)^{l_1'}z_{0,l'-l_1'}\right]. \tag{B2}$$

We next turn our attention to z_{l0} for $l>0$. We have

$$\frac{dz_{l0}}{dt}=\frac{k}{l!}\left(\frac{\alpha\epsilon_r\mu}{\alpha+1}\right)^{l}e^{-\epsilon_r\mu}z_{00}+e^{-\epsilon_r\mu}\sum_{l_1=1}^{l-1}\frac{1}{l_1!}\left(\frac{\alpha\epsilon_r\mu}{\alpha+1}\right)^{l_1}z_{l-l_1,0}+e^{-\epsilon_r\mu}z_{l0}-[(k-1)z_0+1]z_{l0}, \tag{B3}$$

so at equilibrium we have

$$z_{l0}=\frac{1}{(k-1)z_0+1-e^{-\epsilon_r\mu}}\left[\frac{k}{l!}\left(\frac{\alpha\epsilon_r\mu}{\alpha+1}\right)^l e^{-\epsilon_r\mu}z_{00}+e^{-\epsilon_r\mu}\sum_{l_1=1}^{l-1}\frac{1}{l_1!}\left(\frac{\alpha\epsilon_r\mu}{\alpha+1}\right)^{l_1}z_{l-l_1,0}\right]. \tag{B4}$$

Finally, we compute the equilibrium value of $z_{ll'}$ recursively for $l,l'>0$. The result is

$$\begin{aligned}z_{ll'}=\frac{1}{(k-1)z_0+1-e^{-\mu}}\Bigg\{&\frac{k}{l!l'!}\alpha^l\left(\frac{\epsilon_r\mu}{\alpha+1}\right)^{l+l'}e^{-\epsilon_r\mu}z_{00}+\frac{k}{l!}\left(\frac{\alpha\mu}{\alpha+1}\right)^l e^{-\mu}\sum_{l_1'=0}^{l'-1}\frac{1}{l_1'!}\left(\frac{\mu}{\alpha+1}\right)^{l_1'}z_{0,l'-l_1'}\\&+\frac{1}{l'!}\left(\frac{\epsilon_r\mu}{\alpha+1}\right)^{l'}e^{-\epsilon_r\mu}\sum_{l_1=0}^{l-1}\frac{1}{l_1!}\left(\frac{\alpha\epsilon_r\mu}{\alpha+1}\right)^{l_1}z_{l-l_1,0}+e^{-\mu}\Bigg[\sum_{l_1'=1}^{l'-1}\frac{1}{l_1'!}\left(\frac{\mu}{\alpha+1}\right)^{l_1'}z_{l,l'-l_1'}+\sum_{l_1=1}^{l-1}\frac{1}{l_1!}\left(\frac{\alpha\mu}{\alpha+1}\right)^{l_1}z_{l-l_1,l'}\\&+\sum_{l_1=1}^{l-1}\sum_{l_1'=1}^{l'-1}\frac{1}{l_1!l_1'!}\alpha^{l_1}\left(\frac{\mu}{\alpha+1}\right)^{l_1'}z_{l-l_1,l'-l_1'}\Bigg]\Bigg\}.\end{aligned} \tag{B5}$$

APPENDIX C: NUMERICAL SOLUTION OF THE MODEL FOR FINITE GENOMES

Equation (5) in Sec. III gives the expression for the Hamming class symmetrized dynamics equations of our model. We can put this equation into matrix form

$$\frac{d\vec{z}}{dt}=B\vec{z}-(\vec{\kappa}\cdot\vec{z})\vec{z}, \tag{C1}$$

where $\vec{z}=(z_{ll'})$ is the vector of population fractions in the various Hamming classes, B is the matrix of first-order mutation rate constants between the various Hamming classes, and $\vec{\kappa}$ is the vector of first-order growth rate constants for the various Hamming classes, so that $\vec{\kappa}\cdot\vec{z}=\Sigma_{l=0}^{L_{via}}\Sigma_{l'=0}^{L_{rep}}\kappa_{ll'}z_{ll'}$, where $\kappa_{ll'}$ is simply κ_l in our model.

The equilibrium distribution may then be solved using fixed-point iteration, via the equation

$$\vec{z}_{n+1}=\frac{1}{\vec{\kappa}\cdot\vec{z}_n}B\vec{z}_n. \tag{C2}$$

In principle, the iterations are terminated when the z_n stop changing. We introduce a fractional cutoff parameter δ, and stop iterating when $(z_{n+N_\epsilon,ll'}-z_{n,ll'})/z_{n,ll'}<\delta$. N_ϵ is chosen to be sufficiently large so that of the order of one mutation is allowed to occur after N_ϵ iterations, to ensure that equilibration is being accurately measured. For a large sequence length L, the probability of correct replication is $e^{-L\epsilon}$, so the probability of incorrect replication is $1-e^{-L\epsilon}$. Therefore, taking $N_\epsilon=1/(1-e^{-L\epsilon})$ ensures that of the order of one incorrect replication has occurred, so that if $(z_{n+N_\epsilon,ll'}-z_{n,ll'})/z_{n,ll'}<\delta$ for all l,l', then it is possible to assume that equilibration has been achieved.

Note that what this method does is account for the fact that equilibration takes longer for smaller values of ϵ, i.e., for slower mutation rates. Since $\lim_{\epsilon\to 0}N_\epsilon=\infty$, and $\lim_{\epsilon\to 1}N_\epsilon\approx 1$ for large L, we see that our choice of N_ϵ accounts for the slower equilibration rate by iterating more times before comparing the changes in the $z_{ll'}$. In our numerical simulations, we found that $\delta=10^{-4}$–10^{-3} was sufficient to achieve good convergence. For $\alpha=2$, $k=10$, it was found that for $L=30$ the equilibrium values of z_0 and z_{00} were almost identical to their $L=\infty$ values. For this reason, we did not give figures showing the results of numerical simulations in this paper.

[1] D. Voet and J. Voet, *Biochemistry*, 2nd ed. (Wiley, New York, 1995).
[2] P.D. Sniegowski, P.J. Gerrish, and R. Lenski, Nature (London) **387**, 703 (1997).
[3] J. LeClerc, B. Li, W. Payne, and T. Cebula, Science **274**, 1208 (1996).
[4] F. Taddei, M. Radman, J. Maynard-Smith, B. Toupance, P. Gouyon, and B. Godelle, Nature (London) **387**, 700 (1997).
[5] O. Tenaillon, B. Toupance, H.L. Nagard, F. Taddei, and B. Godelle, Genetics **152**, 485 (1999).
[6] A. Giraud, I. Matic, O. Tenaillon, A. Clara, M. Radman, M. Fons, and F. Taddei, Science **291**, 2606 (2001).
[7] I. Matic, M. Radman, F. Taddei, B. Picard, C. Doit, E. Bingen, E. Denamur, and J. Elion, Science **277**, 1833 (1997).
[8] P.D. Sniegowski, P.J. Gerrish, T. Johnson, and A. Shaver, BioEssays **22**, 1057 (2000).
[9] E. Tannenbaum, E. Deeds, and E. Shakhnovich, Phys. Rev. Lett. **91**, 138105 (2003).
[10] A. Sasaki and M.A. Nowak, J. Theor. Biol. **224**, 241 (2003).
[11] M. Eigen, Naturwissenschaften **58**, 465 (1971).

[12] M. Eigen, J.S. McCaskill, and P. Schuster, Adv. Chem. Phys. **75**, 149 (1989).
[13] S. Galluccio, Phys. Rev. E **56**, 4526 (1997).
[14] M. Nilsson and N. Snoad, Phys. Rev. E **65**, 031901 (2002).
[15] R. Pastor-Satorras and R.V. Sole, Phys. Rev. E **64**, 051909 (2001).
[16] M. Nilsson and N. Snoad, Phys. Rev. Lett. **84**, 191 (2000).
[17] S. Altmeyer and J.S. McCaskill, Phys. Rev. Lett. **86**, 5819 (2001).
[18] P. Tarazona, Phys. Rev. A **45**, 6038 (1992).
[19] S. Franz and L. Peliti, J. Phys. A: Math. Gen. **30**, 4481 (1997).
[20] P.R.A. Campos and J.F. Fontanari, Phys. Rev. E **58**, 2664 (1998).
[21] D. Alves and J.F. Fontanari, Phys. Rev. E **57**, 7008 (1998).
[22] J. Swetina and P. Schuster, Biophys. Chem. **16**, 329 (1982).
[23] P. Schuster and J. Swetina, Bull. Math. Biol. **50**, 635 (1988).
[24] C.O. Wilke, J.L. Wang, and C. Ofria, Nature (London) **412**, 331 (2001).
[25] C.O. Wilke, Evolution (Lawrence, Kans.) **55**, 2412 (2001).
[26] E. van Nimwegen, J.P. Crutchfield, and M. Huynen, Proc. Natl. Acad. Sci. U.S.A. **96**, 9716 (1999).
[27] C.O. Wilke, Bull. Math. Biol. **63**, 715 (2001).
[28] M. Kolar and F. Slanina, Eur. Phys. J. B **31**, 379 (2003).
[29] C.O. Wilke, C. Ronnewinkel, and T. Martinetz, Phys. Rep. **349**, 395 (2001).
[30] S. Crotty, C. Cameron, and R. Andino, Proc. Natl. Acad. Sci. U.S.A. **98**, 6895 (2001).
[31] K. Negishi, D. Loakes, and R. Schaaper, Genetics **161**, 1363 (2002).