Error Thresholds for Quasispecies on Dynamic Fitness Landscapes

Martin Nilsson
Institute of Theoretical Physics, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden

Nigel Snoad
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501
and The Australian National University, ACT 0200, Australia[†]
(Received 29 March 1999)

In this paper we investigate error thresholds on dynamic fitness landscapes. We show that there exists both a lower and an upper threshold, representing limits to the copying fidelity of simple replicators. The lower bound can be expressed as a correction term to the error threshold present on a static landscape. The upper error threshold is a new limit that only exists on dynamic fitness landscapes. We also show that for long genomes and/or highly dynamic fitness landscapes there exists a lower bound on the selection pressure required for the effective selection of genomes with superior fitness independent of mutation rates, i.e. there are distinct nontrivial limits to evolutionary parameters in dynamic environments.

PACS numbers: 87.23.Kg, 87.10.+e, 87.15.Aa

Ever since Eigen's work on replicating molecules in 1971 [1], the concept of quasispecies has proven to be a very fruitful way of modeling the fundamental behavior of evolution. A quasispecies is an equilibrium distribution of closely related gene sequences, localized around one or a few sequences with high fitness (the master sequence, or wild type) [2]. The kinetics of these simple systems has been studied in great detail as this formulation has allowed many of the techniques of statistical physics to be applied to replicator and evolutionary systems. See for instance [1,3–12], and references therein.

The appearance in many of these models of an error threshold (or error catastrophe) as an upper bound on the mutation rate, above which no effective selection can occur, has important implications for biological systems. In particular, it places limits on maintainable amounts of genetic information [1,13,14], thus restricting possible theories for the origins of life. It is interesting to note that some RNA viruses seem to have evolved mutation rates that are close to the error threshold [14,15], and that in many cases the quasispecies concept seems to be a valuable descriptor of viral diversity [16].

Studies of quasispecies until now have focused on static fitness landscapes. Many organisms in nature, however, live in a quickly changing environment [17]. This is especially important for viruses and microbial pathogens that must survive in a host with a highly dynamic immune system for which there only exist tight and temporary niches with high fitness (for the pathogen) [18]. There is a body of extant work dealing with the mathematics of evolution in changing environments in the context of both population genetics models ([19–21], for example) and coevolutionary systems (particularly in ecological frameworks) ([22], for example). Neither of these approaches are particularly suitable for investigating the quickly changing multilocus systems that we examine. In particular, they give little indication of the kinds of dynamics we find and characterize below.

In brief, a quasispecies consists of a population of self-replicating genomes, each represented as a sequence of bases s_k, $(s_1 s_2 \cdots s_n)$. Hereafter we will assume binary bases $\{1, 0\}$ and that all sequences have equal length n. Every genome is thus a binary string $(011001\ldots)$, indexed by an integer k $(0 \leq k < 2^n)$.

To describe how mutations affect a population we define W_k^l as the probability that replication of genome l gives genome k as offspring. We will only consider point mutations which conserve the genome length and occur independent of position in the genome with rate $\mu = 1 - q$ (where q is the copying accuracy per base). The equations describing the dynamics of the population now take a relatively simple form [7]

$$\dot{x}_k = \sum_l W_k^l A_l x_l - f x_k \,, \tag{1}$$

where x_k is the relative concentration and A_k the fitness (replication rate) of genome k. $f = \sum_l A_l x_l$ ensures the total normalization of the population with $\sum_l x_l = 1$.

To create a dynamic landscape we modify the standard single-peaked fitness landscape [8] so that the fitness peak moves in genotype space, resulting in different optimal gene sequences at different times. Formally, we write $A_{k(t)} = \sigma$ and $A_l = 1\ \forall l \neq k(t)$, where the (changing) best genotype $k(t)$ describes how the peak (master sequence) moves through sequence space. In this paper the peak moves to one of its closest neighbors (chosen randomly) at regular time intervals τ. This model is inspired by analogy to the existence of a tight niche for an organism (perhaps viral), where changes to the genotypic composition of the niche are more important

than small modifications to the reproductive advantage the niche confers (the latter case would have $A_k(t)$; see [23]). One could also consider probabilistic movement, and to more distant neighbors.

The independence of mutation rate from genome position imposes a symmetry on the rate equations, enabling us to divide the relative concentrations into error classes Γ_i described by their Hamming distance i from the master sequence (Γ_0). This reduces the effective dimension of the sequence space from 2^n to $n + 1$, thereby making the problem significantly more tractable. The use of asymmetric evolution operators (such as recombination) or fitness landscapes is obviously more problematic and is the subject of ongoing work. When the fitness peak moves, this landscape symmetry will be broken: one sequence in Γ_1 will be singled out as the new master sequence. We assume the dynamics to be slow enough and the probability of the peak quickly returning to a previously optimal genotype is small enough for this to not be a problem, i.e., the (temporary) existence of a large concentration at the previous peak genotype will have little effect on initial concentrations for subsequent moves.

On average then, moving the fitness peak corresponds to applying the following coordinate transformation to the concentration vector (written in terms of error classes i and j):

$$R_{ij} = \frac{n - i}{n}\,\delta_{i,j-1} + \frac{i}{n}\,\delta_{i,j+1}\,. \tag{2}$$

To study the population evolution we divide the dynamics into cycles of length τ, the time between shifts of the fitness peak. Between moves the evolution proceeds as for a static landscape [Eq. (1)]. When a shift occurs we then apply the R_{ij} transformation to the concentration vector. The resulting concentration distribution is used as the initial condition for the rate equations from time τ to 2τ, and so on. The resulting population dynamics are shown in Fig. 1 (after the initial transient). A single unit on the (continuous) time scale of the rate equations is identified as a "generation" as it is obviously the mean population replacement time of the equivalent discretized dynamics.

For a static landscape the existence of an error threshold is intuitively clear: the superior fitness (and hence growth rate) of the master sequence must compensate for the mutational loss of Γ_0 individuals. The picture of the error threshold on a dynamic fitness landscape is different: what determines the critical mutation rate is whether the master sequence will have time to regrow between the shifts of the fitness peak and whether any kind of equilibrium can be reached within this dynamic. The existence and uniqueness of such a fixed cycle can be easily shown by noting that evolution over a full shift cycle is Markovian—both the shift and the evolution in the static part of the cycle are Markovian processes. Thus error thresholds are independent of initial population distribution (though this is not the case for recombination, e.g., [24]).

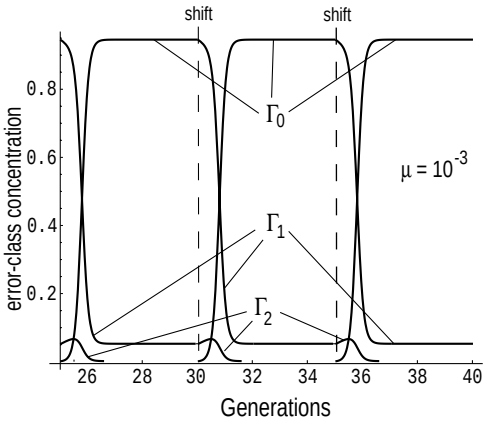

FIG. 1. $n = 50, \tau = 5, \sigma = 10$. The time evolution of the concentrations of the first three error classes, the genotypic composition of which changes as the landscape moves. The mutation rate is such that between shifts the population is dominated by the different master sequences.

To find an analytical approximation for any error threshold we will obviously have to include the dynamics of error class one as well as the master sequence. We can write an approximation to the rate equations for the master sequence $\Gamma_0 \equiv x_{\text{mas}}$ and a representative member j of Γ_1:

$$\begin{aligned} \dot{x}_{\text{mas}} &= (Q\sigma - f)x_{\text{mas}}\,, \\ \dot{x}_{1j} &= \tilde{Q}\sigma x_{\text{mas}} + (Q - f)x_{1j}\,, \end{aligned} \tag{3}$$

where $f = \sigma x_{\text{mas}} + 1 - x_{\text{mas}}$, $Q \equiv q^n$ is the genomic copy fidelity and mutations from one sequence into a neighboring sequence occur with probability $\tilde{Q} = (1 - q)q^{n-1}$. In deriving Eq. (3) back mutations are neglected in the standard way so that the x_{mas} dynamics are decoupled from x_{1j} i.e., only the dominant terms in each equation remain and taking x_{mas} such that $Q\sigma x_{\text{mas}} \gg \tilde{Q}\sum_j x_{1j}$ and $\tilde{Q}\sigma x_{\text{mas}} \gg \tilde{Q}n x_{2i}\ \forall i \in \Gamma_2$. We now assume $x_{1j}(0) = 0$ as x_{1j} is (almost always) in the sparsely populated Γ_2 before the shift. Using this boundary condition we can solve the linearized forms of Eq. (3) [linearized by a change of variables $y(t) = e^{\int^t f(s)\,ds}x(t)$] for the non-normalized concentrations

$$y_{\text{mas}}(t) = y_{\text{mas}}(0)e^{(Q\sigma)t}, \tag{4}$$

$$y_{1j}(t) = y_{\text{mas}}(0)\left(\frac{(e^{(Q\sigma)t} - e^{(Q)t})(1 - q)\sigma}{(\sigma - 1)q}\right). \tag{5}$$

To find an error threshold we are interested in the growth of y_{mas} *relative* to that of a representative genotype y_E some distance from the master sequence (i.e., in the so-called error tail [6]). If we assume that in the error tail a detailed balance holds between mutations into and out of

y_E, then its non-normalized growth will be by a factor of e^t (as $A_E = 1$). In the absence of movement the growth of y_{mas} is given by Eq. (4) and thus the master-sequence occupancy is maintained relative to a member of the error tail when $e^{Q\sigma t} \geq e^t \rightarrow Q_c \geq \sigma^{-1}$, i.e., we find the standard stationary error threshold. When the landscape is made time dependent the initial (non-normalized) concentration of the master sequence at the beginning of a new shift cycle (starting at time τ) will be $y_{1j}(\tau)$. Thus the (non-normalized) growth of the master sequence over a full shift cycle is $\kappa' \equiv y_{1j}(\tau)/y_{\text{mas}}(0)$. To find error threshold(s) we again compare this growth to that of y_E over the same cycle, finding a normalized growth factor for x_{mas}: $\kappa = \kappa'/e^\tau$. If $\kappa < 1$ the quasispecies will die out and the distribution of concentrations will eventually become uniform. Thus

$$\kappa \equiv \frac{(e^{(Q\sigma-1)\tau} - e^{(Q-1)\tau})(1-q)\sigma}{(\sigma-1)q} \geq 1 \qquad (6)$$

gives a condition for the long term maintenance of a nonzero population on the (moving) normalized master sequence x_{mas}. Hence we define the error threshold(s) to be the root(s) of Eq. (6).

Figure 2(b) shows the region where Eq. (6) can be expected to hold. The figure also shows the existence of two error-thresholds, q_{cm} and q_m. The lower threshold q_{cm} is a modified version of the classic error-threshold q_c present on static landscapes, with a perturbation resulting from the movement of the fitness landscape. Figure 2(a) shows the full time-dynamics [Eq. (1)] of x_{mas} when $q_c < q < q_{\text{cm}}$, i.e., a population that should stabilize on a static landscape cannot survive when it moves.

The upper threshold q_m is a new phenomenon that appears only on dynamic or moving (m) fitness landscapes. The existence of such a point intuitively clear—if the mutation rate is very close to zero, there will not be enough individuals present on the new peak position when the shift occurs to grow and maintain a steady occupancy of the master sequence, i.e., the peak moves out from under the quasispecies and the population will not be able to track shifts in the fitness landscape.

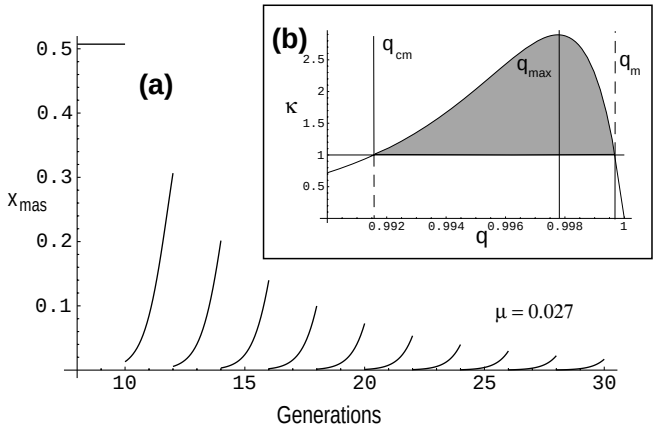

FIG. 2. (a) $q_c = 0.968 < q = 0.99 < q_{\text{cm}} = 0.9916$ the population is below the (dynamic) threshold and so becomes extinct; cf. Fig. 1. (b) κ is plotted as a function of the copying fidelity q. The shaded region is where self-replicating systems are able to maintain themselves. $n = 50, \tau = 2, \sigma = 5$ for both graphs.

Analytical approximations to the error thresholds can be found by assuming different dominant terms. To find the lower threshold q_{cm} we assume the exponential terms in κ determine the growth behavior. We can then find a first order correction in τ to the static threshold by solving Eq. (6) for Q (in terms of q) and then approximating $q \approx \sigma^{-1/n}$

$$Q_{\text{cm}} \approx \frac{1}{\sigma} - \frac{\ln(\sigma^{1/n} - 1)}{\sigma\tau}, \qquad (7)$$

where we also made the approximation $\frac{\sigma}{\sigma-1} \approx 1$. Note that $Q_{\text{cm}} \rightarrow Q_c$ when $\tau \rightarrow \infty$, i.e., we recover the stationary landscape limit.

Figure 3 shows the cm threshold and demonstrates the accuracy of the analytic approximation to q_{cm} given by Eq. (7). This accuracy is observed at the three significant figure level for an order of magnitude change in both σ and τ. Both the qualitative and quantitative dynamics of both error thresholds have been further verified by computer simulations using large populations to approximate the deterministic dynamics.

In Eq. (7) the critical copying fidelity Q_{cm} depends on the genome length. This is not surprising since the fitness peak shifts into a specific member of Γ_1, which consists of n different gene sequences. It is important to note that this is a direct consequence of the dynamic nature of the fitness landscape since Q_c is independent of genome length. The perturbation from the static error threshold increases with genome length and when n becomes large enough the lower error threshold coincides with the upper. We will discuss the consequences of this shortly.

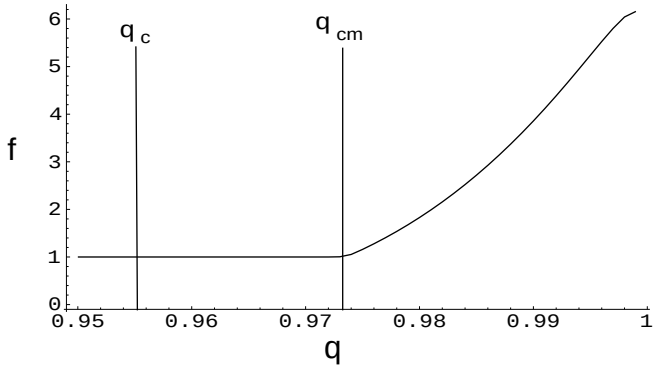

FIG. 3. The mean fitness f (averaged over a full shift cycle once equilibrium had been reached) found by numerically solving the full rate equations [Eq. (1)]. $\tau = 2, n = 50, \sigma = 10$. The error threshold occurs at the approximated analytic value $q_{\text{cm}} = 0.973$.

An analytical approximation to the new upper threshold can be found by taking Q to be very close to 1: therefore the $(1 - q)$ term determines whether $\kappa < 1$. Solving Eq. (6) for this term, then again assuming $\frac{\sigma}{\sigma-1} \approx 1$ and putting $Q \approx 1$, gives

$$q_m \approx 1 - e^{-(\sigma-1)\tau}. \quad (8)$$

Explicit numerical solutions of the full dynamics confirm that this threshold exists and is predicted by Eq. (8). For most values of σ and τ, q_m is very close to 1 [e.g., $(\sigma - 1)\tau = 50 \Rightarrow \mu_m = 10^{-22}$]. Computer simulations indicate that finite population affects are, however, much more significant for the upper error threshold than for the static threshold (for which, see [6,11] and others). In real biological populations this may be important. More detailed studies of these issues are under preparation.

On a static fitness landscape it is always possible to find copying fidelities high enough for evolution to be effective. It turns out that this is no longer the case for dynamic fitness landscapes. The strong dependence of Q_m on n means that as the genome length increases, the region between the two error thresholds where organisms may survive narrows, i.e., the shaded region of Fig. 2(b) contracts. Thus, perhaps surprisingly, there exist dynamic fitness landscapes (σ, τ, n) where solutions to Eq. (6) cease to exist. To find this convergence point we assume the leading behavior of Eq. (6) is dominated by the factor $e^{(q^n\sigma-1)\tau}(1 - q)$. This gives

$$q_{\max} \approx 1 - \frac{1}{\sigma\tau n} \quad (9)$$

as the turning point of $\kappa(q)$ [see Fig. 2(b)] which we can use to find $\kappa(q_{\max}) \geq 1$ as the definition of a survivable (or evolvable) landscape. The essence of this new result is that the selective advantage required for the survival of a quasispecies becomes large for fast-moving landscapes and long genome lengths. Again the effects of finite population sizes may be quite significant and are under investigation.

Thus we have demonstrated the existence of, and derived analytic approximations for, two quasispecies error thresholds on a simple time-dependent fitness landscape. The lower threshold $q_{\rm cm}$ is a perturbation of the well-known (static) error catastrophe, accounting for the destabilizing effect of the changing environment. The existence of an upper bound on the copying fidelity q_m is a new phenomenon, only present in dynamic environments. This upper bound results in the surprising existence of critical regions of the landscape parameters (selection strength σ, genome length n, and the rate of environmental change τ) where the two thresholds coincide (or cross), and therefore no effective selection can occur no matter what copying fidelity can be evolved. Because of the simplicity of our model and its closeness to Eigen's original formulation, we hope that the general nature of these results will be fundamental in many time-dependent landscapes.

The authors would like to thank Claes Andersson, Erik van Nimwegen, and Mats Nordahl for useful discussions. Nigel Snoad and Martin Nilsson were supported by SFI core funding grants. N. S. would also like to acknowledge the support of Marc Feldman and the Center for Computational Genetics and Biological Modeling at Stanford University while preparing this manuscript.

†nigel@santafe.edu

[1] M. Eigen, Naturwissenschaften **58**, 465 (1971).
[2] M. Nowak, Trends Ecol. Evol. **7**, 118 (1992).
[3] P. Schuster and K. Sigmund, Ber. Bunsen-Ges. Phys. Chem. **89**, 668 (1985).
[4] I. Leuthäusser, J. Chem. Phys. **84**, 1884 (1986).
[5] P. Tarazona, Phys. Rev. A **45**, 6038 (1992).
[6] M. Nowak and P. Schuster, J. Theor. Biol. **137**, 375 (1989).
[7] M. Eigen, J. McCaskill, and P. Schuster, Adv. Chem. Phys. **75**, 149 (1989).
[8] J. Swetina and P. Schuster, Biophys. Chem. **16**, 329 (1982).
[9] L. Bonhoeffer and P. Stadler, J. Theor. Biol. **164**, 359 (1993).
[10] P. Higgs, Genet. Res. Camb. **63**, 63 (1994).
[11] D. Alves and J. Fontanari, Phys. Rev. E. **57**, 7008 (1998).
[12] E. Baake, M. Baake, and H. Wagner, Phys. Rev. Lett. **78**, 559 (1997).
[13] M. Eigen and P. Schuster, Naturwissenschaften **64**, 541 (1977).
[14] J. Maynard Smith and E. Szathmáry, *The Major Transitions in Evolution* (W. H. Freeman, Oxford, 1995).
[15] J. Drake, Proc. Natl. Acad. Sci. U.S.A. **90**, 4171 (1993).
[16] E. Domingo *et al.*, J. Fed. Am. Soc. Exp. Biol. **10**, 859 (1996).
[17] L. Van Valen, Evol. Theory **1**, 1 (1973).
[18] B. Levin, M. Lipsitch, and S. Bonhoeffer, Science **283**, 806 (1999).
[19] M. Kimura, Genet. Res. **9**, 23 (1967).
[20] J. Gillespie, *The Causes of Molecular Evolution* (Oxford University Press, Oxford, United Kingdom, 1991).
[21] K. Ishii, H. Matsuda, Y. Iwasa, and A. Saskai, Genetics **121**, 163 (1989).
[22] *Mathematical Ecology: An Introduction,* edited by T. Hallam and S. Levin (Springer-Verlag, New York, 1986).
[23] C. O. Wilke, C. Ronnewinkel, and T. Martinetz, in *Proceedings of ECAL'99,* Lecture Notes in Computer Science, edited by H. Lund and R. Kortmann (Springer-Verlag, Heidelberg, 1999), p. 417; LANL e-print physics/9904028.
[24] M. Boerlijst, S. Bonhoeffer, and M. Nowak, Proc. R. Soc. London B **263**, 1577 (1996).