Analytical study of the effect of recombination on evolution via DNA shuffling

Weiqun Peng, Herbert Levine, and Terence Hwa
Center for Theoretical Biological Physics, University of California San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA

David A. Kessler
Department of Physics, Bar-Ilan University, Ramat-Gan, Israel
(Received 19 July 2003; published 24 May 2004)

DNA shuffling is an evolutionary protocol wherein cycles of selection, recombination, mutation, and amplification are employed to evolve proteins and DNA sequences. Experiments have shown its superiority to traditional protocols which do not employ recombination. Motivated by DNA shuffling, we investigate a multilocus evolutionary model that incorporates selection, recombination, and point mutations. Due to simplicity of the model, for the case of an infinite population we can obtain a full analytical treatment of both its dynamical and equilibrium properties, and study the benefit of recombination explicitly and quantitatively. We also briefly discuss finite-population size corrections.

DOI: 10.1103/PhysRevE.69.051911 PACS number(s): 87.10.+e, 87.23.Cc, 87.23.Kg

I. INTRODUCTION

Recombination, i.e., the exchange of genetic information, is a widespread phenomenon in both prokaryotes and eukaryotes. In prokaryotes, recombination happens occasionally, mediated by phages, direct cell-cell contact, or direct uptake of free DNA from the environment [1]. In eukaryotes, recombination between homologous sequences is a fundamental component underlying sexual reproduction [2]. An enormous body of research has been devoted to understanding the evolutionary benefit of recombination in various circumstances. This effort has led to some general understanding of the circumstances under which recombination helps facilitate evolution, however, many important questions still remain open [3], one important reason lying in the difficulty in theoretical treatments [4].

Inspired by recombination in natural evolution, and propelled by advances in biotechnology, recombination has been employed in *in vitro* molecular evolution experiments to develop proteins and DNA sequences [5,6]. This family of evolutionary protocols, called DNA shuffling (or molecular breeding) [7,8], has been shown experimentally to produce, in terms of the rate of evolutionary progression and final product quality, far superior results as compared to conventional directed evolution methods using only mutagenesis [5,6]. In addition to its widespread practical applications, DNA shuffling has been used to mimic natural evolutionary processes and predict possible evolutionary pathways [9,10].

In spite of its enormous significance, a theoretical understanding of DNA shuffling has been lacking. In this paper, we investigate DNA shuffling from the perspective of evolutionary modeling. Specifically, we aim to find out, quantitatively and analytically, the benefit that the extra step of recombination provides in the evolutionary process, as well as such aspects as the role of mutation and finite-population effects.

Compared to other evolutionary processes, DNA shuffling has two unique features that render it attractive to theoretical study. First, the methods of biotechnology enable a unique recombination scheme that goes well beyond the classical one, i.e., two parents with at most a few crossovers. The multiparent multicrossover nature of DNA shuffling, as we shall see, makes it more effective, and also, incidentally, allows for an exact analytical treatment. Second, in the setting of DNA shuffling the relationship between genotype, phenotype, and fitness is relatively clear and well defined. In addition, such parameters as selection strength, mutation rate, and the amount of recombination are all experimentally controllable. These characteristics should allow theoretical results to be tested directly against experiments.

Specifically, we propose a simple model that incorporates three basic ingredients: selection, recombination, and point mutations. In order to facilitate comparison with various existing evolutionary models, we present our ideas in the language of population genetics. In this language, we study a haploid (i.e., each individual carries a single copy of every gene) multilocus model with multiparent free recombination, and subject to dynamical truncation selection, wherein the fitness of a genotype depends on the population state. We obtain analytical results for both the dynamics and the equilibrium properties, and gain insights into not only why, but also *how* exactly recombination works. This is one of the rare cases in population genetics where exact results can be obtained for a nontrivial multilocus model [11].

This paper is organized as follows. In Sec. II, we present a model of the DNA shuffling process. We first consider the relevant experimental aspects of DNA shuffling, then propose a model that includes all the key ingredients yet is simple enough to allow analytical analysis, at least for the infinite population limit. In Sec. III, we focus on the model's dynamical behavior, studying the role of mutation as well as effects of the amount of diversity in the initial library. In Sec. IV, we move on to the consideration of equilibrium properties, with a discussion of the genetic load under recombination and a comparison of the equilibrium position with and without recombination. In Sec. V, we discuss finite-population effects, focusing on the question of how large must the population size be so that it behaves like one of infinite population.

II. MODELING THE DNA SHUFFLING PROCESS

DNA shuffling involves a directed evolution process wherein a library of homologous DNA sequences is subject to rounds of competitive selection and *in vitro* recombination with multiple parents and multiple crossovers [7,8]. DNA shuffling is a discrete process with nonoverlapping generations, each round of which involves selection, recombination, and point mutation, and is finished by amplification via the polymerase chain reaction (PCR) of the population back to it original size. The typical selection scheme in DNA shuffling experiments is truncation selection, also known as breeding selection, where only a fixed portion of the population (e.g., the top 10 %) is chosen to be retained to participate in later rounds. In truncation selection, whether a particular member of the population is selected or not depends on whether the desired trait exceeds a certain threshold set by the population as a whole and by the selection strength (i.e., the fraction of population selected). As opposed to other evolutionary scenarios, there is no advantage to being better than the cutoff threshold—there is no pressure to excel. Mathematically, this has a dramatic effect on the dynamics, in that the effect of population size is much weaker, as the rare "superstars" found only in a large population do not skew the results.

The recombination step typically involves random fragmentation of homologous DNA sequences by DNase digestion and repeated cycles of reassembly via a self-primed polymerase chain reaction [7,8]. Recombination produces chimeras with a controllable average fragment size (of order 10 base pairs or above). Point mutation is incorporated via the recombination and amplification steps where PCR is utilized and is naturally error prone. To date, the rate of point mutation has been kept very low and only single nucleotide substitution is assumed to be involved. Thus, sequence diversity has come mostly from the diversity present in the initial library instead of being generated by point mutation. This lessens the deleterious effects associated with high rates of mutagenesis but ultimately limits the usefulness of the method. A proper balance of recombination and mutation, as we shall see, would lead to more optimal results.

To make our discussion concrete, we envision the competitive evolution of a library of DNA sequences, selected via binding affinity to certain proteins; an example of such a system is the DNA-histone interaction [12]. As discussed elsewhere [13–15], selection achieved via thermodynamic binding can be mimicked to a high level of accuracy by the simple truncation selection approach. In order to facilitate a theoretical treatment, we construct simplified models of the recombination and selection steps. Let us describe the selection step first. For selection, we need a model that connects genotype and phenotype, which in this case are the DNA sequence and the binding energy of the DNA to the protein, respectively, as well as a model that specifies selection on the phenotype. We adopt the simplest relationship between genotype and phenotype. We assume that each nucleotide contributes to the binding energy independently and additively. Each nucleotide can be one of the $\mathcal{A}(=4)$ nucleotides (representing *A*, *C*, *G*, and *T*), among which one is favorable and the rest are equally deleterious. For simplicity, we denote the contribution of a specific site to the binding energy 1(0) if therein sits a favorable (deleterious) nucleotide. We will refer to a favorable (deleterious) nucleotide as a match (mismatch) with respect to the optimal sequence. This formulation is in fact a two-state model for protein-DNA binding [13,15]. The binding energy of the sequence is simply the number of sites with favorable nucleotides.

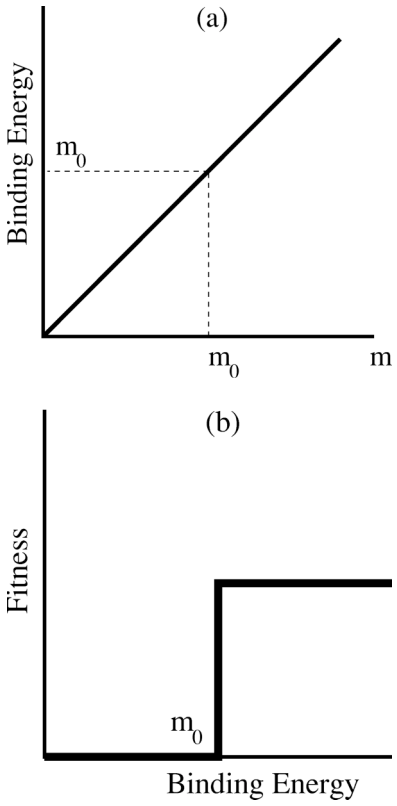

FIG. 1. (a) Phenotypic landscape: Binding energy is exactly the number of favorable active sites (denoted by m). Dash line indicates the corresponding selection threshold m_0 at a particular stage of evolution. (b) The fitness has a dynamic truncation landscape. Sequences with binding energy below the threshold m_0 are discarded, whereas those with binding energy above the threshold are retained to reproduce with equal rate.

Figure 1(a) shows schematically the shape of our phenotypic landscape. Of course, our model reflects just the simplest possibility. In reality, the phenotypic landscape could be much more complicated and is rarely known; in fact, a significant advantage of directed evolution over rational design is that this kind of knowledge is not necessary [16]. Our strategy here is to study the simplest nontrivial model available and obtain a thorough understanding, with the hope of proceeding to more complicated situations to see which results obtained in the simple model are general and which are model specific. Having specified the phenotypic landscape, we next describe the selection protocol. Selection acts on the binding energy of the sequences. As noted above, in (molecular) breeding, truncation selection is used. We define the selection strength via the fraction ϕ of selected members of the total population. Suppose we have a distribution of population P_m in terms of the binding energy $m(m=0,1,\ldots,L)$,

where L is the total number of active sites. The threshold m_0 is self-consistently determined by

$$\phi = \alpha P_{m_0} + \sum_{m=m_0+1}^{L} P_m, \tag{1}$$

where α in the first term takes into account of the partial selection on the threshold state m_0. ϕ varies from $\phi \lesssim 1$ (relatively weak selection) to small ϕ (relatively strong selection). For those sequences whose number of matches are above (below) m_0, they are selected (discarded), and their fitness is 1(0) [see Fig. 1(b)]. In the truncation selection scheme every member's fitness is collectively and dynamically determined by both the phenotypic distribution of the population P_m and the selection strength ϕ. In contrast, most fitness landscapes studied prescribe for each genotype a preset fitness value and its effective fitness value is simply its own fitness over the mean fitness [17]. Truncation selection generates correlations (i.e., linkage) between loci, hence it is epistatic [18,19].

We note in passing that the term truncation selection has also been used to mean a *fixed* steplike landscape. In population genetics literatures, a fixed steplike landscape is also called *hard* truncation selection, and the selection scheme we employed is sometimes termed *soft* truncation selection. Due to the dynamical nature of the selection scheme, using fitness as a yardstick for the evolution is not very useful; for example, the mean fitness of the population is always ϕ by definition. As a consequence, we will focus on the evolution of the phenotypic distribution, i.e., the distribution of binding energies, instead of the fitness distribution.

We now turn to the discussion of recombination. Recombination in general is a nonlinear nonlocal operation in sequence space, hence not easily amenable to theoretical treatment [4]. Our basic approach here will be to make a substantial simplification of the actual situation and assume perfect recombination between all nucleotides. Namely, in building each new sequence after selection, each nucleotide independently samples the nucleotides at the corresponding sites in the entire selected population. Formally, this amounts to assuming that the fragmentation and reassembly process is repeated often enough such that there is no linkage between any of the nucleotides. This is undoubtedly false in detail for the experiments done to date, but we shall see that this approximation exemplifies the benefit of recombination and is quite good for describing the result of a scenario involving the more feasible case of a finite number of crossovers. The advantage of this "maximal" recombination is that it allows for an analytic solution of the model, as will be presented in the remainder of this paper. Finally, we assume simple point mutation for each site. Namely, each nucleotide is independently subject to mutation rate μ_0 per generation. This process includes both beneficial and deleterious mutations.

It is worth pointing out that our approach to recombination may have more general applicability than merely as an approximation for the DNA binding problem. In many cases of DNA shuffling, the initial library of genes coding for an interesting protein is chosen from closely related species so as to ensure proved functionality and sufficient homology to allow recombination. Let us imagine that the majority of the nucleotides along those DNA sequences are already optimal, and (nonsynonymous) mutations of these nucleotides are lethal. Hence, we can consider these nucleotides to be fixed during the entire evolutionary process and ignore them except insofar as they provide enough homology so that overlapping single-stranded fragments from different DNA sequences can anneal with each other (and do not anneal with other fragments by accident) during the self-primed PCR reassembly step [7,8]. For the remaining *active* sites, we assume that they are (a) far enough from each other so that recombination happens freely between any two; (b) they are subject to point mutations with rate μ_0 per nucleotide per generation. Therefore, each active site, along with its fixed flanking homologous regions (whose size or exact delineation does not matter), constitutes a segment. To build a new sequence from recombination, a number of overlapping fragments (which when put together cover the entire sequence) are assembled, with each fragment obtained from randomly sampling the corresponding fragments (i.e., having the same active site) in the selected population. This idea is depicted in Fig. 2 and is an exact realization of our maximal recombination model. Of course, one has to be careful to intersperse an experimental selection step that will eliminate all lethal variants (due to mutation) before proceeding to an actual selection based on useful variation. To proceed, we would then have to explicitly take into account the transformation from gene sequence to amino acid, as the selection would be on the basis of some desired activity of the protein. We do not pursue this line of investigation any further in this work.

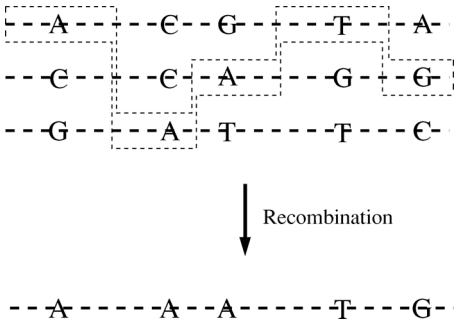

FIG. 2. Maximal recombination under a general interpretation. Each line represents a DNA sequence. Letters represent active sites with variable length fixed regions in between. In building a new sequence, each active site independently samples the corresponding sites in the entire selected population.

Finally, we can rephrase our model in terms of the standard language of population genetics. Each site can be referred to as a locus, which can have $\mathcal{A}$ alleles, one favorable (match) and the rest (mismatches) equally unfavorable. The set of all L loci forms a chromosome. The recombination scheme is such that the allele of each locus of every new chromosome is chosen by randomly sampling the alleles at the corresponding locus of all the selected chromosomes. The fitness value of each chromosome is 1(0), when the number of matches it has is above (below) the threshold m_0. For clarity, we list the correspondence in Table I.

TABLE I. The correspondence between two languages: directed molecular evolution and population genetics.

DNA-protein binding	Population genetics
DNA sequence	Chromosome
Nucleotide	Locus
Binding energy	Phenotypic value:matches
Selection via binding to protein	Dynamical truncation selection
Correlation of nucleotides	Linkage

Despite its enormous practical success, theoretical analysis of the evolutionary dynamics of DNA shuffling has thus far been lacking. Sun [20], Moore and Maranas [21] proposed predictive models for various experimental steps of DNA shuffling experiments, addressing issues of recombination efficiency and distribution of fragment size during the reassembly involved in a single round of evolution. We concern ourselves instead with the evolutionary consequences of multiple rounds. Several aspects of our evolutionary approach have been studied analytically in population genetics [17]. Response to truncation selection in one round, under linkage-free condition, has been characterized in the context of classical breeding [17,22]. The effects of various recombination schemes have been studied in the special case of evolution without selection [23]. Kondrashov [24] studied a model with a different type of dynamical fitness landscape, where the fitness value of a genotype depends only on the difference between its phenotypic value and the mean of the phenotypic distribution, in units of variance of the phenotypic distribution. He derived evolutionary recursion relations and obtained analytical expressions that characterize the equilibrium position, with the assumption of only deleterious mutations, conventional recombination, and a Gaussian distribution before selection. The assumption of unidirectional mutation is only true when the system is close to the optimal state; the Gaussian approximation, as we shall see, is in fact equivalent to using maximal recombination. In genetic algorithms and evolutionary strategies, a research area in computer science where the principles of evolution are employed to find optimal solutions to complex problems [25], a similar setting has been investigated by Mühlenbein and Schlierkamp-Voosen [26], mostly via computer simulation only.

III. MAXIMAL RECOMBINATION: DYNAMICS

To summarize the above discussion, in the language of population genetics which we will adopt from here on, we have a population of N chromosomes each with L loci, and we study the evolution with dynamical truncation selection, maximal recombination, and point mutation as specified above. We fix the order of operation to be selection, recombination, and mutation, and keep the population size N constant at each generation. This model is easily simulated on a computer. The evolutionary protocols are realized in simulations as follows: In producing the population in each generation, truncation selection is first used on the population of the previous generation to find the parental group. Then, each member chromosome of the population in the current generation is built by randomly sampling the parental group to find its parents, inheriting the parents' chromosomes according to the recombination scheme used, and mutating the resulting chromosome. Figure 3 shows simulation results for evolutionary trajectories of the average and variance of the match distribution under weak selection. For comparison, we show results from four different evolutionary protocols: no recombination, single random crossover, multiple crossovers and multiple parents (which is the situation closest to experiments), and maximal recombination. It is evident that recombination improves both the dynamics and equilibrium state as compared to the case with no recombination. For our phenotypic landscape and selection scheme, the more recombination the better. Furthermore, the maximal recombination scheme captures the essential effect of recombination and provides a reasonable approximation to the more readily achievable case of multiple crossovers from multiple parents. In the alternative case of strong selection (not shown), all evolutionary protocols propel the population to an equilibrium state very close to the optimal state; however the population reaches the equilibrium state much faster when recombination is applied.

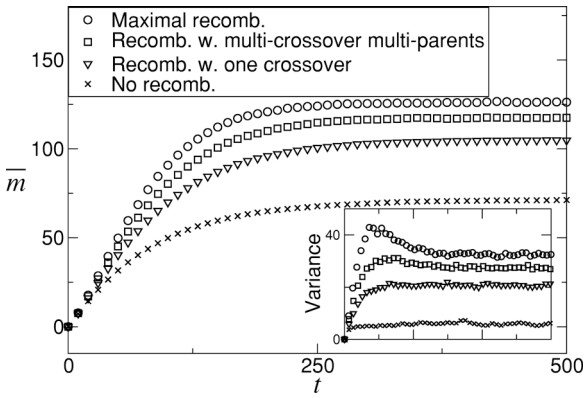

FIG. 3. Comparison of evolutionary trajectories for four different evolutionary protocols under weak selection. Shown are simulation results of the evolutionary trajectory of the average number of matches $\overline{m}$ for the population, with an initial clonal population (i.e., a population consisting of clones of a single chromosome) with no matches at any locus along the chromosome. X, no recombination; triangle, recombination with one random crossover; square, recombination with multiple crossover and multiple parents, using a per bond crossover probability of 0.025. Circle: maximal recombination. Shown in the inset are the corresponding variances of the distributions vs time, ordered in the same way as the matches vs generations curves. Here sequence length $L=170$, selection strength $\phi=0.9$, mutation rate $\mu_0=0.01$, and population size $N=10^4$.

In our analytical work we will assume, unless specifically noted otherwise, the population size N to be very large so that random genetic drift is negligible; we will briefly discuss finite-size effects at the end. We focus on the evolution of such macroscopic characteristics of the population as mean and variance, as opposed to "microscopic" properties such as the fate of individual mutations. As already noted, the fitness

distribution itself does not describe the evolution. We instead study the evolution of the phenotypic (match) distribution, focusing on its first and second moments. For simplicity we assume linkage equilibrium at the beginning of evolution. This is in fact not a stringent assumption, since linkage equilibrium is achieved anyway right after one round of recombination.

A. Dynamics without mutation

We start from the simplest case where the mutation rate μ_0 is set to be 0. We assume that each locus has the same binary distribution characterized by the probability of being favorable (i.e., a match) $p(0)(\neq 0)$. This is in fact the case most relevant to the DNA shuffling experiments to date, where μ_0 is kept extremely small and the diversity is almost entirely provided by the diversity which existed in the initial library [5]. Starting from the homogeneous initial condition, we expect that every locus follows the same evolution trajectory, as the evolution dynamics preserve permutation symmetry of different loci. With this in mind, we focus on the evolution of the probability $p(t)$ for one locus to be favorable at the end of generation t. The phenotypic probability distribution for the number of matches of a chromosome at the end of round t is then a binomial distribution characterized by mean $Lp(t)$. Assuming that L is large (For an exact mean-field treatment without the assumption of large L, see Appendix A), the binomial distribution is well approximated by a Gaussian distribution $P(m,t)$:

$$P(m,t) = \frac{1}{\sqrt{2\pi\sigma^2(t)}}\exp\left(-\frac{[m-\overline{m}(t)]^2}{2\sigma^2(t)}\right), \tag{2}$$

$$\overline{m}(t) = Lp(t), \tag{3}$$

$$\sigma^2(t) = Lp(t)[1-p(t)]. \tag{4}$$

Given such a distribution at the end of generation t, we now discuss step by step the effects on the distribution due to various operations in the evolutionary protocol. In generation $t+1$, selection cuts out the low m tail of the Gaussian distribution. It is straightforward to calculate that the mean match number of the selected population $\overline{m}_s$ is

$$\overline{m} \to \overline{m}_s = \overline{m} + \sigma G(\phi). \tag{5}$$

The new mean can thus be expressed as the old mean plus an improvement due to selection, which is simply the product of the old standard deviation and a factor $G(\phi)$ that solely encodes the strength of selection [$G(\phi)$ is called intensity of selection in population genetics literature] [27]. Here $G(\phi)$ is the mean for the normalized distribution resulting from a standard Gaussian distribution truncated by a $1-\phi$ fraction taken off the tail, namely,

$$G(\phi) = \frac{1}{\sqrt{2\pi}\phi}\exp\left(-\frac{1}{2}X(\phi)^2\right), \tag{6}$$

where $X(\phi)$ is the match threshold defined through

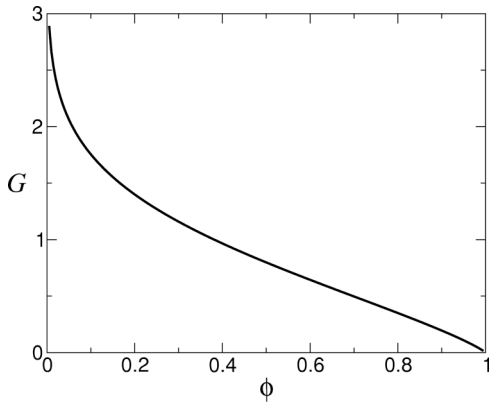

FIG. 4. The selection factor $G(\phi)$. $G(\phi)$ diverges at strong selection $\phi \to 0$ and goes to zero at weak selection $\phi \to 1$.

$$\phi = \int_{X(\phi)}^{\infty} \frac{dx}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right). \tag{7}$$

For completion, we show the behavior of $G(\phi)$ in Fig. 4. For $0.3 < \phi < 1$, $G(\phi)$ is approximately a linear function of ϕ (with slope roughly -1.5), and $G(\phi) \to (1-\phi)\sqrt{2\ln(1-\phi)^{-1}}$ as $\phi \to 1$. In the strong selection limit, i.e., $\phi \to 0$, $G(\phi) \to \sqrt{2\ln\phi^{-1}}$, which diverges.

After selection, the recombination step restores the independence of each locus. The population distribution of matches returns to a binomial (Gaussian) distribution characterized by mean $\overline{m}_s$:

$$\overline{m}_r = \overline{m}_s, \tag{8}$$

$$\sigma_r^2 = \overline{m}_r(1-\overline{m}_r/L). \tag{9}$$

Because of the independence of each locus, we can reexpress Eq. (8) in terms of individual match probability p_r and $p(t)$:

$$p_r = p(t) + \frac{\sqrt{p(t)[1-p(t)]}}{\sqrt{L}}G(\phi), \tag{10}$$

and the variance of the distribution is simply $Lp_r(1-p_r)$. Equation (10) has two features: First, the scaled combination of $G(\phi)/\sqrt{L}$ is the single control parameter. Second, the change of p is proportional to the square root of the variance.

In the case of weak selection, Eq. (10) can be approximated by its continuous-time version:

$$\frac{dp}{dt} = \frac{G(\phi)}{\sqrt{L}}\sqrt{p(1-p)}. \tag{11}$$

The evolutionary dynamics is governed by two fixed points: $p=0$ and $p=1$. $p=0$ is a trivial unstable fixed point. When $p(0)>0$, the population moves toward $p=1$. The entire solution is

$$p(t) = \begin{cases} 0, & p(0) = 0 \\ \frac{1}{2}\left[1 - \sin\left(\beta - \frac{G(\phi)}{\sqrt{L}}t\right)\right], & p(0) > 0, \end{cases} \quad (12)$$

where $\beta = \sin^{-1}[1-2p(0)]$. Equation (11) and its solution have also been derived in Mühlenbein and Schlierkamp-Voosen [26]. From this we see that the system actually reaches the optimal state in a finite time, rather than approaching exponentially. This is due to the square-root behavior of the velocity near $p=1$ noted above. T is given by

$$T = \left(\frac{\pi}{2} + \beta\right)\frac{\sqrt{L}}{G(\phi)}. \quad (13)$$

To appreciate this result and the benefit of recombination, it is helpful to compare it with that of pure enrichment (i.e., selection only), given the same initial condition. Starting from $p(0)$, the population distribution initially is a binomial distribution. Selection keeps narrowing down the distribution by chopping off its low tail round by round, and stops when the distribution contains only the perfect state with L matches. Note that since here we care about the extreme tail of the distribution, the Gaussian approximation is no longer adequate. Based on this scenario, to determine the evolution time T, we look at the fraction of population in the perfect state. In the beginning, the fraction is $p(0)^L$. At round t, the fraction becomes $p(0)^L/\phi^t$. Therefore, the evolution time T_{A} is determined by the condition that $p(0)^L/\phi^{T_{\mathrm{A}}} = 1$. Therefore

$$T_{\mathrm{A}} = L\frac{\ln p(0)}{\ln \phi}. \quad (14)$$

Equations (13) and (14) show that the evolution time scales differently with L in the two cases; as $\sqrt{L}$ in the case with recombination and as L in the case of pure enrichment. This means that the longer the chromosome, the greater the benefit of recombination. When selection is weak but not extremely close to 1 (weak so that continuous-time approximation is appropriate, but not too close to 1 so that around the threshold the tail that is cut off is still Gaussian-like), the chromosome is long and $p(0)$ is neither close to 0 nor close to 1 (so that Gaussian approximation is appropriate), the estimate of evolution time given in Eq. (13) is applicable. In this case,

$$\frac{T}{T_A} \approx \frac{\pi/2 + \sin^{-1}[1-2p(0)]}{-\ln p(0)}\frac{1}{\sqrt{L}}\frac{1}{\sqrt{-2\ln(1-\phi)}}, \quad (15)$$

clearly showing the superiority of the evolution protocol involving maximal recombination. In other cases, Eq. (13) is no longer appropriate. For example, Eq. (13) predicts a finite evolution time even in the limit of $p(0) \to 0$, an artifact of the approximations involved. In Appendix A, we discuss the benefit of recombination under this limit, making use of the formalism developed there, which is not limited by the approximations mentioned above.

We have shown that recombination provides a significant improvement when $L \gg 1$ and selection is weak. In general, at the mean-field level, recombination is at least not worse and in most cases it is beneficial. Though a rigorous proof is not yet available, the basic mechanism at work, as has been presented by population geneticists (see, e.g., Ref. [3] and references therein), is as follows: Selection, as it operates on the phenotype, generally introduces correlation between different loci on a chromosome. In the current case where the selection is via truncation, the correlation shows up as a narrowing of the population distribution in terms of matches. After selection, maximal recombination completely breaks up the correlation between loci, resulting in a broader distribution (see Fig. 3 inset). A broader distribution leads to better response to subsequent selection and hence faster evolution. In other words, without recombination, the unit of selection is the chromosome; with recombination, selection unit goes down to locus level, so that different loci evolve in a more parallel fashion.

B. Dynamics with mutation: homogeneous initial condition

So far, we have studied the dynamics of maximal recombination in the absence of mutation. Now we insert the point mutation process into the evolutionary dynamics. As a first step, we again assume a homogeneous initial condition such that each locus has the same probability of being favorable (i.e., a match) $p(0)(\neq 0)$. Mutation is the final step of the round. Since we only consider single base mutations, linkage is not introduced in the process. The mutation process we consider here includes both beneficial and deleterious mutations; by itself, it drives the chromosomes toward the maximum entropy point $L/\mathcal{A}$ (or $1/\mathcal{A}$ in terms of p), which is in general opposite to the direction of selection. A simple calculation yields

$$p_{\mathrm{m}} = p_{\mathrm{r}} + \frac{\mu_0}{\mathcal{A}-1}(1 - \mathcal{A}p_{\mathrm{r}}). \quad (16)$$

Combining this equation with Eq. (10), we obtain a recursion relation for the evolution of p:

$$p(t+1) = p(t) + \frac{\mu_0}{\mathcal{A}-1}[1 - \mathcal{A}p(t)] + \frac{G(\phi)}{\sqrt{L}}\sqrt{p(t)[1-p(t)]}, \quad (17)$$

where we have dropped a correction factor $[1+\mu_0\mathcal{A}/(\mathcal{A}-1)]$ in the last term, since $\mu_0 \ll 1$. It is clear that the second term on the right-hand side of the recursion relation is due to mutation, and the third term to selection and recombination. When the third term dominates the second term, the dynamics is essentially the same as that with no mutation as discussed above. When p exceeds the maximum entropy point $1/\mathcal{A}$, the mutational contribution becomes harmful to the evolution.

The recursion relation, Eq. (17), has an interesting feature. If we divide Eq. (17) by μ_0 on both sides, we have

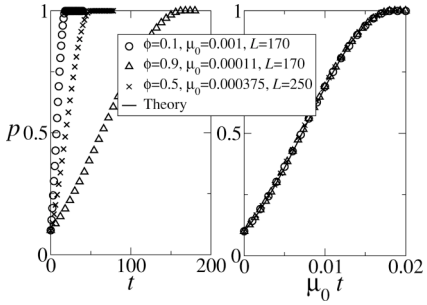

FIG. 5. Comparison of simulation results and theory. The evolutionary trajectories of three simulations with totally different parameters [but same scaling variable $G(\phi)/(\mu_0\sqrt{L})$] collapse into a single curve when time is properly rescaled, and this curve agrees well with the theoretical result obtained from Eq. (17). Here the three simulations all start from a homogeneous initial condition of $p(0)=0.1$ and population size $N=10^4$.

$$\frac{p(t+1)-p(t)}{\mu_0}=\frac{1}{\mathcal{A}-1}[1-\mathcal{A}p(t)]+\frac{G(\phi)}{\mu_0\sqrt{L}}\sqrt{p(t)[1-p(t)]}. \tag{18}$$

Equation (18) says that the scaled combination $G(\phi)/(\mu_0\sqrt{L})$ is the single control parameter. In other words, if different choices of parameters result in the same combination $G(\phi)/(\mu_0\sqrt{L})$, the corresponding dynamics are exactly the same as long as the time scale in each case is rescaled by its respective mutation rate μ_0. Note that in Eq. (17) or Eq. (18), $p(t)$ can go above 1 when selection is strong; this unrealistic result comes from the Gaussian approximation to the population distribution that becomes inaccurate when the population reaches the neighborhood of the optimal state. We will further address the error due to the Gaussian approximation in our discussion of the equilibrium state.

As a test of our theory, Fig. 5 shows that the theoretical predictions, including the scaling with $G(\phi)/(\mu_0\sqrt{L})$, agree extremely well with simulation data. This indicates that the finite-population effect in this landscape is insignificant, and the Gaussian approximation to the binomial distribution is appropriate for sequences of long length L.

When selection is weak so that the change in p in each round is small, the evolution equation (17) can again be accurately approximated by its continuous-time version,

$$\frac{dp}{dt'}=(1-\mathcal{A}p)+C\sqrt{p(1-p)}, \tag{19}$$

where $t'\equiv[\mu_0/(\mathcal{A}-1)]t$ and $C\equiv(\mathcal{A}-1)G(\phi)/(\mu_0\sqrt{L})$. The explicit solution of this equation can be found in Appendix B.

An explicit analytical comparison of the evolutionary performance between the evolutionary protocol with and without recombination is not available. Both recombination and mutation can serve to generate diversity, but the greater benefit of recombination is due to that facts that (a) recombination breaks up the correlation of loci along the chromosome introduced by selection much more effectively than mutation does; breaking of linkage helps broaden the distribution (as selection narrows the distribution) and in turn facilitates more efficient future selection, hence speeding up the evolution, and (b) recombination keeps the mean of the population unchanged, whereas mutation goes against selection (once the population goes beyond the maximum entropy point). Therefore, as has been recognized for a long time, recombination is able to generate variety without the excessive baggage of deleterious mutations.

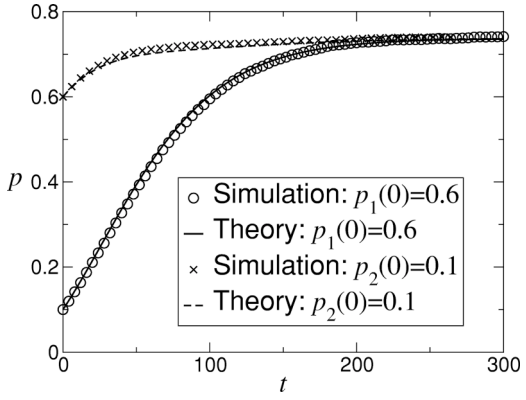

FIG. 6. Evolution of individual match probability p with inhomogeneous initial condition. The initial condition is chosen so that half of the loci have initial match probability of $p_1(0)=0.1$, while the remaining half has $p_2(0)=0.6$. The two simulation curves track the two different groups of loci. The solid curves are theoretical results from Eq. (20). Simulations use the following parameters: $L=170$, $\phi=0.9$, $\mu_0=0.01$, and $N=10^4$.

C. Dynamics with mutation: inhomogeneous initial condition

In the previous discussions, we assumed a simple homogeneous initial condition. An immediate question concerns what happens with a inhomogeneous initial condition, i.e., would different loci synchronize with each other after a short transient period or would they go their separate ways and only meet at the end of the process? To answer this question, we choose a linkage-free initial condition where half of the loci have probability $p_1(0)$ of being a match, and the other half have probability $p_2(0)$ of being a match. Assuming again that $L\gg 1$, a similar derivation to the one presented above produces

$$\begin{aligned}p_1(t+1)&=p_1(t)+\frac{\mu_0[1-\mathcal{A}p_1(t)]}{\mathcal{A}-1}\\&\quad+\frac{G(\phi)}{\sqrt{L/2}}\frac{p_1(t)[1-p_1(t)]}{\sqrt{p_1(t)[1-p_1(t)]+p_2(t)[1-p_2(t)]}},\\p_2(t+1)&=p_2(t)+\frac{\mu_0[1-\mathcal{A}p_2(t)]}{\mathcal{A}-1}+\frac{G(\phi)}{\sqrt{L/2}}\\&\quad\times\frac{p_2(t)[1-p_2(t)]}{\sqrt{p_1(t)[1-p_1(t)]+p_2(t)[1-p_2(t)]}}.\end{aligned} \tag{20}$$

Figure 6 compares these results with an evolutionary tra-

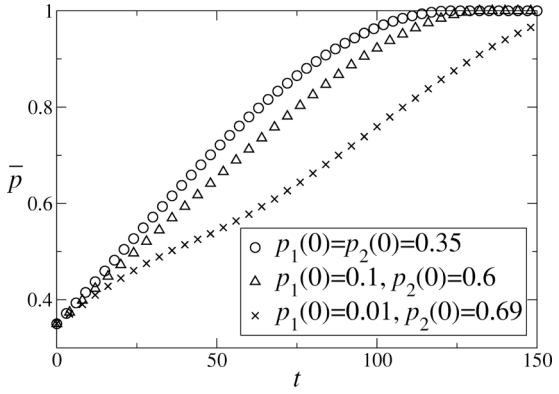

FIG. 7. Comparison of simulation results for the average individual match probability $\bar{p}$ for different initial conditions, in the absence of mutation. $\bar{p}$ is the individual match probability p averaged over loci. Here $L=170$, $\phi=0.9$, and $N=10^4$.

jectory with inhomogeneous initial conditions. It is clear that different loci go their own ways.

If mutation is negligible, one finds that the relative changes in the individual match probabilities are

$$\frac{p_1(t+1)-p_1(t)}{p_2(t+1)-p_2(t)}=\frac{p_1(t)[1-p_1(t)]}{p_2(t)[1-p_2(t)]}, \tag{21}$$

i.e., proportional to the ratio of the variance on each locus. In other words, the selection works on variance; different loci with different p experience different selection pressures depending on their contribution to variance. In the case of inhomogeneous initial conditions, this means that the evolutionary process will be slowed down by loci having very small initial $p(0)$; see Fig. 7 for explicit examples.

It is worth mentioning that although systems with different initial conditions follow different evolutionary paths, they end up with the same equilibrium state. Recombination is beneficial to the evolutionary progress even when the initial condition is inhomogeneous.

D. Dynamics with mutation: clonal initial condition

In the above theoretical analysis we have assumed an initial condition which already has all the favorable alleles available in the population, and mutation was relatively unimportant. In many situations, mutations are absolutely necessary for the system to find the optimal state. This scenario could happen for a system prepared with a clonelike initial condition with some loci lacking the beneficial alleles. In this section we focus on this case and study the effect of different mutation rates on the evolutionary dynamics. We use a clonal initial condition, where all members of the population are clones to each other, with no beneficial allele at any locus [i.e., $p(0)=0$]. For this simple case, the obviously best strategy is to subject the system to maximal mutation rate $\mu_0=(\mathcal{A}-1)/\mathcal{A}$ until the system reaches $p=1/\mathcal{A}$, and then to stop mutation altogether. However, this naive strategy does not apply to other clonal initial conditions which are inhomogeneous; here it is not *a priori* clear when one should turn

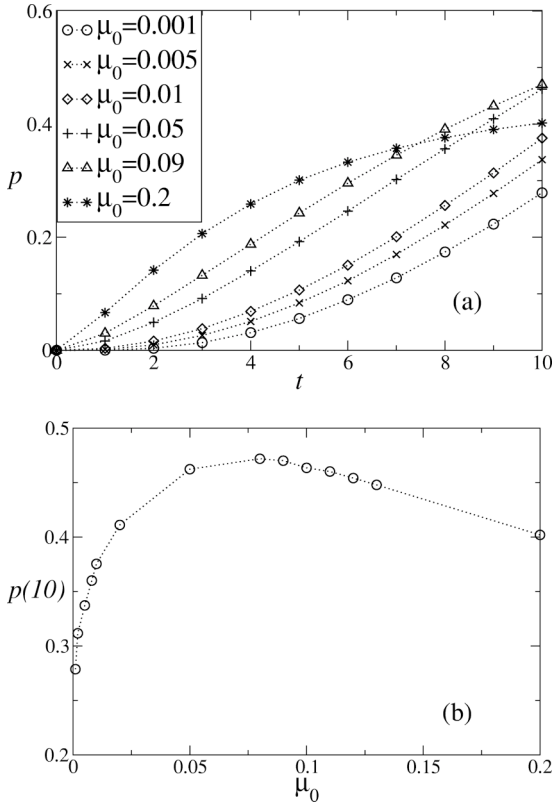

FIG. 8. Simulation results of different mutation rates given a clone initial condition of $p(0)=0$. (a) Evolutionary trajectories for different choices of mutation rate μ_0. (b) Comparison of individual match probability at the tenth generation $p(10)$ for different mutation rates. The "optimal" mutation rate is around $\mu_0=0.08$ under this condition. Here $L=170$, $\phi=0.1$, and $N=10^4$.

off mutation, since we have access only to measurements of the full phenotype.

Assuming that mutation rate is fixed, we can gauge the performance of different choices of mutation rate by measuring p after a given number of generations. Figure (8) shows a comparison of simulations starting at $p(0)=0$ with different mutation rates using this criterion. It is clear that there is an "optimal" mutation rate. Of course, the optimal mutation rate depends on when evolution is terminated. In the more common situation, a majority of the loci are already occupied by beneficial alleles, and for the rest, mutation is required to create the beneficial allele. In this case as well, we can see the existence of an optimal mutation rate, as in Fig. 9.

In conclusion, mutation is necessary when there is limited diversity in the initial library. Even in this case, mutation is only helpful in the short term dynamics, but harmful to the equilibrium state. In a realistic experiment, since breeding usually proceeds for a only few generations there would exist an optimal mutation rate.

IV. MAXIMAL RECOMBINATION: EQUILIBRIUM STATE

Having studied the dynamical behavior of the model in the previous sections, we now turn to the study of its equi-

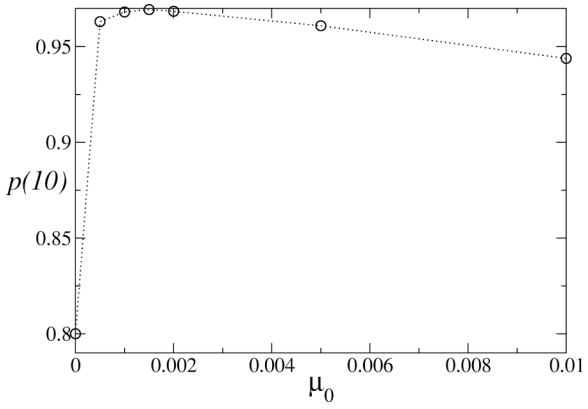

FIG. 9. Simulation results for different mutation rates given an inhomogeneous clonal initial condition. Initially, all the members of the population are clones to each other; 1/5 of the loci are occupied by mismatches and the rest by matches. Here $L=170$, $\phi=0.1$, and $N=5\times10^5$.

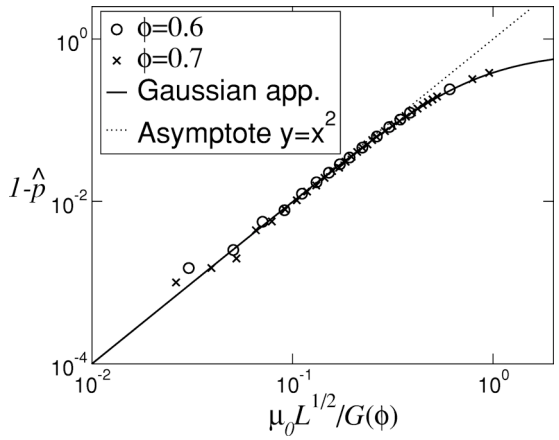

FIG. 10. Equilibrium value of individual match probability p as function of $\mu_0\sqrt{L}/G(\phi)$. The $y=x^2$ curve denotes the asymptotic behavior of p in strong selection (low mutation) resulting from the Gaussian approximation, Eq. (24). For simulation results, $L=170$, $\mu_0=0.01$, $N=10^4$.

librium properties, focusing on the mean number of matches in equilibrium. This quantity plays the role here that genetic load does in normal evolution problems. Strictly speaking, genetic load is the difference in fitness between the equilibrium population and the optimal fitness state. In our case, the optimal state has fitness value 1, and the mean fitness of the population is simply ϕ, so the difference is trivially $1-\phi$. However, since the purpose of the experiments is to maximize the phenotypic value, i.e., the number of matches, the mean number of matches is a reasonable way to characterize the equilibrium state, and the "cost" of mutations. Going back to the language of DNA-protein binding, we are using the mean binding affinity of the DNA sequence to characterize the equilibrium.

Under the Gaussian approximation, the equilibrium value $\hat{p}$ can be found by setting $p(t+1)=p(t)$ in Eq. (17), yielding

$$\frac{\hat{p}(1-\hat{p})}{(\mathcal{A}\hat{p}-1)^2}=\frac{1}{(\mathcal{A}-1)^2}\left(\frac{\mu_0\sqrt{L}}{G(\phi)}\right)^2. \tag{22}$$

As shown in Fig. 10, for weak and intermediate selection strength this result agrees with the simulation data. For the parameter regime of strong selection (or very small mutation rate), $\mu_0\sqrt{L}/G(\phi)$ is no longer a good scaling variable, and the Gaussian approximation begins to deviate from the exact result.

For weak selection **(**i.e., $[\mu_0\sqrt{L}/G(\phi)]^2\gg1$**)** the equilibrium probability is near $1/\mathcal{A}$. We have

$$\hat{p}-\frac{1}{\mathcal{A}}=\frac{(\mathcal{A}-1)^{3/2}}{\mathcal{A}^2}\frac{G(\phi)}{\mu_0\sqrt{L}}. \tag{23}$$

When selection is relatively strong, i.e., when $[\mu_0\sqrt{L}/G(\phi)]^2\ll1$, the equilibrium position is

$$1-\hat{p}=\left(\frac{\mu_0\sqrt{L}}{G(\phi)}\right)^2. \tag{24}$$

As shown in Fig. 10, for the parameters employed there, this power law is accurate in the region of $0.1\leqslant\mu_0\sqrt{L}/G(\phi)<1$. [A more precise statement about the lower cutoff for the validity of Eq. (24) is $\mu_0 L>\ln\phi^{-1}$; see discussion below]. This $\mu_0{}^2$ dependence has also been derived by Kondrashov [24] for a different type of dynamical selection scheme.

It is well known that in a fixed smooth landscape, the mutational load is $\mu_0 L$ in the strong selection limit. We see from the above that for our breeding problem, strong selection, at least in the Gaussian approximation, produces a load that scales with $\mu_0{}^2$ rather than the usual μ_0. The reason for this scaling is easy to understand: With recombination, the change due to selection of pL, the number of matches, is proportional to the width of the distribution $\sqrt{Lp(1-p)}\approx\sqrt{L(1-p)}$ (which is independent of the Gaussian approximation), and mutation reduces pL by $-\mu_0 L$. Balancing the two effects we arrive at the $\mu_0{}^2$ scaling. This simple argument shows that the $\mu_0{}^2$ law is a generic result for the genetic load when recombination is at work (which happens whenever the selected population still occupies a number of different states). When selection strength is sufficiently strong that the selected population lies almost entirely in the optimal state, recombination becomes ineffective as there is no diversity within the selected population. As a result [as shown in Fig. 10], the Gaussian result is no longer valid. In fact, as we shall show below, in this case the usual mutational load of $O(\mu_0)$ takes over.

For small mutation rates, as we shall see, the mean mismatch number is small at equilibrium, except for very weak selection. This is in contradistinction to the dynamic case, where we focused on the case where there were a large number (order L) of initial mismatches. There, for most of the time, the Gaussian approximation is quite adequate. In equi-

librium, however, this is generically not the case, and a more careful treatment is warranted.

To study this issue in more detail, we can make use of a more powerful approximation scheme that accurately covers both the strong selection and extremely strong selection cases so as to find the equilibrium position. Since $Lp \sim L$, we shift our focus to the mismatch probability $q \equiv 1-p \sim 0$. Before selection, the population distribution in terms of mismatch number k should follow a Poisson distribution $P_k = \mathcal{Q}^k e^{-\mathcal{Q}}/k!$, where $\mathcal{Q} \equiv Lq$. Call C_{k_0} the cumulative probability of being in a state with $k \leqslant k_0$. Selection results in

$$\phi = C_{k_0-1} + \alpha P_{k_0}, \tag{25}$$

where k_0 is the threshold, which is determined by the condition that $C_{k_0-1} < \phi \leqslant C_{k_0}$. $\alpha < 1$ counts the partial selection on the members of population with k_0 mismatches.

The results of selection, recombination, and mutation are

$$q_r = \frac{1}{L\phi}\left(\sum_{k=0}^{k_0-1} kP_k + \alpha k_0 P_{k_0}\right), \tag{26}$$

$$q = q_r + \mu_0 - \frac{\mathcal{A}}{\mathcal{A}-1}\mu_0 q_r, \tag{27}$$

where q_r (q) is the individual mismatch probability right after recombination (mutation). The last term is negligible as it arises from extremely rare compensatory beneficial mutations. Combining Eqs. (25)–(27), we arrive at the following equation that determines the equilibrium average mismatch number $\hat{\mathcal{Q}}$:

$$\phi = \frac{e^{-\hat{\mathcal{Q}}}}{k_0 + U - \hat{\mathcal{Q}}} \sum_{k=0}^{k_0-1} \frac{k_0 - k}{k!}\hat{\mathcal{Q}}^k, \tag{28}$$

where $U \equiv \mu_0 L$ is the genomic mutation rate. The physical $\hat{\mathcal{Q}}(\phi)$ curve is given by the envelope of the various $\hat{q}(\phi, k_0)$ for different k_0's, and is piecewise analytic. Figure 11 shows the difference between this solution and the solution given by the Gaussian approximation. Even though the Gaussian approximation has generally the right trend, for this small mutation rate the Gaussian approximation is numerically off by a large percentage, except for extremely weak selection. Note also that for $\phi < e^{-U}$, selection only preserves the optimal state at equilibrium, recombination stops working, and Eq. (28) produces $\hat{\mathcal{Q}} = L - L\hat{p} = U$, the conventional mutational load, which is a result that is unobtainable from the Gaussian approximation.

Now we come back to the question of how much of an improvement is conferred by recombination. Figure 12 shows a direct comparison of the mean-field equilibrium state $\hat{p}$ with and without recombination, for the same value of mutation rate μ_0 and sequence length L. We see that recombination improves the number of matches in equilibrium, especially for intermediate selection strength. In the strong selection limit (i.e., when $\phi < e^{-U}$), as discussed above, the process of recombination is in fact neutral because of a lack of diversity in the selected population. For the regime of intermediate selection strength, the equilibrium position for the maximal recombination protocol is given by Eq. (22). For the no-recombination case, we use a result from Cohen and Kessler [28]:

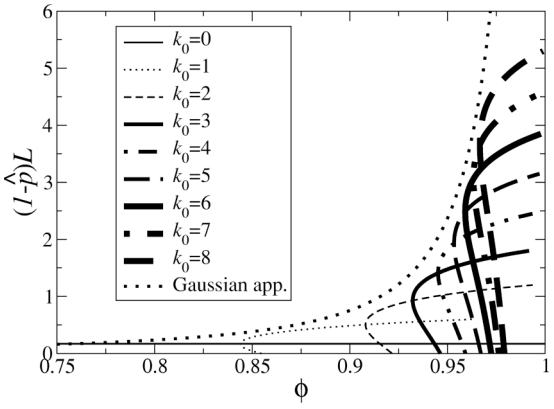

FIG. 11. Theoretical evaluation of equilibrium value of mean mismatches: comparison of Gaussian approximation and Poisson approximation. Curves corresponding to different k_0 values are trial solutions of Eq. (28) with different choices of threshold k_0. The envelope of the trial solutions is the physical solution. For comparison, the solution given by Gaussian approximation is also plotted. Here $L=170$, $\mu_0=0.001$.

$$-\frac{\ln \phi}{U} = \frac{1}{\mathcal{A}-1} + \frac{\mathcal{A}-2}{\mathcal{A}-1}\hat{p}_A - \sqrt{\frac{4}{\mathcal{A}-1}\hat{p}_A(1-\hat{p}_A)}, \tag{29}$$

when $\hat{p}_A$ denotes the equilibrium value of total number of matches divided by chromosome length L for the no-recombination case. This expression is appropriate for $L(1-\hat{p}_A) \gg 1$, so that selection is not too strong; and $\hat{p}_A - 1/\mathcal{A} \gg L^{-1/2}$, which defines the weak selection regime for the no-recombination case. To better understand the difference

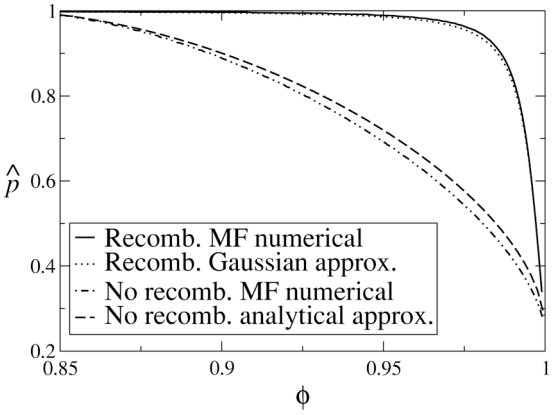

FIG. 12. Comparison of equilibrium positions for two evolutionary protocols: maximal recombination [mean-field numerical result from Eqs. (A6), (A8), and (A9); Gaussian approximation result from Eq. (22)] and no recombination [mean-field numerical result from Cohen and Kessler [28]; analytical approximation from Eq. (29)]. Here $L=170$ and $\mu_0=0.001$.

shown in Fig. 12, we reexpress Eq. (24) to simplify the comparison with Eq. (29),

$$1-\hat{p}=\mu_0\frac{U}{G(\phi)^2}. \tag{30}$$

For intermediate selection, so that $-\ln\phi/U\sim 1$ and $U/G(\phi)^2\sim 1$, we have that $\hat{p}_{\mathrm{A}}$ is an $O(1)$ function whereas $1-\hat{p}\sim\mu_0$. Thus the recombination curve remains near $\hat{p}=1$ until ϕ is close to 1, at which point it takes a sharp dive; the no-recombination curve moves downward continuously. Thus, if selection is not very weak, the equilibrium state under recombination is much less sensitive to the degree of selection, and lies very close to the optimal state.

Now, we move on to consider the region where the recombination $\hat{p}$ takes its dive. Here the picture is most clear in the limit where we take U to be a constant of order 1 and L to be asymptotically large (or equivalently, μ_0 to be small $\sim L^{-1}$ for large L). From Eq. (22), we know that the dive takes place at $1-\phi\sim L^{-1/2}$ (up to logarithmic corrections in $1-\phi$), where $\hat{p}$ is of order 1 and away from both 1 and $1/\mathcal{A}$. On the other hand, when $1-\phi\sim L^{-1/2}$, $1-\phi$ is still big enough for Eq. (29) to be valid for the no-recombination case. Using the smallness of $(1-\phi)/U\sim L^{-1/2}$, we find that

$$\frac{(1-\phi)}{U}\sim L^{-1/2}\approx\frac{A(\hat{p}_{\mathrm{A}}\mathcal{A}-1)^2}{4(\mathcal{A}-1)^2}, \tag{31}$$

showing that $\hat{p}_{\mathrm{A}}$ is close to $1/\mathcal{A}$ with corrections of order $L^{-1/4}$. Thus, we have shown (in the limit of small μ_0 and large L) that, as long as selection in the maximal recombination case is not very weak (i.e., $1-e^{-U}>1-\phi\geqslant L^{-1/2}$), $\hat{p}$ is a finite distance from the random result $1/\mathcal{A}$, and recombination is beneficial.

Now we move on to the weak selection regime, again focusing on the limit of small μ_0 and large L (while maintaining $U\sim 1$). We first discuss the regime where $L^{-1/2}\gg 1-\phi\gg L^{-1}$. In fact, in this regime $\hat{p}$ is close to $1/\mathcal{A}$. Using the weak selection approximation for $\hat{p}$, i.e., Eq. (23), we obtain that $\hat{p}-1/\mathcal{A}\sim(1-\phi)L^{1/2}$. In the same regime, Eqs. (29) and (31) are still valid for the no-recombination case, hence $\hat{p}_{\mathrm{A}}-1/\mathcal{A}\sim\sqrt{1-\phi}\ll\hat{p}-1/\mathcal{A}$, showing that here as well recombination is beneficial. When selection is even weaker so that $1-\phi\sim L^{-1}$, $\hat{p}-1/\mathcal{A}\sim L^{-1/2}\sqrt{\ln L}$. On the other hand, the no-recombination system exhibits the scaling [28]

$$\hat{p}_{\mathrm{A}}-\frac{1}{\mathcal{A}}\sim L^{-1/2}f\left(\frac{(1-\phi)}{\mu_0}\right), \tag{32}$$

where f denotes a scaling function of $(1-\phi)/\mu_0$. For $1-\phi\sim L^{-1}\sim\mu_0$, we have

$$\frac{\hat{p}-1/\mathcal{A}}{\hat{p}_{\mathrm{A}}-1/\mathcal{A}}\sim\sqrt{\ln L}. \tag{33}$$

Thus, in this somewhat unrealistic regime as well, the recombination equilibrium position is much closer to the optimal state. Finally, in the ultraweak selection limit, both $\hat{p}_{\mathrm{A}}$ and $\hat{p}$ converge to $1/\mathcal{A}$.

V. FINITE-POPULATION EFFECT

In the previous sections, we have assumed an infinite population, and comparison with simulations has shown that this approximation is appropriate most of the time. However, under the experimentally relevant situations, where very strong selection and very low mutation rate are employed, finite-population effects can show up and be potentially important. As an extreme example, if we choose the best member from the population, i.e., $N\phi=1$, then recombination has nothing to work with and offers no benefit. In this section we discuss the population size above which the evolutionary dynamics and equilibrium can essentially be deemed the same as those of infinite populations. To rid ourselves of the impact of initial conditions, we assume an initially diversified population. Two population sizes are relevant for this problem, the total population size N at the beginning of each generation, and the selected population size $N\phi$. We will focus on the case most often encountered in *in vitro* evolution experiments, namely N very large, but $N\phi$ is rather small. In the following analysis, we set ϕ constant, and compare the evolutionary trajectory for different population sizes N. We start again from the simplest case of no point mutation, shown in Fig. 13(a). It can be seen that in general the smaller the population size, the worse the evolutionary efficiency and the equilibrium state. Under these conditions, the finite-population effect comes from the genetic association between loci that results in the loss of favorable alleles from the entire population when the population is subject to selection.

We can estimate the population size N_0 above which the population dynamic approaches infinite-population results by estimating the average number of lost favorable alleles at all loci in the selected population. As the individual match probability p is lowest at the beginning of evolution, if there is any loss of favorable allele, it is most likely to happen at the first round of selection, hence we focus on the first round of selection and recombination. Right after the first round of selection, the probability for a specific locus to lose the favorable allele is $[1-p(1)]^{N\phi}$. Therefore the average number of loci losing the favorable allele is $L(1-p)^{N\phi}$. The opposite way of saying this is that for the population to retain favorable alleles at all loci after the first round of selection, we require $L(1-p)^{N_0\phi}<1$. This gives the following estimate of N_0:

$$N_0=\frac{1}{\phi}\frac{-\ln L}{\ln[1-p(1)]}, \tag{34}$$

where $p(1)=p(0)+G(\phi)\sqrt{p(0)[1-p(0)]/L}$. If in the first round the above criterion is met, then it is less likely for the population to lose any favorable alleles at successive rounds when subject to selection, as $p(t)$ increases in later rounds and $L(1-p)^{N\phi}$ becomes increasingly smaller than 1. Comparison with simulation results shows that Eq. (34) gives a reasonable estimate (at least correct in order of magnitude).

When a (weak) point mutation is involved, the dynamics can be separated into two regimes: (a) recombination dominated evolution, where for loci that have a favorable allele in

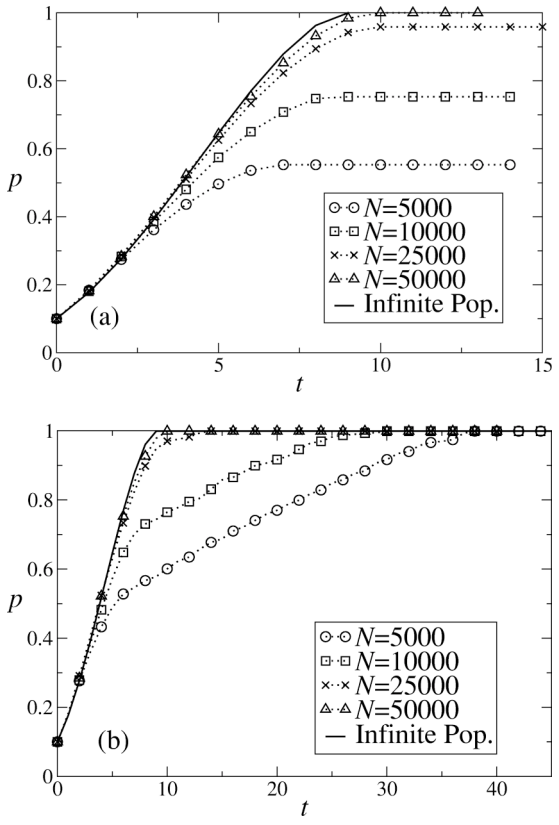

FIG. 13. Finite-population effect on the dynamics and equilibrium of maximal recombination: Simulation results for the case of (a) no mutation ($\mu_0=0$) and (b) weak mutation ($\mu_0=0.001$). With mutation, two regimes exist: (a) at the beginning recombination dominates evolution, as in the mutation-free case and (b) at later stages mutations are essential, as they are needed to find lost favorable alleles. Here evolution is constrained by the rate for mutation to find beneficial alleles, thus slower. Here $L=170$, $\phi=0.001$. Infinite-population trajectories are theoretical results taken from Eq. (17).

them, recombination helps those favorable alleles to spread to the whole population, as in the mutation-free case, and (b) mutation/recombination evolution, where loci that have lost their favorable alleles due to selection need mutations to find them again. The dynamics in this regime is significantly slower, as seen in Fig. 13(b). In this case, Eq. (34) still identifies the smallest population size N_0 that behaves more or less like an infinite population. As a side note, the flip side of the finite-population issue is that there exists an optimal selection pressure for a particular population size; an estimate of this optimal ϕ can be found from the condition $L(1-p)^{N\phi}<1$.

The estimate of N_0 given ϕ (or equivalently the optimal ϕ_0 given N) has practical implications. It gives a population size that is just enough for the evolution to proceed at its fastest rate (given a selection strength). Since in real-life experiments selection routinely involves screening the population, which is costly and time consuming, a smaller while still equally effective population size would be helpful. Of course, the above arguments assumed a random initial library. If one starts from a population of a single clone, then as already explained above, even at the beginning mutation is absolutely necessary to generate diversity. For this case it is more difficult to directly estimate the finite-population effects.

VI. DISCUSSION

In this work we proposed and studied a simplified model of DNA shuffling, a very important evolutionary protocol for directed evolution. We investigated this model from a population genetics' point of view, as a multilocus evolutionary model incorporating both recombination and point mutation and subject to dynamical truncation selection. Our specific recombination scheme, as an extreme limit of multiparent multicrossover genetic mixing employed in DNA shuffling, enables us to pursue analytical results for the dynamical and equilibrium features. To summarize, we derived the recursion relations that completely characterize the evolutionary process, which shows explicitly how recombination helps to speed up the evolution. Assuming a large number of loci, we found that selection and mutation affect the evolutionary process only through a combination of their respective parameters. When the evolution is relatively slow so that the discrete recursion relation can be approximated by a continuous differential equation, we could solve for the evolutionary trajectory, and rigorously prove in special cases that it is indeed faster than that for the case without recombination. We also investigated how different initial conditions and mutation rates affect the evolutionary trajectory. As to the equilibrium properties of the system, we found that the genetic load has a scaling form different from that without recombination, as long as the selection is not too strong. We also showed that, in terms of equilibrium state, recombination will help the most when selection is neither too strong nor too weak. At the end, we discussed the finite-size effect under conditions relevant to the shuffling experiments, and estimated the minimal population size above which the population behaves like an infinite population.

We have been focusing on a model with specific conditions, but our results are actually applicable to broader situations. Let us first discuss selection.

(a) For simplicity, we employed a dynamical truncation selection. In the case of DNA sequence evolution via binding to protein, selection is "smoother" (i.e., the step corners are rounded [13,15]). This simplification does not cause any substantial difference in our analytical results as long as the change in binding affinity is large (i.e., the step is steep enough).

(b) Our approach and results are also applicable to the type of dynamical (*soft*) selection studied by Kondrashov [24]. There, the selection is implemented as a function $W(X)$, where $X=(m-\bar{m})/\sigma$, i.e., the phenotypic difference between the phenotype and the mean $\bar{m}$, divided by the standard deviation σ of the phenotypic distribution. The effect of this type of selection on a population with a Gaussian distribution is

$$m_s - m = \sigma\delta, \tag{35}$$

where δ is a quantity solely characterized by the selection function $W(X)$ and does not depend on the population phenotypic distribution [see Eqs. (7) and (14) of Kondrashov [24]]. It is evident that Eq. (35) is analogous to Eq. (5) except that one needs to replace $G(\phi)$ with δ. Therefore, when one exchanges $G(\phi)$ with δ, all our results under the Gaussian approximation remain valid, including various scaling arguments and solutions to the continuous evolutionary equations.

Now we move to recombination. One immediate question is what would change when the recombination scheme is more realistic. A straightforward generalization of the maximal recombination toward the realistic situation is that we can cut each DNA sequence into segments of equal length, with the maximum recombination as the extreme case of length equal to 1. Computer simulations show that in general the longer the segment length, i.e., the less crossovers for each chromosome, the worse the performance. For a general segment length, one can derive a hierarchy of recursion relations that relate the cumulants of the population distribution at the current generation to those of previous generation, an approach that has been applied to a multilocus system on a fixed landscape [29,30]. How to close the chain of recursion relations remains to be studied. The difficulty is associated with the remaining linkage within each segment. In any event, the maximal recombination model can serve as a theoretical upper limit which is qualitatively correct and even reasonably accurate quantitatively.

Our study has been guided by the *in vitro* evolution process, where the intensity of recombination and mutation are both prescribed by the experimenter. This might not be the case in a more natural setting, where the evolution of recombination (mutation) itself can be an important aspect of the problem [31]. Our model does not incorporate a modifier gene that explicitly controls recombination rate. In such a modifier approach, the recombination rate can itself evolve because the modifier gene is under indirect selection due to its association with other genes on the same chromosome under direct selection [31].

Finally, we believe that our model may be relevant to some examples of natural evolution: (a) dynamical truncation selection can also exist in nature, for example, the mutual selection in a host-parasite system and (b) recombinations of multiparent, multicrossover type do exist in nature: certain RNA viruses have multiple segments in their genome, and when multiple viruses infect the same host and new virus particles are made, recombinations of the type we study can happen [32]. With the ease of analytical treatment of this model, we hope it can serve as a theoretical starting point for understanding interplay among various aspects such as recombination, mutation, and selection, and as a testing ground where general statements can be tried out.

ACKNOWLEDGMENTS

We would like to give special thanks to L. Chao for stimulating discussions. We also acknowledge interactions with S. P. Otto, A. Poon, W. P. C. Stemmer, and J. Widom. In addition, we thank the anonymous reviewer for very helpful comments. This work was partially funded by the NSF sponsored Center for Theoretical Biological Physics (Grants Nos. PHY-0216576 and 0225630).

APPENDIX A: EVOLUTIONARY RECURSION RELATIONS FOR ARBITRARY CHROMOSOME LENGTH

In the discussion of the evolutionary dynamics and equilibrium state, we make use of a Gaussian (and a Poisson distribution) to approximate a binomial distribution, which is accurate when $L \gg 1$ (except for the tails of the distribution). In fact, one can relax this condition and work directly with the binomial distribution, with the help of special function $I_x(a,b)$, the incomplete Beta function. $I_x(a,b)$ is defined as [33]

$$I_x(a,b) = \frac{1}{B(a,b)} \int_0^x dt\ t^{a-1}(1-t)^{b-1}, \tag{A1}$$

where $B(a,b)$ is the complete Beta function.

The probability density of m matches is a binomial distribution,

$$P_p(m,L) \equiv C_L^m p^m (1-p)^{L-m}. \tag{A2}$$

The cumulative probability density (for $m>0$) is related to the incomplete Beta function via

$$\sum_{i=m}^{L} P_p(i,L) = I_p(m, L+1-m), m > 0. \tag{A3}$$

We have, at the selection,

$$\phi = \alpha P_p(m_0,L) + \sum_{i=m_0+1}^{L} P_p(i,L), \tag{A4}$$

where again m_0 is the threshold and population members at the state with m_0 matches may be partially selected.

The new individual match probability after selection and recombination is

$$Lp^r = \frac{1}{\phi}\left(\alpha m_0 P_p(m_0,L) + \sum_{i=m_0+1}^{L} iP_m(i,L)\right). \tag{A5}$$

Incorporating the mutational process and making use of Eq. (A4), we have, for $0<m_0<L$,

$$p_{t+1} = \frac{\mu_0}{\mathcal{A}-1} + \frac{1 - \dfrac{\mathcal{A}}{\mathcal{A}-1}\mu_0}{L\phi}[m_0\phi + Lp_t I_{p_t}(m_0, L-m_0) - m_0 I_{p_t}(m_0+1, L-m_0)], \quad 0<m_0<L, \tag{A6}$$

where we made use of the following identity:

$$\sum_{i=m_0+1}^{L} iP_p(i,L) = LpI_p(m_0, L-m_0). \tag{A7}$$

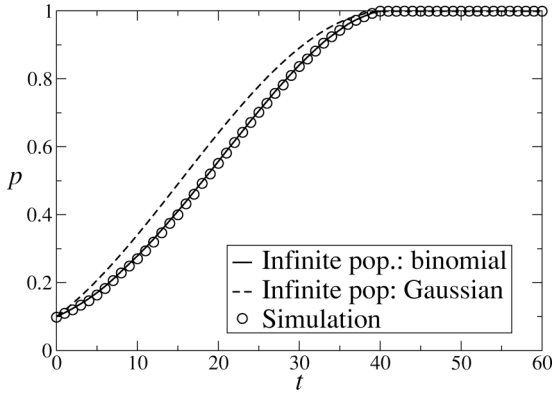

FIG. 14. Comparison of two theoretical approaches with simulation result. Solid curve is the result of Eq. (A6). Dashed curve is the result of Eq. (17). Here $L=10$, $\phi=0.9$, and $\mu_0=0.001$. For the simulation $N=10^4$.

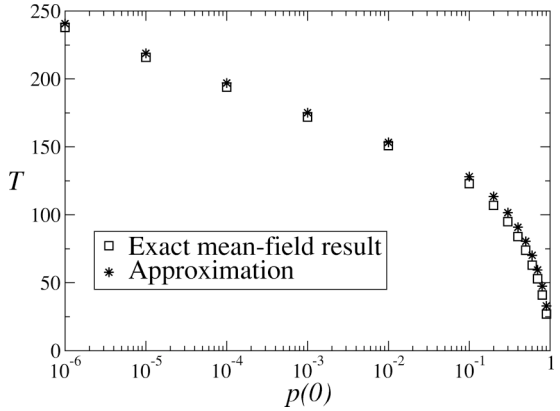

FIG. 15. Comparison of the evolution time for maximal recombination between the exact mean-field result and the approximation. The exact mean-field result is obtained from Eqs. (A6), (A8), and (A9). The approximation is given in Eq. (A12). Here $L=100$, $\phi=0.9$, and $\mu_0=0$.

There are two special cases where the recursion relation in Eq. (A6) needs to be modified. When $m_0=0$,

$$p_{t+1}=\frac{\mu_0}{\mathcal{A}-1}+\left(1-\frac{\mathcal{A}}{\mathcal{A}-1}\mu_0\right)\frac{p_t}{\phi}, m_0=0. \tag{A8}$$

When $m_0=L$,

$$p_{t+1}=1-\mu_0, m_0=L. \tag{A9}$$

The dynamics and equilibrium properties follow from Eqs. (A6), (A8), and (A9). Figure 14 shows that, when the chromosome length L is short, indeed this approach significantly improves the agreement with simulations.

Making use of the formalism developed above, we now revisit the issue of evolution time when the mutation rate is zero. In Sec. III A we used Gaussian approximation to analyze the problem. As mentioned there, the result obtained has a number of serious limitations. Here we focus on the $p(0)\sim 0$ limit. Assuming selection is weak, one would expect that during the early stage of the evolution $m_0=0$. At this stage, we have

$$p_{t+1}=p_t/\phi. \tag{A10}$$

The threshold m_0 increases above 0 when the probability density at state 0, $[1-p(0)/\phi^t]^L$, is less than $1-\phi$. Hence the time T_0 it takes the population to increase its threshold above 0 is

$$T_0=\frac{\ln p(0)}{\ln \phi}-\frac{\ln[1-(1-\phi)^{1/L}]}{\ln \phi}. \tag{A11}$$

Using Eq. (12) for estimation of the evolution time after the early stage, we obtain the total evolution time as

$$T=\begin{cases} T_0+\left[\dfrac{\pi}{2}+\sin^{-1}[2(1-\phi)^{1/L}-1]\right]\dfrac{\sqrt{L}}{G(\phi)}, & 1-p(0)>(1-\phi)^{1/L} \\ \left[\dfrac{\pi}{2}+\sin^{-1}[1-2p(0)]\right]\dfrac{\sqrt{L}}{G(\phi)}, & 1-p(0)\leqslant(1-\phi)^{1/L}. \end{cases} \tag{A12}$$

The first case is for $p(0)$ small, with the initial selection threshold $m_0=0$. The second case is for $p(0)$ not particularly small, with the initial selection threshold $m_0>0$. How good is this estimate? Figure 15 shows the comparison of the evolution time between the exact mean-field result [obtained by following recursion relations in Eqs. (A6), (A8), and (A9)] and the approximation obtained in Eq. (A12). The agreement between the two is reasonably good overall.

Now that we have an estimate that is applicable even for small $p(0)$, let us compare the evolution time for maximal recombination with that for pure enrichment, in the limit of $p(0)\to 0$. T and T_A are both proportional to $\ln p(0)/\ln \phi$. However, in the case of enrichment, the singularity is associated with chromosome length L [see Eq. (14)], whereas for maximal recombination, L only appears in the nonsingular

part. Therefore, in the small $p(0)$ limit, maximal recombination is again beneficial.

APPENDIX B: EXPLICIT SOLUTION OF THE EVOLUTION EQUATION

Assuming that at $t=0$ the individual match probability is $p(0)$, the continuous-time approximation of Eq. (18) is

$$\frac{dp}{dt'} = (1 - \mathcal{A}p) + C\sqrt{p(1-p)}, \tag{B1}$$

where $t' \equiv [\mu_0/(\mathcal{A}-1)]t$ and $C \equiv (\mathcal{A}-1)G(\phi)/(\mu_0\sqrt{L})$. This continuous-time approximation is appropriate when change of p between consecutive rounds is small (i.e., when selection is not very strong). For most parameter values the right-hand side is positive, therefore p keeps increasing. The general solution of Eq. (B1) is

$$\begin{aligned}\frac{\mathcal{A}^2 + C^2}{2}t' = &-\frac{\mathcal{A}}{2}\ln\frac{|1 - \mathcal{A}p + C\sqrt{p(1-p)}|}{|1 - \mathcal{A}p(0) + C\sqrt{p(0)[1-p(0)]}|} \\ &+ \frac{C(1-\mathcal{A}/2)}{\sqrt{\Delta}}\left[\ln\frac{|2\sqrt{(1-p)/p} + C - \sqrt{\Delta}|}{2\sqrt{(1-p)/p} + C + \sqrt{\Delta}}\right. \\ &\left. - \ln\frac{|2\sqrt{[1-p(0)]/p(0)} + C - \sqrt{\Delta}|}{2\sqrt{[1-p(0)]/p(0)} + C + \sqrt{\Delta}}\right] \\ &- C\left(\tan^{-1}\sqrt{\frac{1-p}{p}} - \tan^{-1}\sqrt{\frac{1-p(0)}{p(0)}}\right),\end{aligned} \tag{B2}$$

where $\Delta \equiv C^2 - 4(1-\mathcal{A})$. In the absence of mutation the above equation has been solved, the solution a sinusoidal function [see Eq. (12)]. The other extreme situation is when the selection strength ϕ is 1, which leads to $G(\phi)=0$ and $C=0$. In this case p goes to $1/\mathcal{A}$ exponentially with time constant $-\ln[1-\mathcal{A}\mu_0/(\mathcal{A}-1)] \approx \mathcal{A}\mu_0/(\mathcal{A}-1)$.

[1] R. J. Redfield, Nat. Rev. Genet. **2**, 634 (2001).
[2] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, *Molecular Cell Biology*, 4th ed. (Freeman, New York, 1999).
[3] S. P. Otto and T. Lenormand, Nat. Rev. Genet. **3**, 252 (2002).
[4] E. Baake, J. Math. Biol. **42**, 455 (2001).
[5] A. L. Kurtzman, S. Govindarajan, K. Vahle, J. T. Jones, V. Heinrichs, and P. A. Patten, Curr. Opin. Biotechnol. **12**, 361 (2001).
[6] E. T. Farinas, T. Bulter, and F. H. Arnold, Curr. Opin. Biotechnol. **12**, 545 (2001).
[7] W. P. C. Stemmer, Nature (London) **370**, 389 (1994).
[8] W. P. C. Stemmer, Proc. Natl. Acad. Sci. U.S.A. **91**, 10 747 (1994).
[9] M. C. Orencia, J. S. Yoon, J. E. Ness, W. P. C. Stemmer, and R. C. Stevens, Nat. Struct. Biol. **8**, 238 (2001).
[10] M. Barlow and B. G. Hall, Genetics **160**, 823 (2002).
[11] D. A. Kessler, H. Levine, D. Ridgway, and L. Tsimring, J. Stat. Phys. **87**, 519 (1997).
[12] A. Thastrom, P. T. Lowary, H. R. Widlund, H. Cao, M. Kubista, and J. Widom, J. Mol. Biol. **288**, 213 (1999).
[13] P. H. Von Hippel and O. G. Berg, Proc. Natl. Acad. Sci. U.S.A. **83**, 1608 (1986).
[14] W. Peng, U. Gerland, T. Hwa, and H. Levine, Phys. Rev. Lett. **90**, 088103 (2003).
[15] U. Gerland and T. Hwa, J. Mol. Evol. **55**, 386 (2002).
[16] F. H. Arnold, Nature (London) **409**, 253 (2001).
[17] J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).
[18] S. P. Otto and N. H. Barton, Evolution Int. J. Org. Evolution **55**, 1921 (2001).
[19] E. E. Shnol and A. S. Kondrashov, Genetics **134**, 995 (1993).
[20] F. Sun, J. Comput. Biol. **6**, 77 (1999).
[21] G. L. Moore and C. D. Maranas, J. Theor. Biol. **205**, 483 (2000).
[22] M. Kimura and J. F. Crow, Proc. Natl. Acad. Sci. U.S.A. **75**, 6168 (1978).
[23] K. J. Dawson, Linear Algebr. Appl. **348**, 115 (2002).
[24] A. S. Kondrashov, Genet. Res. **65**, 113 (1995).
[25] H.-G. Beyer, *The Theory of Evolution Strategies* (Springer-Verlag, Berlin, 2001).
[26] H. Mühlenbein and D. Schlierkamp-Voosen, Evol. Comput. **1**, 335 (1993).
[27] D. S. Falconer and T. F. C. Mackay, *Introduction to Quantative Genetics*, 4th ed. (Addison-Wesley, England, 1996).
[28] E. Cohen and D. A. Kessler, e-print cond-mat/0301503.
[29] N. H. Barton and M. Turelli, Genetics **127**, 229 (1991).
[30] R. Büger, *The Mathematical Theory of Selection, Recombination and Mutation* (Wiley, Chichester, 2000).
[31] M. W. Feldman, S. P. Otto, and F. B. Christiansen, Annu. Rev. Genet. **30**, 261 (1996).
[32] L. Chao, J. Theor. Biol. **133**, 99 (1988).
[33] *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, edited by M. Abramowitz and I. A. Stegun (Dover, New York, 1970).