Role of finite populations in determining evolutionary dynamics

Tane S. Ray, Karl A. Payne, and L. Leo Moseley
Computational Physics Laboratory, Department of Computer Science, Mathematics and Physics, University of the West Indies, Cave Hill Campus, Barbados

(Received 16 November 2007; published 14 February 2008)

The connection between the finite size of an evolving population and its dynamical behavior is examined through analytical and computational studies of a simple model of evolution. The infinite population limit of the model is shown to be governed by a special case of the quasispecies equations. A flat fitness landscape yields identical results for the dynamics of infinite and finite populations. On the other hand, a monotonically increasing fitness landscape shows "epochs" in the dynamics of finite populations that become more pronounced as the rate of mutation decreases. The details of the dynamics are profoundly different for any two simulation runs in that events arising from the stochastic noise in the pseudorandom number sequence are amplified. As the population size is increased or, equivalently, the mutation rate is increased, these epochs become smaller but do not entirely disappear.

DOI: 10.1103/PhysRevE.77.021909 PACS number(s): 87.23.−n, 07.05.Tp, 05.10.−a

INTRODUCTION

The modern theory of the evolution of biological organisms is based upon simple and demonstrable principles. The change in the frequency of alleles in a population resulting from natural selection and the generation of new alleles through imperfect reproduction is well established. Nevertheless, the details of how evolution occurs, including the level at which natural selection acts and the role of finite population size, are current topics of great interest in the research community [1].

Attempts have been made to address some of the complex issues in evolutionary theory by means of simple models that are amenable to analytical and computational studies [2–7]. Evolving populations are modeled without incorporating the enormous amount of detail that dictates the course of real biological evolution. The model introduced in this paper is a particularly simple one which effectively bridges the gap between several other models of evolution. Indeed, it is morphologically identical, for example, to a computational version of the quasispecies model described by Campos and Fontanari [8] where the population is made up of a set of binary strings that undergo asexual reproduction. A particular binary string representing a single member can be interpreted as a portion of that member's genome.

The model may also be viewed as a modified, multiple-allele Wright-Fisher model with mutations and weighted selection [9–11], or, alternatively, as a drastically simplified version of the "artificial life" platform, AVIDA [12,13]. Furthermore, it is similar to the "Royal Road" genetic algorithm where several blocks comprising a string are replaced by single bits [14,15]. Despite the fact that the model is simpler than either the Royal Road algorithm or AVIDA, it retains the basic elements that are necessary for a population to evolve and many features seen in these other models are also evident here. For example, the average fitness at low mutation rates exhibits plateau structures for nontrivial fitness landscapes.

As is the case with some other computational models of evolution, the binary string model examined in this paper is governed by quasispecies equations in the large population limit [16–18]. More precisely, when the product of the population size M and mutation rate μ is large ($\mu M \gg 1$), the dynamics of the simulations approach those given by the quasispecies equations. The departure from the quasispecies equations when the above relation is not satisfied and the subsequent transition from infinite to finite population behavior is a major focus of the present paper. The transition appears especially relevant to determining the course of evolution. In the large population limit, all genomes are present with their correct respective frequencies. Simulations become deterministic in the sense that they can be made to reproduce the quasispecies equations as closely as desired, independent of the pseudorandom number sequence, by increasing the population to a suitably large size while keeping the mutation rate fixed.

On the other hand, a small population does not generally follow the quasispecies equations, as can be seen through the following argument. Let $p_s(t)$ represent the frequency of string s at time t as given by the quasispecies equations. Now consider a simulation where the size of the population is denoted by M. The condition $Mp_s(t) \ll 1$ implies that the string is not usually present in any given simulation run. More importantly, the absence of the string influences the future of the distribution. The finite size of the system can thereby radically change the dynamics from those of quasispecies theory. Furthermore, it will be seen later that this effect is not removed when many simulation runs are averaged.

The results here support the work of several other authors. For example, analytic scaling arguments by Zhang [19] indicate that solutions to the quasispecies equations have a behavior reminiscent of the punctuated equilibrium of Gould *et al.* [20]. Zhang has found that a decrease in the population size while keeping the mutation rate fixed amplifies this effect so that for long periods of time the genome distribution remains in a quiescent or metastable state. Rare events triggered by stochastic noise cause the population to shift to another metastable distribution with higher fitness. These metastable configurations have also been noted in analytical

work on the "strong selection weak mutation" limit [21,22]. Finally, they have been seen in the work of Nimwegen *et al.* [14] using the Royal Road genetic algorithm. They are called "fitness epochs" by these latter authors and that term will be adopted here.

In the present paper, fitness epochs are shown to exist also in the binary string model; even for large populations that are governed by quasispecies equations, where $\mu M \gg 1$, in agreement with the work of Zhang mentioned above. When the inequality does not hold, the fitness epochs become much more pronounced and are of longer duration. The mechanism causing the fitness epochs, their impact on the evolutionary history of the population and the crossover from quasispecies behavior to finite populations will be examined within the context of the binary string model.

It is important to distinguish between the finite population work here and that of Alves and Fonatanari [23] where the authors have been concerned with the so-called "error threshold" phenomenon that pervades the quasispecies literature. The error threshold arises in a system having a fitness landscape that is flat with the exception of one master sequence that has a larger fitness. In the limit where the length of each string as well as the population size is infinite, one observes a critical value of the mutation rate below which the system forms a cloud of quasispecies around the master sequence. For larger mutation rates than the critical one, the distribution of genomes is uniform. By contrast, the work presented here is primarily concerned with short strings that typically have only four bits. In terms of biology, this is equivalent to a gene locus which has a finite number of alleles. In addition, the mutation rate is usually kept low so that the regime is far from the error threshold.

The paper is organized as follows. The binary string model is carefully defined and compared with other evolutionary models. The particular quasispecies equations describing the model in the large population limit $\mu M \gg 1$ are then derived. These are solved analytically for the special case of a flat fitness landscape; i.e., the case of random "genetic drift." Simulation results are presented showing that, even when the above limit does not hold, the dynamics continue to be governed by the quasispecies equations.

This is in stark contrast to the case of populations that are under selective pressure, which is the next topic. Here, a nontrivial fitness landscape is introduced that is a monotonically decreasing function of the Hamming distance between any string in the population and an ideal string. The ideal string can be thought of as effectively representing a fixed environment. In the large population limit where the above condition is satisfied, the system dynamics are again governed by the quasispecies equations. When it is not satisfied, large fitness epochs are observed that cause the dynamics to deviate from quasispecies theory.

DESCRIPTION OF THE MODEL

The model uses a binary string to represent each organism. All of the binary strings comprising the population have the same number of bits (N). For the most part, the data in this paper come from simulations using strings with $N=4$. However, most of the results and conclusions do not depend strongly upon the particular value of N. Also, a single fixed binary string, which has the same length as that of the population strings, represents the environmental niche and can be regarded as the ideal string that has the maximum possible fitness.

Reproduction occurs asexually so that the offspring of a particular binary string are copies of that string. Mutations are introduced into each offspring with a fixed probability. If a mutation occurs a single bit of the offspring, chosen randomly, is changed. This is the only type of mutation present and it is meant to mimic point mutations in a biological genome where one base pair is altered. Most other models in the literature allow any number of simultaneous mutation events so that all bits of any one offspring could be changed, allowing a nonzero transition probability to any other string. Here, by contrast, an offspring can only differ from the parent by a single bit. In this respect, there is a negligible difference between this binary string model and the other models when the mutation rate is low and the string length is small.

The number of offspring produced by a given parent string differs according to a fixed fitness landscape. The simplest case is a flat fitness landscape where all parents have the same fitness and therefore produce the same number of offspring. On the other hand, a monotonically increasing fitness landscape models selection pressure from the environment. The Hamming distance k between the environment string and the parent string (i.e., the total number of corresponding bits between the strings that differ) determines the number of offspring. The smaller the Hamming distance, or the closer the match between the two strings, the more offspring are produced. The number of offspring c_k is calculated using the formula

$$c_k = c_0 e^{-k/n}. \tag{1}$$

The value of c_k generally lies between the two integers, $m < c_k < m+1$. The number of offspring is chosen randomly to be either m or $m+1$ and it is done proportionately so as to make the average c_k equal to that obtained from Eq. (1).

The above equation is designed to give reasonable behavior under the considerations of the model because it produces a sharp decrease in the fecundity with increasing Hamming distance. The prefactor c_0 and the integer n are arbitrary. In the sections that follow, simulation data are presented using four bit strings ($N=4$) where $c_0=3$ and $n=2$.

All members of the population are allowed to reproduce and the population is therefore sampled for the parent strings without replacement. This procedure is different than the standard sampling with replacement that is done, for example, in the unmodified Wright-Fisher model, where noise introduced through the sampling procedure is necessary to cause drift. It has been found that the simulation results of the present model are insensitive to the sampling procedure, as stochasticity is introduced through the culling process described in the next paragraph.

The offspring along with the parents are collected together into an intermediate population which is then randomly culled to reduce it to a size equal to that of the pre-

vious generation. Thus the size of the population remains constant from generation to generation. The restriction of constant population size is the only spatial constraint imposed upon the model and is similar in this respect to a finite-sized system under the mean field approximation of statistical mechanics.

The model contains the basic ingredients essential for evolution. Imperfect replication is modeled through the reproduction algorithm outlined above and natural selection is modeled by means of the difference in fecundity determined by the fitness landscape along with the culling process. The following is a summary of the binary string algorithm:

(1) Begin a generation by listing, in random order, all of the member strings of the population. These are called the "parent" strings. Select the first parent from the list.

(2) Generate offspring by making exact copies of the parent. The number of offspring is determined by the Hamming distance between the parent string and the environment string according to Eq. (1). For the case of a flat fitness landscape, the number of offspring is constant.

(3) Determine if each offspring is a mutant by drawing a random number between zero and one and comparing it with a fixed probability for a mutant string (μ). If in following this procedure an offspring is designated a mutant, choose a single bit of the offspring at random and change it to the opposite value.

(4) If the list of parent strings has not been exhausted select the next parent on the list and go back to step (2). If none are left, randomly select a fixed number M of strings from the intermediate population of parents plus offspring. These M strings comprise the next generation.

(5) The iteration of a generation has been completed. Go to step (1) to proceed to the next generation.

ANALYSIS OF THE MODEL THROUGH RECURRENCE RELATIONS

One can analyze the large population behavior of the model using recurrence relations for the probability distribution. The recurrence relations are equivalent to a special case of the discrete quasispecies equations [18]. Here, the distribution is written in terms of the probability that a string has a Hamming distance k measured with respect to a fixed string representing the environment ($0 \leqslant k \leqslant N$). In writing the recurrence relations an assumption is made that any stochastic noise will be averaged out. It will be shown that this assumption holds true for the case of genetic drift. However, it fails badly for the case of selection pressure until the system reaches equilibrium. The equilibrium distribution given by the recurrence relations is the same as that observed in simulation data.

The recurrence relations for the model will now be constructed. Consider a population of M binary strings. Let the number of strings with Hamming distance k at time t be given by $n_k(t)$. Let the next generation occur at time $t+\Delta t$. Further, let c_k be the number of offspring for a string with Hamming distance k and let the mutation rate μ be the probability that a particular offspring undergoes a point mutation. For a value of k within the range ($0<k<N$), the new value of n_k after one generation will be given by

$$n_k(t+\Delta t) = \frac{[1+(1-\mu)c_k]n_k(t) + \mu\frac{N-(k-1)}{N}c_{k-1}n_{k-1}(t) + \mu\frac{k+1}{N}c_{k+1}n_{k+1}(t)}{M + \sum_{k=0}^{N} c_k n_k(t)} M. \tag{2}$$

The three terms in the numerator represent the average contributions from strings of the parent generation with the respective Hamming distances of k, $k-1$, and $k+1$. The first term is the number of parent strings plus those offspring that are without mutations and are hence exact replicas of the parents.

The second term is the offspring of the $k-1$ parent strings with a mutation that increases the Hamming distance to k. The factor $[N-(k-1)]/N$ gives the probability that the particular bit chosen for mutation is one that *agrees* with the corresponding bit in the environmental string. When it is changed, it will then disagree with the environment string and therefore the Hamming distance will have increased by one.

Analogous reasoning applies to the third term, which represents those offspring of parent strings with Hamming distance $k+1$ that undergo a mutation which decreases the Hamming distance to k. The probability that the particular bit chosen to be altered is one that *disagrees* with the corresponding environmental string bit is $(k+1)/N$. Finally, the denominator in Eq. (2), along with the population size M multiplying the numerator, represents the effect of culling.

The equations for $n_k(t)$, where $k=0$ and $k=N$, are of the same form as Eq. (2) except that there are only two terms in the numerator. The $n_0(t)$ equation does not have the $k-1$ term and the $n_N(t)$ equation does not have the $k+1$ term.

The frequency of a string with Hamming distance k is $p_k(t)=n_k(t)/M$. Writing Eq. (2) in terms of $p_k(t)$ gives the recurrence relations for the distribution of strings, Eq. (3), with k in the range ($0<k<N$). Equations (4) and (5) for $k=0$ and $k=N$ are included for convenience of reference.

$$p_k(t+\Delta t) = \frac{[1+(1-\mu)c_k]p_k(t) + \mu\frac{N-(k-1)}{N}c_{k-1}p_{k-1}(t) + \mu\frac{k+1}{N}c_{k+1}p_{k+1}(t)}{1+\sum_{k=0}^{N} c_k p_k(t)}, \tag{3}$$

$$p_0(t+\Delta t) = \frac{[1+(1-\mu)c_0]p_0(t) + \frac{\mu}{N}c_1 p_1(t)}{1+\sum_{k=0}^{N} c_k p_k(t)}, \tag{4}$$

$$p_N(t+\Delta t) = \frac{[1+(1-\mu)c_N]p_N(t) + \frac{\mu}{N}c_{N-1}p_{N-1}(t)}{1+\sum_{k=0}^{N} c_k p_k(t)}. \tag{5}$$

It is difficult to solve these nonlinear equations for the time dependence of the distribution for general k. However, one can solve for the equilibrium distribution by diagonalizing the evolution matrix constructed from the numerators of Eqs. (3)–(5). The dominant eigenvector is then the equilibrium state [3,18]. In lieu of an analytical solution for the complete dynamics, these equations may be iterated numerically to give the full history of the distribution and the result may then be compared with simulation data.

CASE I: FLAT FITNESS LANDSCAPE

On a flat fitness landscape, the number of offspring produced by each parent is independent of the Hamming distance, $c_k \to c$. Equations (3)–(5) then reduce to linear recurrence relations. Further insight can be obtained through approximating $\Delta p_k(t) = p_k(t+\Delta t) - p_k(t)$ as the derivative dp_k/dt. This will be a good approximation to the recurrence equations when $|\Delta p_k(t)| \ll 1$; a condition that holds when the mutation rate is small ($\mu \ll 1$) and/or when the system is close to equilibrium. The result is the following set of differential equations:

$$\frac{dp_k}{dt} = \frac{\mu c}{(1+c)}\left\{-p_k(t) + \frac{N-(k-1)}{N}p_{k-1}(t) + \frac{k+1}{N}p_{k+1}(t)\right\}, \tag{6}$$

$$\frac{dp_0}{dt} = \frac{\mu c}{(1+c)}\left\{-p_0(t) + \frac{1}{N}p_1(t)\right\}, \tag{7}$$

$$\frac{dp_N}{dt} = \frac{\mu c}{(1+c)}\left\{-p_N(t) + \frac{1}{N}p_{N-1}(t)\right\}. \tag{8}$$

These equations can be solved exactly using standard techniques for a homogenous linear system with constant coefficients [24]. Writing all $p_k(t)$ as a column vector of length $N+1$ puts Eqs. (6)–(8) in the form

$$\frac{d}{dt}\mathbf{p}(t) = \frac{\mu c}{(1+c)}\mathbf{A}\mathbf{p}(t). \tag{9}$$

The notation used here is the following:

$$\mathbf{A} = \begin{pmatrix} -1 & \frac{1}{N} & 0 & 0 & 0 & \cdots & 0 \\ \frac{N}{N} & -1 & \frac{2}{N} & 0 & 0 & & \vdots \\ 0 & \frac{N-1}{N} & -1 & \frac{3}{N} & 0 & & \vdots \\ 0 & 0 & \frac{N-2}{N} & -1 & \ddots & 0 & 0 \\ \vdots & & & \ddots & \ddots & \frac{N-1}{N} & 0 \\ \vdots & & & 0 & \frac{2}{N} & -1 & \frac{N}{N} \\ 0 & \cdots & \cdots & 0 & 0 & \frac{1}{N} & -1 \end{pmatrix},$$

$$\mathbf{p}(t) = \begin{pmatrix} p_0(t) \\ p_1(t) \\ p_2(t) \\ \vdots \\ \vdots \\ p_{N-1}(t) \\ p_N(t) \end{pmatrix}.$$

The aim is to write the vector $\mathbf{p}(t)$ as a linear combination of eigenvectors using the following decomposition:

$$\mathbf{p}(t) = \sum_n a_n \boldsymbol{\varepsilon}^{(n)} e^{-r^{(n)}t}. \tag{10}$$

The sum runs over all of the eigenvectors $\boldsymbol{\varepsilon}^{(n)}$ and the constants a_n are determined from the initial conditions. Substituting the form $\boldsymbol{\varepsilon}e^{-rt}$ into Eq. (9) gives the secular equation

$$\det[\mathbf{A}' - \lambda\mathbf{I}] = 0. \tag{11}$$

The eigenvalues for this particular equation are related to the $r^{(n)}$ by the following equation:

$$r^{(n)} = \frac{\mu c}{(1+c)}\left[1 - \frac{\lambda^{(n)}}{N}\right]. \tag{12}$$

The matrix $\mathbf{A}'$ has the form

$$\mathbf{A}' = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ N & 0 & 2 & 0 & 0 & & \vdots \\ 0 & N-1 & 0 & 3 & 0 & & \vdots \\ 0 & 0 & N-2 & 0 & \ddots & 0 & 0 \\ \vdots & & & \ddots & \ddots & N-1 & 0 \\ \vdots & & & 0 & 2 & 0 & N \\ 0 & \cdots & \cdots & 0 & 0 & 1 & 0 \end{pmatrix}, \tag{13}$$

which can be shown to have the following characteristic equation:

$$\lambda \prod_{n=2,4,6,\ldots}^{N} [\lambda^2 - n^2] = 0, \tag{14}$$

when N is even, and

$$\prod_{n=1,3,5,\ldots}^{N} [\lambda^2 - n^2] = 0, \tag{15}$$

when N is odd. The eigenvalues are then

$$\lambda^{(\pm n)} = 0, \pm 2, \pm 4, \pm 6, \ldots, \pm N \tag{16}$$

for even N, and

$$\lambda^{(\pm n)} = \pm 1, \pm 3, \pm 5, \ldots, \pm N \tag{17}$$

for odd N.

It is clear that the eigenvalue corresponding to the eigenvector representing the system at equilibrium is given by $\lambda^{(+N)} = +N$, as this leads to $r^{(+N)} = 0$ by Eq. (12). The equilibrium eigenvector has elements proportional to the binomial coefficients C_k^N, as can be shown directly from Eqs. (6)–(8) when the left hand side of each is zero using mathematical induction. Note that k is again the Hamming distance between strings and the environmental string, as well as the row index for the eigenvector.

The binomial form of the equilibrium eigenvector is expected because all strings should occur with the same frequency when the system reaches the equilibrium state. No particular string has an advantage over any other as they all have the same number of offspring. A uniform distribution of strings means that the fraction having a Hamming distance k is simply the total number of strings that can exist having that Hamming distance, given by the combination C_k^N, divided by the number of possible strings 2^N.

The remaining eigenvectors can all be calculated by substituting the eigenvalues (16) and (17) into Eq. (11) to give the general distribution. As an illustration, consider the case of four-bit strings ($N=4$). The general solution is

$$\mathbf{p}(t) = a_{+4}\begin{pmatrix} 1 \\ 4 \\ 6 \\ 4 \\ 1 \end{pmatrix} + a_{+2}\begin{pmatrix} 1 \\ 2 \\ 0 \\ -2 \\ -1 \end{pmatrix} e^{-r^{(+2)}t} + a_0\begin{pmatrix} 1 \\ 0 \\ -2 \\ 0 \\ 1 \end{pmatrix} e^{-r^{(0)}t} + a_{-2}\begin{pmatrix} 1 \\ -2 \\ 0 \\ 2 \\ -1 \end{pmatrix} e^{-r^{(-2)}t} + a_{-4}\begin{pmatrix} 1 \\ -4 \\ 6 \\ -4 \\ 1 \end{pmatrix} e^{-r^{(-4)}t}. \tag{18}$$

The eigenvectors have been ordered from left to right with an increasing magnitude of $r^{(n)}$. The leftmost eigenvector represents the equilibrium state where $r^{(+4)}=0$. As the system approaches equilibrium the second eigenvector will tend to dominate the dynamics, unless its coefficient is zero (i.e., $a_{+2}=0$).

The coefficients a_n may be determined from the initial distribution of strings $\mathbf{p}(0)$ by setting $t=0$ in Eq. (18) and then solving the resulting system of linear equations. For example, if one begins with identical strings, all with Hamming distance $k=4$, the resulting coefficients are $a_{+4}=1/16$, $a_{+2}=-1/4$, $a_0=3/8$, $a_{-2}=-1/4$, and $a_{-4}=1/16$. It is worth noting that the coefficient of the equilibrium eigenvector a_{+4} must always be equal to $1/16$, independent of the initial conditions, in order to give the correct equilibrium distribution. For a general N-bit string model, the coefficient of the equilibrium state eigenvector will be $a_{+N}=2^{-N}$ when the eigenvector is written with its elements equal to the binary coefficients.

Figure 1 shows that the analytical solution is in excellent agreement with simulation data. Furthermore, the data resulting from many averaged runs are found to be independent of population size, although the behavior of very small populations is too erratic to obtain sufficiently accurate results. In the language of statistical mechanics, the simulation is self-averaging.

CASE II: MONOTONICALLY INCREASING FITNESS LANDSCAPE

Selection is imposed by changing the fitness landscape so that strings having a small Hamming distance from the environment string are favored. The number of offspring created from a parent string is now obtained from the exponential form of the fitness landscape, Eq. (1). For the four-bit strings examined in this paper, the average number of offspring varies from 3.0 for parents with a Hamming distance 0 to a value of 0.4 for parents with a Hamming distance of 4, as previously explained.

Figures 2(a)–2(d) show how the probability distribution behaves for four different mutation rates. The discrete points are the simulation data averaged over several runs. The solid

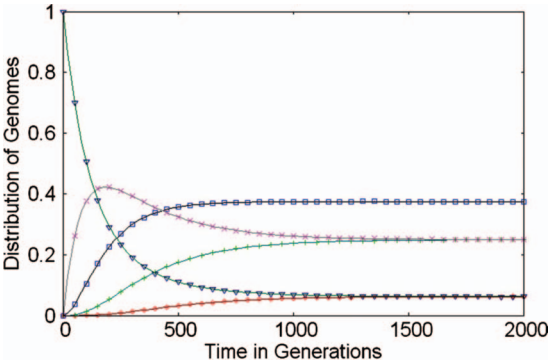

FIG. 1. (Color) Distribution of genomes for a flat fitness landscape. The points show simulation data for a population of $M=3600$ four-bit strings, where the initial conditions are set with all strings having a Hamming distance equal to 4. The mutation rate is set to $\mu=0.01$ and the number of offspring produced by each parent is $c=3$. The data represent an average over 10^4 independent runs. The solid lines are the results predicted from Eq. (18) and the agreement is excellent. Key: $\diamond$, $p_0(t)$; $+$, $p_1(t)$; $\sqcup$, $p_2(t)$; $\times$, $p_3(t)$; and $\triangledown$, $p_4(t)$.

lines are the results obtained from numerically iterating the recurrence relations (3)–(5). Lines are used to make a distinction between recurrence relations and simulations, although it should be understood that the recurrence relations also give one discrete point per generation. For the highly nonbiological limit of a mutation rate $\mu=1.0$ (i.e., every offspring contains a mutation), there is good agreement between the simulation and the recurrence relations, as illustrated by Fig. 2(a).

As the mutation rate is decreased, the simulation data deviates from the recurrence relations and hence from quasispecies theory. Figures 2(b)–2(d) show an increasing discrepancy; especially in the last two plots where the mutation rates are small (0.01 and 0.001).

The disagreement between the recurrence relations and the simulation data is a result of the finite size of the population. Consider the data displayed in Fig. 3. Here, the quantity $p_1(t)$ (the probability that a string has a Hamming distance $k=1$) for a fixed mutation rate of $\mu=0.05$ is plotted for several population sizes. The curve without any data points is that given by the recurrence relation for $p_1(t)$. As the size of the system increases, $p_1(t)$ is seen to be collapsing to the values given by its recurrence relation. The recurrence relations are thus seen to represent the behavior of a system having a population large enough so that effects due to its finite size are negligible, regardless of the value of the mutation rate.

It is instructive to investigate the origin of this finite population effect. Consider the simulation data shown in Figs. 4(a)–4(d). In contrast to all of the data in the previous figures, each plot is the result of a single run. The plots have the same mutation rate $\mu=0.001$ and the same population size $M=3600$. The only difference between them is the sequence of pseudorandom numbers used to decide when mutations occurred and which members are preserved when the population is culled.

The form of the plots is roughly the same. They begin with identical initial populations where all strings have a

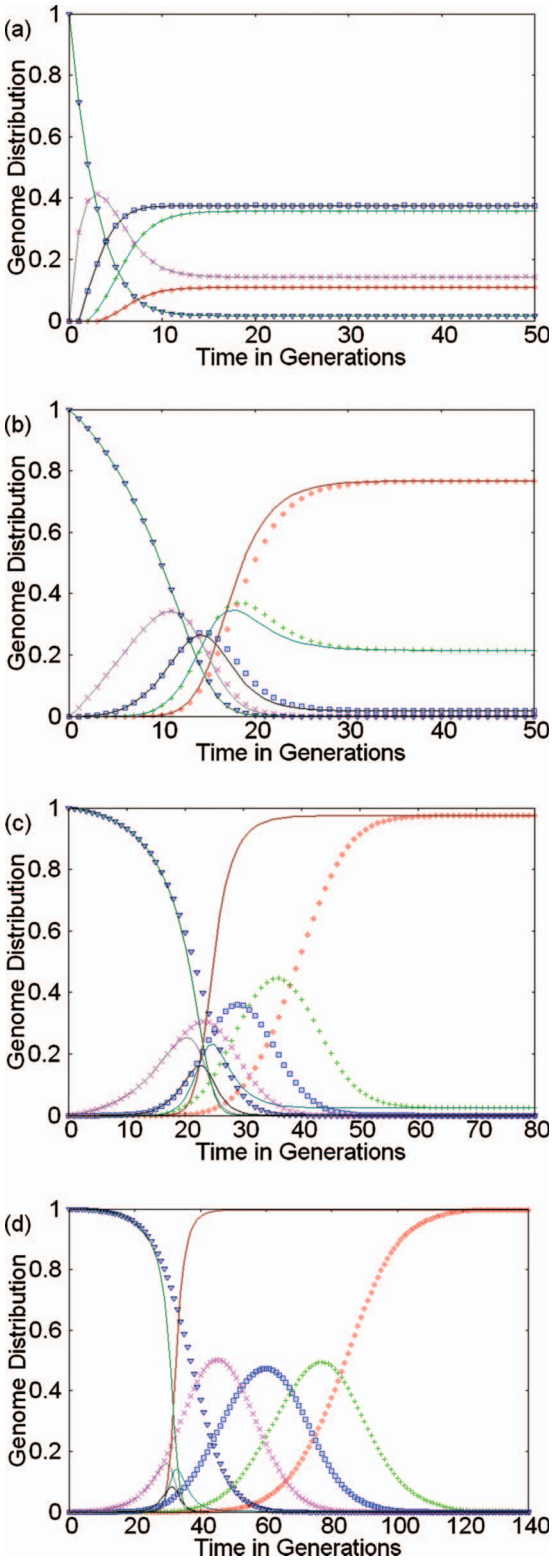

FIG. 2. (Color) Distribution of genomes for exponential fitness landscape. Points are simulation data averaged over 10^3 independent runs. Lines are theoretical curves from recurrence relations (3)–(5). Population size $M=3600$; mutation rate (a) $\mu=1.00$, (b) $\mu=0.10$, (c) $\mu=0.01$, and (d) $\mu=0.001$. Key: $\diamond$, $p_0(t)$; $+$, $p_1(t)$; $\sqcup$, $p_2(t)$; $\times$, $p_3(t)$; $\triangledown$, $p_4(t)$.

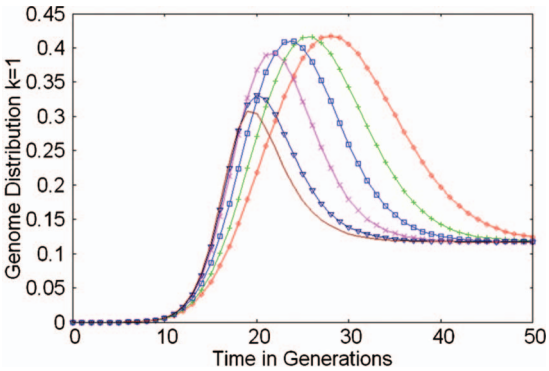

FIG. 3. (Color) Distribution of genomes with a Hamming distance $k=1$ $[p_1(t)]$ for several population sizes. Each curve represents an average of 10^3 independent runs. The curve without points is given by the recurrence equation (3) putting $k=1$. Mutation rate $\mu=0.05$. Key: $\diamond$, $M=180$; $+$, $M=360$; $\square$, $M=900$; $\times$, $M=3600$; ∇, $M=36000$.

Hamming distance equal to four (i.e., all strings are equal to "0000"). After a few generations, one or more mutant offspring are produced having a Hamming distance equal to 3 (i.e., either "0001," "0010," "0100," or "1000"). Any of these mutants that survive the culling process reproduce at a higher rate than any of the other strings. Eventually, after several more generations, the new genomes dominate the population. This produces the leftmost peak in each of the plots.

The next peak in the distribution plot occurs when a parent with a Hamming distance of 3 produces a mutant offspring having a Hamming distance of 2. The genome of this mutant then dominates the population, and so on until finally an "ideal" string with a Hamming distance of 0 arises. The population then reaches an equilibrium distribution where most strings will have a Hamming distance of 0 and a few will have larger Hamming distances as a result of occasional mutations. Unlike the dynamics, the equilibrium distribution is found to be independent of the system size and agrees with that obtained from the recurrence relations.

The details of the evolution for the populations displayed in Figs. 4(a)–4(d) are quite different even though their general form is similar. For example, the peaks where the genomes dominate are of different heights and widths and they also occur at different times. This is a consequence of the fact that relatively rare stochastic mutation events are controlling the dynamics. The frequency of these events depends on the mutation rate and the size of the population. Clearly, the evolution of the genome distribution for many typical runs like those shown in Figs. 4(a)–4(d) cannot yield an average behavior in agreement with that of quasispecies theory. The relatively small population size along with a small mutation rate will result in long periods where the population is dominated by genomes of intermediate fitness before the equilibrium distribution is reached.

The possible evolutionary paths that the dynamics can take are conveniently illustrated by the graph in Fig. 5. All 16 possible genomes are shown and those genomes that differ from one another by a single mutation are connected by a link. Genomes with the same Hamming distance form the

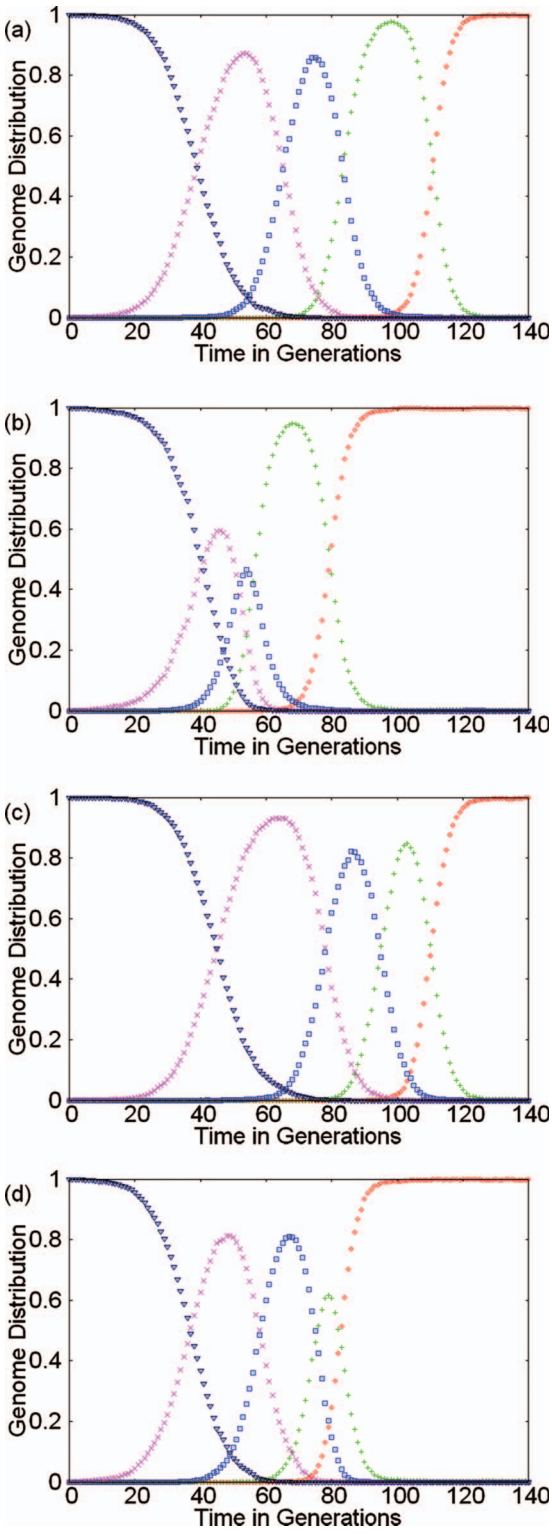

FIG. 4. (Color) Examples of the evolution of the distribution of genomes. Each peak represents a fitness epoch. Starting from the left-hand side, the first peak indicates the period when strings with $k=3$ dominate, the second where strings with $k=2$ dominate, and the third where strings with $k=1$ dominate. Each of the figures shows data for a single run using a different sequence of pseudorandom numbers but otherwise having the same parameters. Population size $M=3600$, mutation rate $\mu=0.001$. Key: $\diamond$, $p_0(t)$; $+$, $p_1(t)$; $\sqcup$, $p_2(t)$; $\times$, $p_3(t)$; ∇, $p_4(t)$.

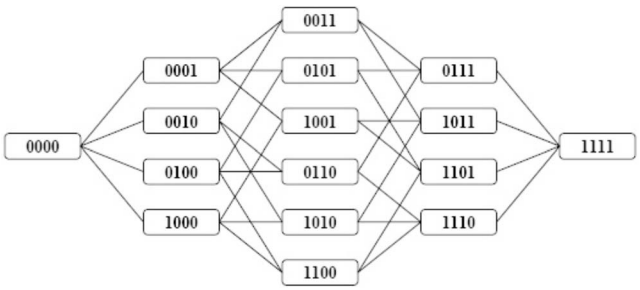

FIG. 5. All possible genomes for the four-bit model. Links exist between genomes that are connected by a single mutation, hence only strings having genomes differing by one bit can be parents of one another. Initial conditions have all strings with identical genomes "0000." As the simulation proceeds, mutant offspring are produced and genomes further to the right are favored. There are 24 (4!) possible paths to the string with the greatest fitness "1111." For low mutation rates some of these paths will dominate. This is called "evolutionary path fixing."

columns in Fig. 5. The evolution of the population can be seen as a progression from one where the leftmost genome ("0000") dominates to one where the rightmost genome ("1111") dominates.

The reader will note that those other evolutionary models based on quasispecies which allow multiple simultaneous mutations are not restricted to follow these paths. In these models it is possible, for example, to produce an ideal string "1111" from a parent string "0000" as a result of four mutations in the one offspring. Allowing multiple mutations facilitates analytical treatment but when the population is small and the mutation rate is low, they are extremely rare.

The transition from large population to small population behavior for fixed (but low) mutation rate will now be analyzed for the binary string model introduced in this paper. Let us first consider, from the standpoint of a simulation, the evolution of the model with a very large population, starting from an initial condition where all strings are identical and have a Hamming distance of 4. After one generation, a small but finite fraction of mutant offspring with a Hamming distance equal to 3 are present. Furthermore, all four genomes with $k=3$, shown in the second leftmost column in Fig. 5, are represented with equal frequencies. After two generations, a tiny fraction of strings with Hamming distance $k=2$ exist.

Following this line of reasoning, ideal strings with a Hamming distance of 0 are present in the population after only four generations. They do not immediately dominate the system however, as the strings with larger Hamming distance are much more numerous. The number of generations necessary to reach equilibrium is proportional to the logarithm of the inverse mutation rate, in accordance with Malthusian growth. In the intermediate generations, strings with larger Hamming distances undergo a brief increase in their frequencies and then quickly decrease.

One can visualize the dynamics as sweeping through the graph in Fig. 5 horizontally from left to right without any vertical variation, because all genomes with the same Hamming distance must occur with equal frequencies. The genome distribution is seen to follow a dynamics that is independent of the particular pseudorandom number sequence,

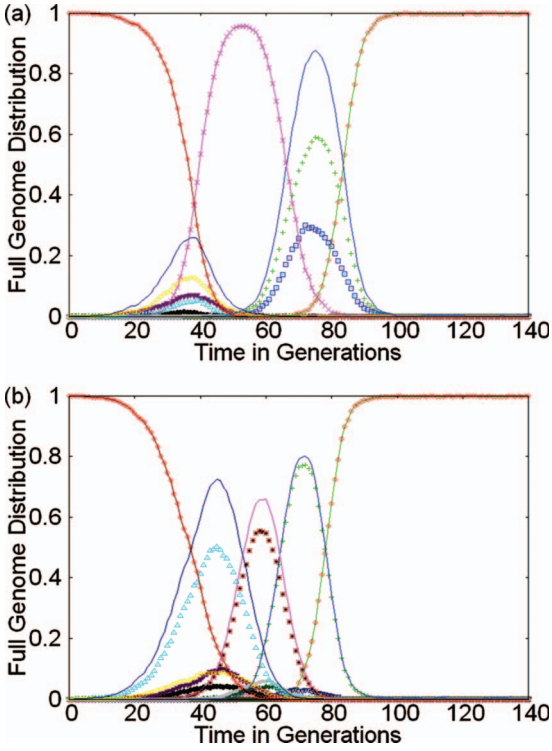

FIG. 6. (Color) Lines are the genome distribution written in terms of Hamming distance $p_k(t)$. Points represent all of the 16 possible genomes. Each $p_k(t)$ is represented by a line and is seen to be dominated to a certain degree by a particular genome, demonstrating evolutionary path fixing. Both figures are data from single runs having the same parameters but different sequences of pseudorandom numbers. Population size $M=3600$; mutation rate $\mu=0.001$.

which is consistent with the fact that the dynamics is governed by the quasispecies equations.

Consider now the behavior of a smaller population. In this case, the simulation data shows that the evolutionary paths in Fig. 5 are not all sampled evenly. Instead, certain paths dominate others and in doing so actively inhibit them. This phenomenon is the principal one that determines the dynamics and causes the model to deviate from quasispecies behavior. The present authors call this important effect "evolutionary path fixing" and it is quite possibly relevant to biological evolution.

The mechanism of evolutionary path fixing is twofold. First, the advent of a novel mutation produces a string with a higher reproductive rate that rapidly displaces members carrying inferior genomes in the population and thereby decreases the genetic diversity. The smallness of the population makes it highly probable that rare genomes disappear completely. In fact, the well-established counterpart to this effect observed in the evolution of microbial colonies is called "periodic selection" [25].

The second mechanism of evolutionary path fixing has some overlap with the first. In the process of displacing inferior members of the population, direct ancestors of the novel mutants are also made to go extinct. This means that other mutant genomes with the same fitness as the novel mutants are suppressed so that they are rare or never occur at

all. Referring again to Fig. 5, the result of this effect is to favor one member of each vertical column. This is in sharp contrast to quasispecies behavior where all members of a single column are equally represented in the population.

An extreme case of evolutionary path fixing can be visualized by considering the domination of the dynamics by a *single* path through the graph in Fig. 5. The question as to which of the paths is actually selected can only be answered through detailed knowledge of the sequence of pseudorandom numbers employed in a particular simulation. For a fixed population size, it is found that there is a smooth transition from all paths to a single path as the mutation rate is decreased.

Evolutionary path fixing through the random dominance of particular genomes is directly observable from the simulation data. Figures 6(a) and 6(b) show the data from two independent simulation runs. The individual genomes are shown with points and the distribution $p_k(t)$ with lines. Here, the sum of the individual genome plots with a Hamming distance k must add up to give $p_k(t)$. Of course, $p_4(t)$ and $p_0(t)$ are identical to the plots for the genomes "0000" and "1111," respectively, as they are the only ones contributing to the Hamming distances $k=4$ and $k=0$.

In Fig. 6(a) one can see that $p_2(t)$ is clearly dominated by a single genome out of the six possible genomes with $k=2$. On the other hand, $p_1(t)$ has two out of the possible four genomes dominating it, with one of them having roughly twice the frequency of the other. Finally, $p_3(t)$ has all four genomes with $k=3$ present. However, the frequency of the most successful one is twice times that of its nearest competitor. Figure 6(b) shows a similar pattern where a single genome dominates each peak.

DISCUSSION

It has been shown that finite population size has a profound effect on evolutionary dynamics when there is selection pressure. Fitness epochs occur that are markedly different from the behavior predicted by quasispecies theory. However, a system with a flat fitness landscape where there is no selection pressure has a dynamics that is independent of population size.

The preceding analysis of the binary string model serves to clarify some aspects of evolutionary modeling and the connection to microbial ecology. In a minimalist sense the model captures the essential features of a microbial population, exhibiting evolutionary path fixing which includes the biological mechanism of periodic selection. In addition, the equilibrium state of the model resembles a "climax community" in microbial ecology, where the relative frequencies of species do not change for long periods of time. One can also study the connection between a climax community and equilibrium statistical mechanics. This connection has already been partially explored [26].

Much of the behavior of the binary string model is similar to more complicated models such as the AVIDA platform and the Royal Road genetic algorithm. Features such as fitness epochs are exhibited by all these models and are presumably caused by the same evolutionary path fixing mechanism.

The inherent simplicity of the model so far does not allow for the consideration of the interactions between species essential to microbial populations. Also, modifications of the environment by the population are not within the scope of the model as it currently stands. Finally, important cooperative interactions between similar organisms are neglected, not the least of which is the exchange of genetic material through recombination. A large amount of work remains to be done in these areas.

ACKNOWLEDGMENTS

The authors would like to thank Professor Marc Lavoie (UWI), Dr. Smail Mahdi (UWI), Professor Dietrich Stauffer (Cologne), and Professor Craig Sargent (UKY) for helpful comments and suggestions. One of us (T. R.) wishes to thank Professor Peter Poole of St. Francis Xavier University, Nova Scotia for support under an NSERC grant. This work was done in part while one of us (L.M.) was at the Stanford University Department of Psychology on a visiting scholar appointment.

[1] John Maynard Smith and Eors Szathmary, *The Major Transitions in Evolution* (Oxford University Press, New York, 1995).
[2] D. Gillespie, J. Comput. Phys. **22**, 403 (1976).
[3] L. Demetrius, J. Stat. Phys. **30**, 3 (1983).
[4] A. Kowald and L. Demetrius, Proc. R. Soc. London, Ser. B **272**, 741 (2005).
[5] T. J. P. Penna, S. M. de Oliveira, and D. Stauffer, Phys. Rev. E **52**, R3309 (1995).
[6] B. F. De Blasio and F. V. De Blasio, Phys. Rev. E **72**, 031916 (2005).
[7] T. S. Ray, J. Stat. Phys. **74**, 929 (1994).
[8] P. R. A. Campos and J. F. Fontanari, Phys. Rev. E **58**, 2664 (1998).
[9] S. Wright, Genetics **16**, 97 (1931).
[10] R. A. Fisher, Proc. R. Soc. Edinburgh **42**, 321 (1922).
[11] J. R. Norris, *Markov Chains* (Cambridge University Press, New York, 1997).
[12] C. Adami, C. Ofria, and T. C. Collier, Proc. Natl. Acad. Sci. U.S.A. **97**, 4463 (2000).
[13] Christoph Adami, *Artificial Life* (Springer-Verlag, New York, 1998).
[14] E. van Nimwegen, J. P. Crutchfield, and M. Mitchell, Theor. Comput. Sci. **229**, 41 (1999).
[15] E. van Nimwegen, J. P. Crutchfield, and M. Mitchell, Phys. Lett. A **229**, 144 (1997).
[16] M. Eigen, Naturwiss. **58**, 456 (1971).

[17] M. Eigen and P. Schuster, Naturwiss. **64**, 541 (1977).
[18] C. O. Wilke, BMC Evol. Biol. **5**, 44 (2005).
[19] Y. C. Zhang, Phys. Rev. E **55**, R3817 (1997).
[20] S. J. Gould and N. Eldredge, Paleobiology **3**, 115 (1977).
[21] J. H. Gillespie, Am. Nat. **121**, 691 (1983).
[22] A. Traulsen, Y. Iwasa, and M. Nowak, J. Theor. Biol. **249**, 617 (2007).
[23] D. Alves and J. Fontanari, Phys. Rev. E **57**, 7008 (1998).
[24] See, for example, William E. Boyce and Richard C. DiPrima, *Elementary Differential Equations and Boundary Value Problems*, 3rd ed. (John Wiley & Sons, New York, 1977).
[25] Ronald M. Atlas and Richard Bartha, *Microbial Ecology* (Benjamin-Cummings Publishing, Menlo Park, CA, 1998).
[26] T. S. Ray, Int. J. Mod. Phys. C (to be published).