Population genetics approach to the quasispecies model

D. Alves and J. F. Fontanari
Instituto de Física de São Carlos, Universidade de São Paulo, Caixa Postal 369, 13560-970 São Carlos, São Paulo, Brazil
(Received 28 May 1996)

A population genetics formulation of Eigen's molecular quasispecies model [Naturwissenchaften **58**, 465 (1971)] is proposed and several simple replication landscapes are investigated analytically. Our results show a remarkable similarity to those obtained with the original kinetics formulation of the quasispecies model. However, due to the simplicity of our approach, the space of the parameters that define the model can be thoroughly explored. In particular, for the single-sharp-peak landscape our analysis yields some interesting predictions such as the existence of a maximum peak height and a minimum molecule length for the onset of the error threshold transition. [S1063-651X(96)00410-2]

PACS number(s): 87.10.+e, 64.60.Cn

I. INTRODUCTION

The molecular quasispecies model introduced by Eigen more than 20 years ago [1] has become a major framework of the research on the dynamics of competing self-reproducing macromolecules (see [2] for a review). In this model, a molecule is represented by a string of ν digits $(s_1, s_2, \ldots, s_\nu)$, with the variables s_i allowed to take on κ different values, each representing a different type of monomer used to build the molecule. The number of different types of molecules is thus κ^ν. The concentrations x_i of molecules of type $i = 1,2, \ldots, \kappa^\nu$ evolve in time according to the differential equations

$$\frac{dx_i}{dt} = \sum_j W_{ij} x_j - [D_i + \Phi(t)] x_i , \tag{1}$$

where the constants D_i stand for the death probability of molecules of type i and $\Phi(t)$ is a dilution flux that keeps the total concentration constant. The elements of the replication matrix W_{ij} depend on the replication rate A_i of molecules of type i as well as on the Hamming distance $d(i,j)$ between strings i and j. They are given by

$$W_{ii} = A_i q^\nu \tag{2}$$

and

$$W_{ij} = \frac{A_i}{(\kappa - 1)^{d(i,j)}} q^{\nu - d(i,j)} (1-q)^{d(i,j)}, \quad i \neq j, \tag{3}$$

where $0 \leqslant q \leqslant 1$ is the single-digit replication accuracy, which is assumed to be the same for all digits. Perhaps the main outcome of the quasispecies model is that, for a given replication accuracy, there exists a maximum string length that selection can maintain. This phenomenon, termed the error threshold, poses a serious difficulty in envisioning life as an emergent property of systems of competing self-replicating macromolecules. It seems that some sort of cooperation between the macromolecules must be incorporated in the model in order to avoid this error catastrophe [3,4].

In this paper we employ a classic population genetics approach [5] to investigate the evolution of an infinite population of self-replicating molecules. To accomplish that we have to make two simplifying assumptions to the original quasispecies model. First, we assume that molecules composed of the same number of monomers of each type are equivalent, i.e., possess the same replication rate, regardless of the particular positions of the monomers inside the molecules. Hence a given molecule is characterized solely by the vector $\vec{P} = (P_1, P_2, \ldots, P_\kappa)$, where P_α is the number of monomers of type α in that molecule. Since $\Sigma_\alpha^\kappa P_\alpha = \nu$, the number of different types of molecules is reduced to $(\nu + \kappa - 1)!/\nu!(\kappa - 1)!$. Second, in the population genetics approach we focus on the evolution of the monomer frequencies rather than on the evolution of the molecule frequencies or concentrations. Henceforth the variable t will denote the number of nonoverlapping generations or simply the generation number. We assume then that, given the monomer frequencies in generation t, $p_\alpha(t)$ with $\Sigma_\alpha^\kappa p_\alpha(t) = 1$, the molecule frequencies are given by the multinomial distribution

$$\Pi_t(\vec{P}) = C_{\vec{P}}^\nu [p_1(t)]^{P_1} [p_2(t)]^{P_2} \cdots [p_\kappa(t)]^{P_\kappa}, \tag{4}$$

where $C_{\vec{P}}^\nu = \nu!/P_1!P_2! \cdots P_\kappa!$. Thus, in each generation the monomers are sampled with replacement from an infinite monomer pool. The effects of random drift are neglectable because the population of molecules is also infinite. The changes in the monomer frequencies are due then to the driving of natural selection, modeled by the replication rate $A(\vec{P})$, and to mutations, modeled by the error rate per digit $1-q$. A similar assumption was employed in the analytical study of the effects of learning on evolution [6]. With these assumptions we are able to study analytically the dynamical behavior of the model in the full space of the control parameters ν, q, κ and replication landscapes $A(\vec{P})$. In particular, while previous investigations [2] have almost exclusively dealt with binary strings ($\kappa = 2$), our population genetics approach readily applies to the analysis of more complex strings.

A result worth mentioning concerning the quasispecies model is the existence of a correspondence between the ordinary differential equations (1) and the equilibrium properties of a surface lattice systems [7]. However, from an operational viewpoint, it seems easier to solve directly the

ordinary differential equations than to use cumbersome statistical mechanics tools to obtain the surface equilibrium properties of the corresponding lattice system for finite ν [8]. Actually, most of the statistical mechanics analyses of the quasispecies model are restricted to the limit $\nu\to\infty$ [7–9].

The remainder of this paper is organized as follows. In Sec. II we derive the equations governing the evolution of the monomer frequencies. To better appreciate the consequences of our simplifying assumptions, these equations are solved for several simple replication landscapes that have already been thoroughly analyzed in the literature [2,8]. In Sec. III we discuss our results and present some concluding remarks. In particular, we point out how the model can be generalized so as to include sexual reproduction between the molecules.

II. MODEL

We now proceed with the derivation of the basic recursion relations for the monomer frequencies. The fraction of monomers of type α that a molecule characterized by $\vec{P}$ contributes to generation $t+1$ is proportional to the product of three factors: (a) its frequency in the population $\Pi_t(\vec{P})$ in generation t, (b) its replication rate $A(\vec{P})$, and (c) the average number of monomers α that replicate correctly, qP_α, plus the average number of monomers $\beta\neq\alpha$ that due to replication errors mutate to α, $[(1-q)/(\kappa-1)]\Sigma_{\beta\neq\alpha}P_\beta$. After some simple algebra it yields

$$p_\alpha(t+1)=\frac{1}{\kappa-1}\left(1-q+\frac{\kappa q-1}{w_t}\sum_{\vec{P}}\Pi_t(\vec{P})A(\vec{P})P_\alpha\right), \tag{5}$$

where the normalization factor w_t is the average replication rate of the entire population in generation t,

$$w_t=\nu\sum_{\vec{P}}\Pi_t(\vec{P})A(\vec{P}). \tag{6}$$

Here the notation $\Sigma_{\vec{P}}$ stands for $\Sigma_{P_1=0}^{\nu}\cdots\Sigma_{P_\kappa=0}^{\nu}\delta(\nu,\Sigma_\alpha^\kappa P_\alpha)$, where $\delta(k,l)$ is the Kronecker delta. To proceed further we must specify the replication rate $A(\vec{P})$ of each molecule type, i.e., specify the replication landscape.

A. Single sharp maximum

In this case we ascribe replication rate a to the molecule composed of ν monomers of type 1 and replication rate 1 to all the remaining molecules, i.e., $A(\vec{P})=a$ if $\vec{P}=(\nu,0,\ldots,0)$ and $A(\vec{P})=1$ otherwise. This is the simplest and probably the most studied replication landscape [2] because it clearly shows that although the so-called master string $(\nu,0,\ldots,0)$ has no match, its chance of successfully taking over the population depends nontrivially on the values of the control parameters as well as on the initial monomer frequencies $p_\alpha(0)$. Hence Eq. (5) reduces to

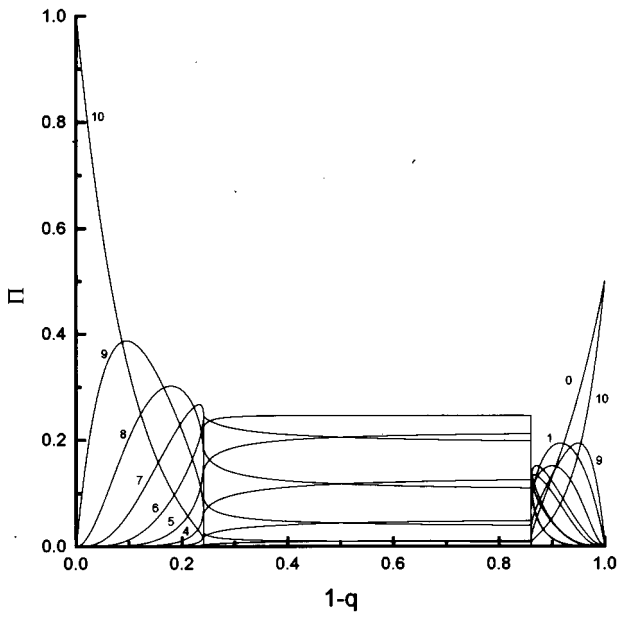

FIG. 1. Steady-state frequencies of molecules of type $P_1=10$ (master string), 9, 8, 7, 6, 5, 4, and 0 as a function of the error rate per digit $1-q$ for $\nu=10$, $a=50$, and $\kappa=2$. The error threshold transition occurs at $1-q_t\approx0.241$ and the complementary replication regime sets in for $1-q>0.86$.

$$p_1(t+1)=\frac{1}{\kappa-1}\left[1-q+(\kappa q-1)\frac{p_1(t)+(a-1)[p_1(t)]^\nu}{1+(a-1)[p_1(t)]^\nu}\right] \tag{7}$$

and

$$p_\alpha(t+1)=\frac{1}{\kappa-1}\left[1-q+(\kappa q-1)\frac{p_\alpha(t)}{1+(a-1)[p_1(t)]^\nu}\right],\qquad \alpha\neq1. \tag{8}$$

For simplicity, we keep the symmetry between monomers of type $\alpha\neq1$ by setting their initial frequencies to $p_\alpha(0)=[1-p_1(0)]/(\kappa-1)$. Furthermore, we bias the initial population towards the master string by choosing $p_1(0)\approx1$.

In Fig. 1 we present the steady-state molecule frequencies, obtained by solving the recursion relation (7) for $\nu=10$, $a=50$, and $\kappa=2$, as a function of the error rate per digit $1-q$. Three distinct regimes can be identified. First, the direct replication regime (DR), which occurs for $1-q\leqslant1-q_t\approx0.241$, is characterized by a molecular population composed of a cloud of mutants around the master string, termed a *quasispecies*. In this regime there is a high proportion of type 1 monomers, i.e., the fixed point is $p_1^*\approx1$. This fixed point disappears discontinuously at the error threshold $1-q_t$. We note that, in contrast to the original quasispecies model, the error threshold transition is discontinuous. Second, the stochastic replication regime (SR), which occurs for $1-q>1-q_t$, is characterized by the fixed point $p_1^*\approx1/2$, which corresponds to an almost uniform distribution of monomer types. Third, the complementary replication regime (CR) sets in when the replication error is so high $(1-q>0.86)$ that the monomers are almost certain to

mutate, so that the population oscillates between the quasispecies and its complement. This regime exists only for binary strings since only in this case is the complement of a string unique.

Some comments regarding the role of the initial monomer frequencies $p_\alpha(0)$ are in order. For $1-q<0.239$ the high-p_1 fixed point is the only stable fixed point of (7). Above that value, a second stable fixed point $p_1^*\approx 1/2$ appears. These fixed points compete such that there is an all-or-none selection. The winner, however, is not determined by the replication rate only, but also by the initial monomer frequency $p_1(0)$. As mentioned above, the high-p_1 fixed point disappears at the error threshold transition. We will return to this issue in the analysis of the competition between two sharp maxima. The behavior pattern of the molecule frequencies for $\kappa>2$ is qualitatively similar to that discussed above: the DR and SR regimes are characterized by the fixed points $p_1^*\approx 1$ and $p_1^*\approx 1/\kappa$, respectively, while the CR regime is absent.

In the following we will focus on the dependence of the error threshold $1-q_t$ on the control parameters. The fixed points $p_1(t+1)=p_1(t)=p_1^*$ of the recursion relation (7) are the roots of $f(p)=0$, where

$$f(p)=(1-q)[\kappa p-1+(\kappa-1)(a-1)p^\nu] -(a-1)(1-p)p^\nu. \tag{9}$$

Numerical analysis of this function indicates that the discontinuous disappearance of the fixed point $p_1^*\approx 1$, which is the cause of the error threshold phenomenon in our model, coincides with the appearance of a double root of $f(p)$. Hence the error threshold $1-q_t$ can be easily determined by solving $f(p)=0$ and $df(p)/dp=0$ for p and $q=q_t$ simultaneously. Eliminating the term p^L of these two equations yields the following quadratic equation for p:

$$\kappa\nu p^2-[1+\nu+q\kappa(\nu-1)]p+q\nu=0, \tag{10}$$

which possesses real roots for either $q\kappa\leqslant 1$ or $q\kappa\geqslant[(\nu+1)/(\nu-1)]^2$. Only the latter is relevant for the analysis of the error threshold since this phenomenon occurs in the high replication accuracy region. The disappearance of the high-p_1 fixed point is associated with the larger root of (10), while the smaller root is related to the appearance of the uniform fixed point $p_1^*\approx 1/\kappa$. In order to avoid the error threshold discontinuous transition we must set the control parameters so as to violate the second inequality. In particular, for ν and κ fixed, the discontinuous transition line $q_t=q_t(a)$ terminates at the critical point

$$q_c=\frac{1}{\kappa}\left(\frac{\nu+1}{\nu-1}\right)^2, \tag{11}$$

$$a_c=1+\kappa^\nu\left(\frac{\nu-1}{\nu+1}\right)^\nu\frac{(\kappa-1)(\nu-1)^2-4\nu}{(\kappa-1)(\nu^2-1)}, \tag{12}$$

which for large ν become $q_c\approx 1/\kappa$ and $a_c\approx e^{-2}\kappa^\nu$, respectively. We note that the critical point coordinates p_c, q_c and a_c are obtained by solving the three equations $f(p)=0$, $df(p)/dp=0$, and $d^2f(p)/dp^2=0$ simultaneously. The condition $q_c\leqslant 1$ implies that there is a minimum string length

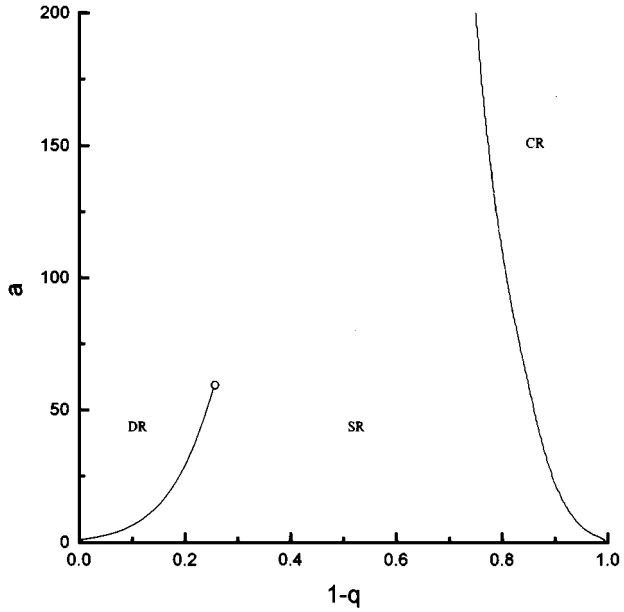

FIG. 2. Phase diagram in the space $(1-q,a)$ for $\nu=10$ and $\kappa=2$. The discontinuous transition between the phases DR and SR ends at the critical point $1-q_c=0.251$ and $a_c=58.01$.

$$\nu_{\min}=\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}, \tag{13}$$

below which the error catastrophe does not occur. In Fig. 2 we present the phase diagram in the space $(1-q,a)$ for $\nu=10$ and $\kappa=2$. The discontinuous transition between the phases DR and SR ends at the critical point $1-q_c=0.251$ and $a_c=58.01$, while the transition between the phases SR and CR seems to never disappear. Thus, for a given value of ν it is always possible to choose a sufficiently large value of a so that the phases CR and SR are no longer distinguishable. To the best of our knowledge, there is no similar result reported for the original quasispecies model. It must be noted, however, that due to the numerical difficulty of solving the set of κ^ν ordinary differential equations (1), the space of parameters has not been adequately explored for that model. In fact, the computational effort needed to study the evolution of a population of molecules composed of more than two types of monomers ($\kappa>2$) is so large that the important problem of the dependence of the error threshold $1-q_t$ on the number of monomer types κ has remained unaddressed so far. In the population genetics framework, however, the number of recursion relations increases linearly with κ, so this parameter does not introduce any particular difficulty to our analysis. Moreover, for the replication landscapes considered in this paper, in which the replication rates of the molecules are determined by one type of monomer only, the problem is reduced to the solution of a single recursion relation. In Fig. 3 we present the dependence of $1-q_t$ on κ for $a=10$ and several values of ν. Note that beyond $\kappa\approx 4$ the error threshold is almost insensitive to further increase of κ. Hence, in order to maximize the information content of the quasispecies, it is advantageous to choose κ as large as possible. Different values of a do not produce any qualitative change in this figure.

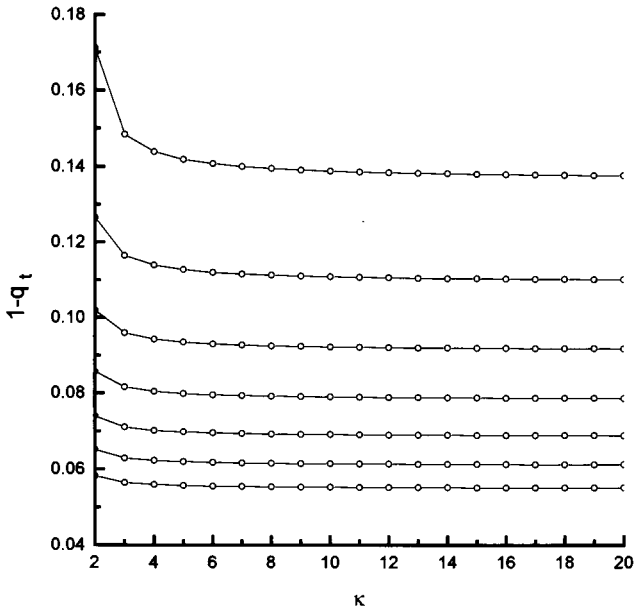

FIG. 3. Error threshold $1-q_t$ as a function of the number of monomer types κ for $a=10$ and (from top to bottom) $\nu=8$, 10, 12, 14, 16, 18, and 20. The lines are guides to the eyes.

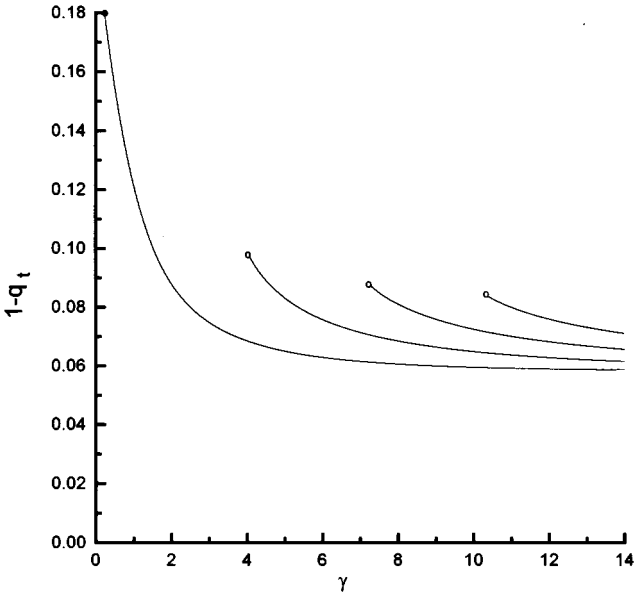

FIG. 4. Error threshold $1-q_t$ as a function of the exponent γ for $\nu=20$, $a=10$, and (from left to right) $\mu/\nu=0.75$, 0.5, 0.25, and 0. The curves begin at $\gamma_c=\gamma_c(\mu)$.

B. Single smooth maximum

In what follows we will consider the case $\kappa=2$ only. We assume that the replication rates of the molecules increase with the number of monomers of type 1 they possess, irrespective of the other monomer types. More specifically,

$$A(P_1)=1+(a-1)\left(\frac{P_1-\mu}{\nu-\mu}\right)^{\gamma} \tag{14}$$

if $P_1\geqslant\mu$ and $A(P_1)=1$ otherwise. Here μ is an integer that can take on the values $0,1,\ldots,\nu-1$ and γ is a real, positive variable. Clearly, μ measures the size of the flat region of the replication landscape, while the parameter γ determines the smoothness of the landscape near μ: the larger γ, the smoother the landscape. The same procedure employed in the analysis of the single-sharp maximum, which is recovered for $\mu=\nu-1$, can be used to investigate the error threshold transition for the smooth replication landscape (14). In particular, in Fig. 4 we show the error rate per digit at the threshold transition $1-q_t$ as a function of the exponent γ for $\nu=20$, $a=10$, and several values of μ. For a given μ there is a critical value γ_c below which the error threshold phenomenon does not occur. This exponent is shown in Fig. 5 as a function of the ratio μ/ν. It is clear from these figures that broad maxima (small μ) can resist the error catastrophe longer, or even avoid it, depending on the value of the exponent γ. Large values of γ actually increase the size of the flat region and so they favor the appearance of the error threshold. Our results are in agreement with a comment by Tarazona [8] that the exponent with which the replication landscape goes flat is germane to the onset of the error threshold transition. Different values of ν and a do not change qualitatively these results.

C. Two sharp maxima

As before, we assume that $A(\vec{P})$ depends on P_1 only. In this case the replication landscape consists of two sharp maxima $A(P_1=0)=A_0$, $A(P_1=\nu)=A_\nu$, and $A(P_1)=1$ otherwise. In order to illustrate the role of the initial monomer frequencies we present in Figs. 6–8 the frequency of type 1 monomers as a function of the generation number t for $\nu=20$, $\kappa=2$, $A_0=200$, $A_{20}=10$, and several initial frequencies. The evolution for $1-q=0$ is shown in Fig. 6. There are only two stable fixed points, namely, $p_1^*=1$ and 0.

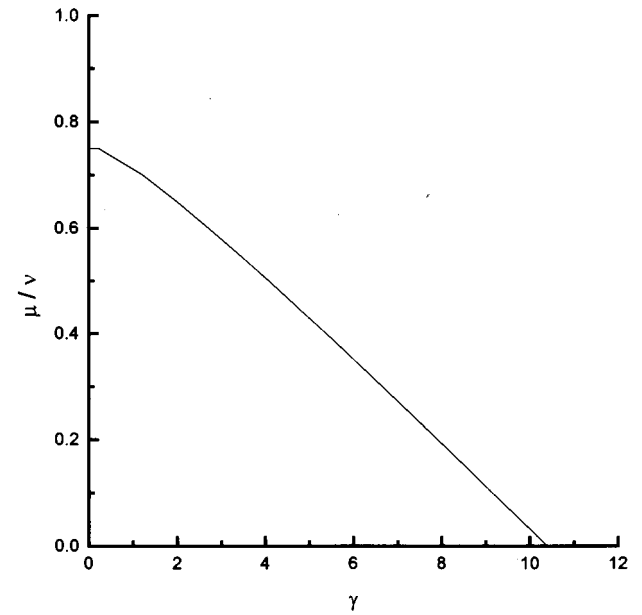

FIG. 5. Critical value of the exponent γ as a function of the ratio μ/ν for $\nu=20$ and $a=10$. Below the curve the error threshold phenomenon does not occur.

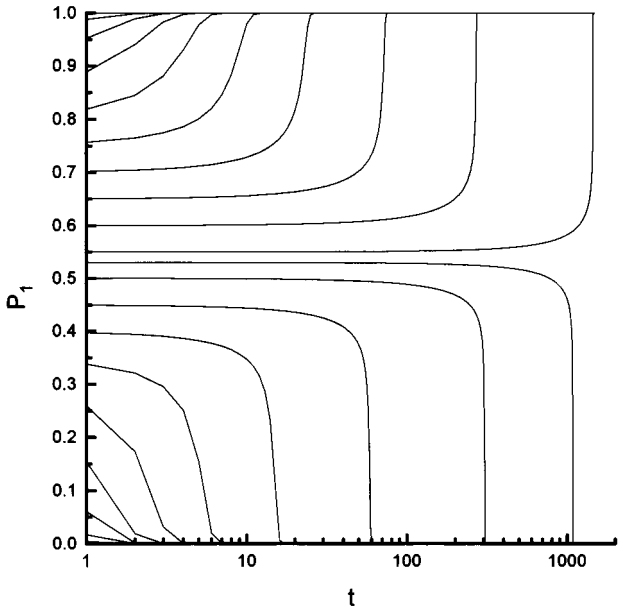

FIG. 6. Frequency of monomers of type 1 as a function of the generation number for the two-sharp-maxima replication landscape and several initial frequencies $p_1(0)$. The parameters are $\nu=20$, $\kappa=2$, $A_0=200$, $A_{20}=10$, and $1-q=0$.

Despite the large difference between the replication rates of the molecules associated to these fixed points, their basins of attraction are practically of the same size. They would be strictly equal if $A_0=A_{20}$. The main effect of a large replication rate in this case is to speed up the convergence to the low-p_1 fixed point. By increasing the error rate a new stable fixed point $p_1^*\approx 1/2$, associated with the stochastic replication regime, appears. The interplay of the three stable fixed points is shown in Fig. 7 for $1-q=0.01$. For nonzero replication error rates, the basin of attraction of the low-p_1 fixed point is considerably larger than that of the high-p_1 fixed point. Of course, their basins of attraction have actually decreased as compared with the case $1-q=0$. We note that the two quasispecies do no coexist: for a given initial population there is an all-or-none selection. Finally, in Fig. 8 we present the evolution for $1-q=0.06$. The high-p_1 fixed point, associated with the molecule with the smaller replication rate, has disappeared and the stochastic replication fixed point has taken over its basin of attraction. A further increase of the error rate $1-q$ will eventually lead to the disappearance of the low-p_1 fixed point too.

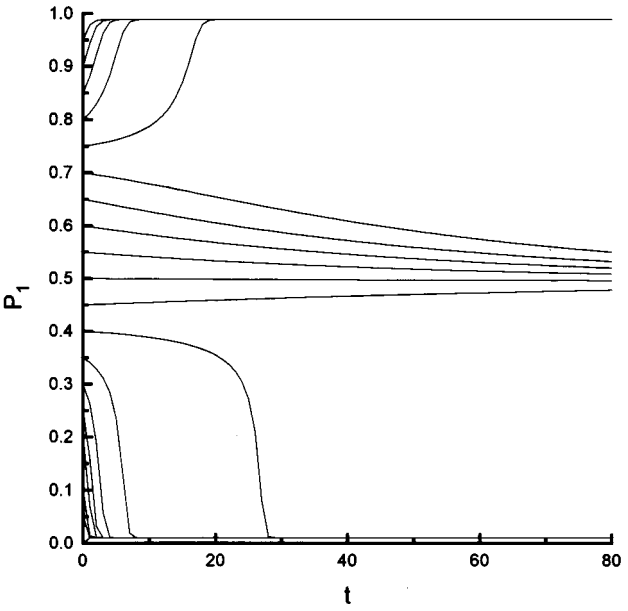

FIG. 7. Same as Fig. 6, but for $1-q=0.01$.

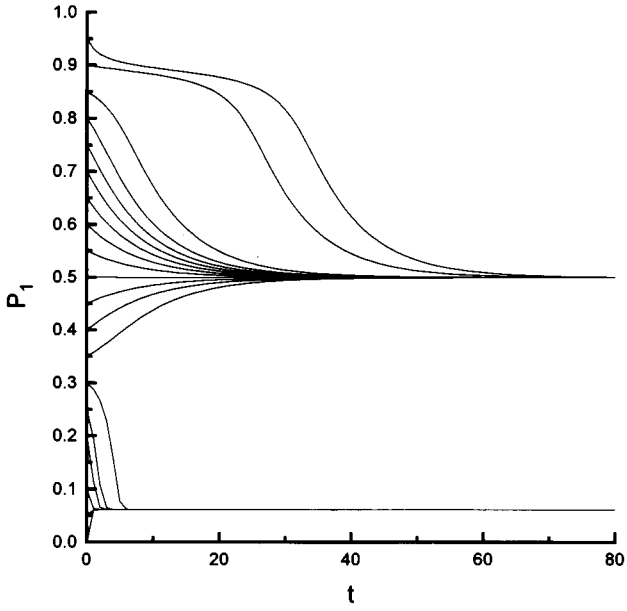

FIG. 8. Same as Fig. 6, but for $1-q=0.06$.

Within this framework we can easily study the competition between a sharp maximum and a broad or smooth maximum [10]. The results show the same qualitative features as those presented above. In particular, since the broader maximum possesses the larger error threshold $1-q_t$ (see Fig. 4) it plays the same role as the larger replication rate maximum. We note that, in contrast to the original quasispecies model, there is no selection transition in our model [10], which would amount to a discontinuous transition between the low- and the high-p_1 fixed points, i.e., the former should take over the basin of attraction of the latter.

III. DISCUSSION

An interesting extension of the quasispecies model is the possibility of two molecules exchanging matter during a collision. Clearly, the analog to this phenomenon in the population genetics approach is sexual reproduction. More specifically, the collision (mating) between the molecules (parents) $(s_1^f,\ldots,s_\nu^f)$ and $(s_1^m,\ldots,s_\nu^m)$ produces the new molecules (offspring) $(s_1^f,\ldots,s_{c-1}^f,s_c^m,\ldots,s_\nu^m)$ and $(s_1^m,\ldots,s_{c-1}^m,s_c^f,\ldots,s_\nu^f)$, where the digit $0\leqslant c\leqslant\nu$ is the so-called crossover point. The number of offspring depends, of course, on the replication rate of the parent molecules. Using the assumptions presented in Sec. I, it is straightforward to derive the following recursion relations for the evolution of the monomer frequencies:

$$p_\alpha(t+1)=\frac{1}{\kappa-1}\left[1-q+\frac{\kappa q-1}{2w_t}\sum_{\vec{P}^f}\sum_{\vec{P}^m}\Pi_t(\vec{P}^f,\vec{P}^m)\times A(\vec{P}^f,\vec{P}^m)(P^f_\alpha+P^m_\alpha)\right], \tag{15}$$

where

$$w_t=\nu\sum_{\vec{P}^f}\sum_{\vec{P}^m}\Pi_t(\vec{P}^f,\vec{P}^m)A(\vec{P}^f,\vec{P}^m) \tag{16}$$

is the average replication rate of the entire population and

$$\Pi_t(\vec{P}^f,\vec{P}^m)=\Pi_t(\vec{P}^f)\Pi_t(\vec{P}^m) \tag{17}$$

is the frequency of the collisions or matings between the molecules $\vec{P}^f$ and $\vec{P}^m$. Here $A(\vec{P}^f,\vec{P}^m)$ determines the number of offspring generated by the mating between these two molecules. As expected, since the positions of the monomers inside the molecules play no role in our population genetics approach, the basic recursion relations (15) are independent of the crossover point c. It is interesting to note that this equation reduces to Eq. (5) in the case that $A(\vec{P}^f,\vec{P}^m)=A(\vec{P}^f)A(\vec{P}^m)$ and so the two reproduction modes, asexual and sexual, yield the same results. We have investigated the steady-state solutions of (15) under a variety of conditions, but found no noteworthy difference from the previously presented results. We only mention that by penalizing matings within the same class, i.e., $A(\vec{P}^f,\vec{P}^m)=1$ if $\vec{P}^f=\vec{P}^m$, we can obtain the formation of a quasispecies (a master string surrounded by a cloud of mutants) even in the regime of perfect replication accuracy $q=1$.

The critical, though natural, assumption of the population genetics approach proposed in this paper is the use of the multinomial distribution (4) for the molecule frequencies. Since this is a single-peaked distribution, the coexistence of two or more quasispecies, which could only be described by a multipeaked distribution, is prevented *a priori*. In the original quasispecies model such a coexistence is possible only in the case of degenerate quasispecies. This is an important issue since it would be highly desirable to study the spontaneous formation of hypercycles within the framework of the quasispecies model [3,11].

IV. CONCLUSION

In this paper we have presented a population genetics formulation of the classic quasispecies model proposed by Eigen [1]. Owing to its extreme simplicity, this formulation may be useful, in the sense of having the value of an approximation, to tackle problems for which the numerical difficulty of solving the ordinary differential equations (1) or employing the statistical mechanics approach [7] makes the analysis prohibitive. Furthermore, even for the well-studied replication landscape that consists of a single sharp peak, our population genetics analysis has yielded some interesting and unexpected results such as the existence of a maximum peak height (12) and a minimum string length (13) for the onset of the error catastrophe. It would be interesting to investigate whether similar bounds exist for the original quasispecies model.

ACKNOWLEDGMENTS

This work was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

[1] M. Eigen, Naturwissenchaften **58**, 465 (1971).
[2] M. Eigen, J. McCaskill, and P. Schuster, J. Phys. Chem. **92**, 6881 (1988); Adv. Chem. Phys. **75**, 149 (1989).
[3] M. Eigen and P. Schuster, *The Hypercycle—A Principle of Natural Self-Organization* (Springer-Verlag, Berlin, 1979).
[4] S. A. Kauffman, *The Origins of Order* (Oxford University Press, Oxford, 1993).
[5] D. L. Hartl and A. G. Clark, *Principles of Population Genetics* (Sinauer, Sunderland, 1989).
[6] J. F. Fontanari and R. Meir, Complex Syst. **4**, 401 (1990).
[7] I. Leuthäusser, J. Chem. Phys. **84**, 1884 (1986); J. Stat. Phys. **48**, 343 (1987).
[8] P. Tarazona, Phys. Rev. A **45**, 6038 (1992).
[9] S. Franz, L. Peliti, and M. Sellitto, J. Phys. A **26**, L1195 (1993).
[10] P. Schuster and J. Swetina, Bull. Math. Biol. **50**, 635 (1988).
[11] I. R. Epstein, J. Theor. Biol. **78**, 271 (1979).