RNA Virus Evolution via a Fitness-Space Model

Lev S. Tsimring and Herbert Levine
Institute for Nonlinear Science, University of California, San Diego, La Jolla, California 92093-0402

David A. Kessler
Department of Physics, Bar-Ilan University, Ramat Gan 52900, Israel
(Received 16 January 1996)

We present a mean-field theory for the evolution of RNA virus populations. The theory operates with a distribution of the population in a one-dimensional fitness space, and is valid for sufficiently smooth fitness landscapes. Our approach explains naturally the recent experimental observation [I. S. Novella *et al.*, Proc. Natl. Acad. Sci. U.S.A. **92**, 5841–5844 (1995)] of two distinct stages in the growth of virus fitness. [S0031-9007(96)00371-7]

PACS numbers: 87.10.+e, 82.20.Mj

RNA viruses offer a unique opportunity for the experimental study of molecular evolution. These viruses exhibit both high replication rates (10^5 day^{-1}) and high mutation rates [$10^{-4} - 10^{-5}$ mutation/(nucleotide/replication)]; hence, evolutionary dynamics which would take years to unfold in even relatively simple bacteria occur within days in RNA virus colonies. The temporal dynamics of one such system has been studied in a series of recent experiments [1]. In this paper, we show how a simple model for the motion of the virus population on a fitness space can account for some interesting findings in these studies.

In the experiments by Holland and co-workers, clones of vesicular stomatitis virus (VSV) were carried through a transmission series of up to 100 consecutive daily "passages" (the experimental technique was first described in [2]). Every passage begins with the inoculation of approximately 10^5 viruses in a bottle containing a monolayer of fresh cells. The viruses are allowed to replicate for one full day, with the number of viruses at the end of the day reaching 10^{10}. Then a subsample of approximately 10^5 viruses is taken from the bottle and used for the next passage. During the series an average fitness of the evolving clone (ec) is measured as follows. A small sample of viruses is separated from the main population and mixed with a sample of wild-type (wt) virus; the wild type, which serves as a reference, is taken from a frozen nonevolving stock. This mixture is then carried through a few passages at which the ratio c_r of ec concentration c_e to the wt concentration c_w is measured daily. The logarithm of the relative fitness is determined as a slope of $\log c_r(n)$ vs n, where n is a number of (daily) passage [3].

The specific findings we wish to study concern the evolution of colonies whose fitness had been artificially lowered using a "genetic bottlenecking" procedure; this bottleneck is created by doing passages with only one particle transferred from one bottle to the next, and arises because the repeated small sampling does not allow for the effective suppression of deleterious mutants via selection [4]. A typical plot of the temporal dynamics of the logarithm of fitness gained by a clone with initially low fitness is shown in Fig. 1. The logarithm of fitness grows rapidly until it reaches zero (i.e., the fitness of ec becomes equal to the fitness of wt). After that, the logarithm of fitness continues to grow linearly, however at a slower rate [5]. In about 50 days the relative fitness reached values of the order of 10 which is quite a remarkable increase. For comparison, fitness in *Escherichia coli* [6] bacteria colonies increased by 8% after 400 generations of monitored evolution.

Our purpose is to show that a "mean-field" model of evolution of the virus population on a one-dimensional fitness space can naturally account for this experimental data. Most discussions of evolution are based on the alternative notion of the sequence space originally introduced by Wright [7] (see also [8,9]). In this extremely high-dimensional space the number of dimensions is equal to the number of nucleotides in the genome (for VSV virus over 11 000), and every point represents a particular genetic sequence (genome). Each genome can be labeled with a fitness value (which is related to the replication rate of the corresponding virus); thus a fitness landscape is formed in the sequence space. Typically, one writes down equations governing the population dynamics of each genome taking into account replication and mutation. This

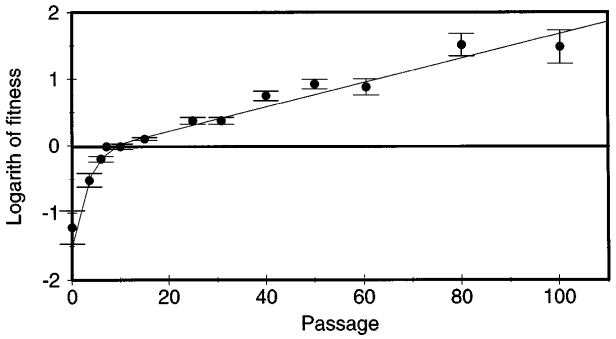

FIG. 1. Evolution of fitness of a monoclonal-antibody resistant clone (MARM) during the transmission series of 80 passages on HeLA cells (Fig. 2b of [1]).

 0031-9007/96/76(23)/4440(4)$10.00 © 1996 The American Physical Society

approach ignores the fact that in reality an average number of species (molecules) per genetic sequence is very small (except for maybe the most common "master" sequence), and fluctuations due to the discreteness of molecules and the stochastic nature of mutations can be very significant. Nonetheless, these models make a nontrivial prediction that a cloud of mutants surrounding the master sequence (the "quasispecies") is formed in the course of an evolutionary process. Unfortunately, a large number of equations and the need for detailed assumptions regarding genome fitness functions and mutation rates make it difficult to obtain simple qualitative insight into the temporal dynamics of the quasispecies. To date, only relatively small-scale simulations have been performed [10] of relevance to RNA molecules (sequence length of about 10^2) as compared with genome length [$O(10^4)$ for RNA viruses].

Our approach to the description of molecular evolution is rather different. As mentioned above, every sequence can be characterized by its replication rate r. There may be different sequences which exhibit similar replication rates; we shall lump them together and introduce a time-dependent density of population per unit of fitness, $p(r,t)$ [11]. Without mutations, the dynamics of the population can be described by

$$\partial p(r,t)/\partial t = rp(r,t). \tag{1}$$

Actually, this continuum description will be valid only if there are a sufficient number of virus particles at a specific r; this assumption is, however, much easier to justify for population dynamics in the one-dimensional fitness space than for high-dimensional sequence space. If on average the population is constrained to have constant size $p_{\rm tot} = \int p(r,t)dr$ (such as is accomplished by the passages technique), Eq. (1) is modified to become

$$\partial p(r,t)/\partial t = (r - \bar{r})p(r,t), \tag{2}$$

where $\bar{r} = \int rp(r,t)dr/p_{\rm tot}$ is an average fitness of the population. It is easy to see that if initially population is distributed arbitrarily over a range of fitnesses $\{r_{\rm min}, r_{\rm max}\}$, its average fitness will grow monotonically until it eventually reaches $r_{\rm max}$. In this final state the limit distribution function $p_\infty(r) = p_{tot}\delta(r - r_{\rm max})$, i.e., all sequences have identical fitness.

Our basic hypothesis is that one can also include mutation effects in the equation for p without explicitly taking into account the underlying genomic transition rates [12]. In the simplest form such an equation one could have is

$$\frac{\partial p(r,t)}{\partial t} = (r - \bar{r})p(r,t) + D\frac{\partial^2 p(r,t)}{\partial r^2} + \frac{\partial}{\partial r}[v_{\rm drift}p(r,t)]. \tag{3}$$

This choice depends on the landscape being smooth, so that single mutations can only lead to small fitness changes, and also requires uniformity of the mutation rate, i.e., the rate does not depend on the fitness itself. The drift term arises due to the predominance of deleterious mutations over beneficial ones, as indicated by the reduction in overall fitness seen in the bottlenecking protocol; $v_{\rm drift}$ depends, in general, on the fitness, although it can be taken to be constant over a sufficiently small fitness range. This model equation (without the drift) is reminiscent of mean-field models for diffusion-limited growth, where p describes a solid cluster growing in a linear gradient of diffusing particles [13].

We expect this type of equation to be valid for systems with smooth landscapes such that the population is always sampling a large number of genomes with similar fitness; it will break down if there is a tendency for the population to get trapped in local maxima and hence become localized at specific genomes with low transition probabilities to more fit variants. To understand how a description based on a mean-field type description can emerge in situations with smooth fitness landscapes, we consider a simple model for which the fitness of a binary-string genome is just the proportion of 1's [14]. The underlying dynamics is a continuous-time Markov process with reproduction and with single-bit flipping mutation at rate λ. It is easy to see that this process can be described by the population equations

$$\frac{dp_n(t)}{dt} = \frac{n}{N}p_n(t) + \lambda\left(\frac{n+1}{N}p_{n+1}(t) + \frac{N-n+1}{N} \times p_{n-1}(t) - p_n(t)\right), \tag{4}$$

where $p_n(t)$ is the time-dependent concentration of strands with n ones, N is the total genome size, and the prefactors come from considering the random choice of which single bit gets flipped. If we define $r = (n - N/2)/\sqrt{N}$, rescale $\lambda = \tilde{\lambda}\sqrt{N}$, $t = \tilde{t}\sqrt{N}$, and take the continuous r limit, we obtain Eq. (3) (with $D = \tilde{\lambda}/2$ and $v_{\rm drift} = 2\tilde{\lambda}r$). This genome model makes the rather unrealistic assumption of independent additive genomic contributions to the overall fitness. We would like to stress, however, that we expect the general idea of a fitness-space equation to remain valid for more complex landscapes as long as trapping in metastable states is irrelevant [15].

The fact that the experiments find constant fitness growth implies that the drift term relevant for the virus population is relatively unimportant. The presence of a small constant drift term has no effect on the qualitative features of the fitness dynamics and hence will be dropped in what follows. Then, as already mentioned, this model is similar to one studied within the context of diffusion-limited growth. There, it has been shown [16] that naively using the continuum equation (3) down to arbitrarily small $p(r,t)$ leads to a divergence of the average value of r (i.e., the mean fitness) in finite time. The essential cause of this effect is the unlimited growth rate of p at large r. Given this fact, any small nonzero value of p at large r will start to grow very quickly. In a real viral dynamics, this cannot occur. A mutant population cannot grow via reproduction

until there actually exists at least one mutant virus—that is, the discreteness of the population provides a cutoff [17]. Furthermore, the transferring of only a small part, 10^{-5}, of the population at the end of each passage will usually eliminate strains which do not have significant numbers. In fact there is experimental evidence that even very highly fit mutants do not grow in a population if their relative concentration is very small [18]. Thus, we assume that there exists a threshold value of concentration p_c below which the autocatalytic growth of concentration ceases. Therefore we replace Eq. (3) by

$$\frac{\partial p(r,t)}{\partial t} = \theta(p - p_c)(r - \bar{r})\, p(r,t) + D\frac{\partial^2 p(r,t)}{\partial r^2}, \tag{5}$$

where $\theta(x)$ is the Heaviside step function.

Figure 2(a) illustrates numerical simulation of Eq. (5) starting from a random distribution of the population within some interval $\{r_{\min}, r_{\max}\}$—this is meant to mimic the "bottlenecked" population. In Fig. 2(b) temporal dynamics of the average replication rate $\bar{r}$ is shown for this run. A crossover from a fast growth of the replication rate to a relatively slow linear growth of $\bar{r}$ is observed, which resembles closely the experimental dependence (Fig. 1). The first (fast) phase involves formation of a pulselike distribution which is localized near $r_{\max}$. In this distribution the selection term is approximately balanced by a mutation (diffusion) term in Eq. (5). This distribution is highly peaked near the most fit of the initial state, which presumably arose from a wild-type virus

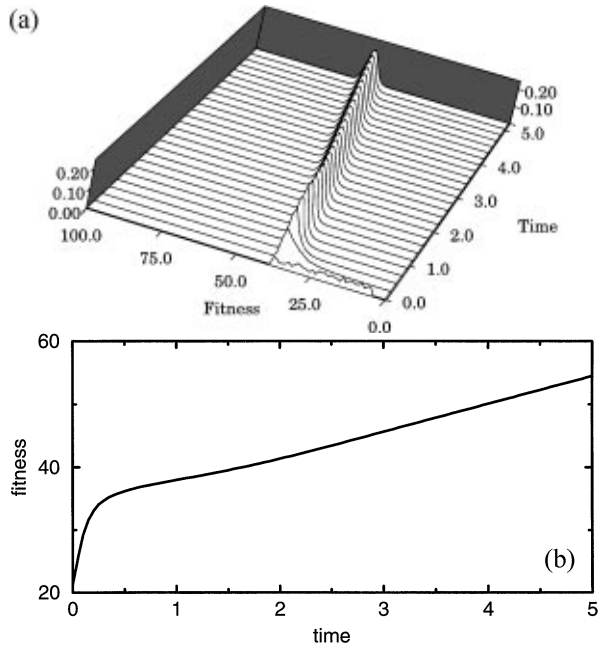

FIG. 2. Evolution of the viral colony concentration in fitness space starting from random distribution within a range $\{5.0 < r < 18.0\}$ in the framework of Eq. (5). The diffusion constant is $D = 1$, $p_{\text{tot}} = 1$, and threshold concentration $p_c = 5 \times 10^{-4}$: (a) space-time diagram $p(r,t)$ and (b) average fitness $\bar{r}$ versus time.

which by chance did not accumulate many deleterious mutations during the bottlenecking phase; indeed, the crossover value of the replication rate is observed to be close to the wild-type value.

At the second (slow) phase the distribution drifts slowly towards higher values of the replication rate while keeping its stationary pulselike form. If the initial condition already corresponds to a pulse distribution, then the dynamics skip the first phase and proceed directly to the second. This is also consistent with observations of the neutral clone evolution [5].

The stationary-moving pulse solution and its selected velocity can be found analytically. First, we introduce new variables $\tilde{t} = D^{1/3}t$ and $\tilde{r} = D^{-1/3}r$ to scale out the diffusion constant D. Dropping tildes and assuming that p in the asymptotic regime depends on one variable $z = r - ct$, we have $\bar{r} = ct + r_0$, where

$$r_0 = \int zp(z)dz \Big/ \int p(z)dz\,. \tag{6}$$

The equation for $p(z)$ (in the region $p > p_c$) reads

$$p'' + cp' + zp = r_0 p\,, \tag{7}$$

(primes denote differentiation with respect to z) which has the solution

$$p(z) = p_0 e^{-cz/2}\mathrm{Ai}(-z)\,, \tag{8}$$

where we introduced a new variable $\tilde{z} = z - r_0 - c^2/4$ and dropped the tilde again. This solution is valid only for $-\infty < z < z_c$, where z_c is defined by the condition $p(z_c) = p_c$. At $z > z_c$, we must match it with a stationary moving solution of the diffusion equation (with $D = 1$), namely,

$$p(z) = p_c e^{-c(z-z_c)}. \tag{9}$$

Matching across the interface $z = z_c$ should satisfy two boundary conditions, $p_+ = p_-$ and $p'_+ = p'_-$, and the normalization condition, $\int_{-\infty}^{\infty} p(z)dz = p_{\text{tot}}$. Here p_- is solution (8) at $z = z_c$, and p_+ is solution (9) at $z = z_c$. These conditions yield a system of three equations for p_0, c, and z_c,

$$p_0 e^{-cz_c/2}\mathrm{Ai}(-z_c) = p_c\,, \tag{10}$$

$$-\frac{c}{2}p_0 e^{-cz_c/2}\mathrm{Ai}(-z_c) + p_0 e^{-cz_c/2}\mathrm{Ai}'_z(-z_c) = -cp_c\,, \tag{11}$$

$$p_0\int_{-\infty}^{z_c} e^{-cz/2}\mathrm{Ai}(-z)dz + \frac{p_c}{c} = p_{\text{tot}}\,, \tag{12}$$

For small p_c/p_0, matching point z_c is close to the first zero of the Airy function, so we can introduce a small correction $\xi = z_0 - z_c$. Here $z_0 \approx 2.3381\ldots$. For small ξ, we find from (10) and (11) that $\xi = 2/c$, and

$$p_0 = (c/2)p_c e^{(c/2)(z_0-1)}|\mathrm{Ai}'(-z_0)|^{-1}. \tag{13}$$

[here $|\mathrm{Ai}'(-z_0)| \approx 0.700\ldots$.] Substituting this into

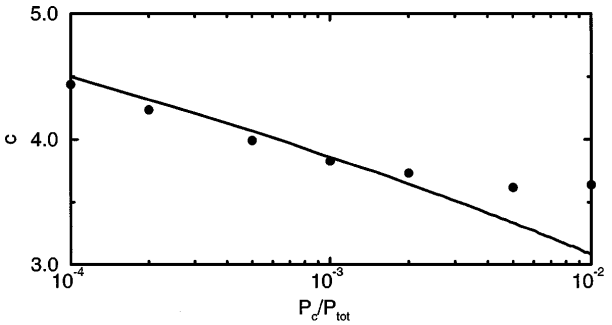

FIG. 3. The rate of fitness increase as a function of the threshold ratio $p_c/p_{\rm tot}$: solid line, theoretical formula (15); dots, numerical simulations of (5) with $D = 1$.

Eq. (12) and neglecting small second term in the left hand side, we obtain a closed equation for the pulse speed c,

$$ce^{(c/2)(z_0-1)}\int_{-\infty}^{z_0} e^{-cz/2}\mathrm{Ai}(-z)dz = 2\frac{p_{\rm tot}}{p_c}|\mathrm{Ai}'(-z_0)|\,. \tag{14}$$

For large c the integral in (14) can be evaluated asymptotically using the Laplace method [19], as $\exp(c^3/24)$, so Eq. (14) simplifies to

$$ce^{c^3/24+(c/2)(z_0-1)} = 2(p_{\rm tot}/p_c)|\mathrm{Ai}'(-z_0)|\,. \tag{15}$$

The solution of this equation is shown in Fig. 3 together with the results of direct numerical simulation of (5); the theoretical estimate agrees quite well with the numerics at small $p_c/p_{\rm tot}$ but underestimates the speed at higher $p_c/p_{\rm tot}$. At small values of the threshold $p_c/p_{\rm tot}$, the fitness growth rate c depends on the threshold value very weakly, as $[-\ln(p_c/p_{\rm tot})]^{1/3}$. Returning to dimensional variables, the rate of fitness growth c is proportional to the two-third power of the diffusion constant D (which is related to the mutation rate of the virus).

In this paper we introduced a model for viral population dynamics which operates in a one-dimensional fitness space rather than in an extremely high-dimensional sequence space. This model is expected to hold for the evolution on smooth fitness landscapes. The resulting equation (5) can quite naturally explain the results of experiments on the evolution of the VSV *in vitro* [1]. The first stage of the exponential growth of fitness (Fig. 1) corresponds to formation of the quasistationary population distribution in the fitness space, which is close to the wild-type distribution. On the second stage, the quasispecies distribution "propagates" towards higher fitness values with a constant speed which has been analytically deduced from Eq. (5). Our findings suggest that it is quite interesting to understand the conditions under which evolution in terms of genomes can be rephrased in terms of fitness and thereby become amenable to simple analysis and simulation.

We wish to thank R. Huerta, I. S. Novella, J. Quer, and J. J. Holland for stimulating discussions. L. T. acknowledges support from U.S. Department of Energy Grants DE-FG03-95ER14516 and DE-FG03-96ER14592; H. L. is supported in part by NSF Grant DMR94-15460.

[1] I. S. Novella, E. A. Duarte, S. F. Elena, A. Moya, E. Domingo, and J. J. Holland, Proc. Natl. Acad. Sci. U.S.A. **92**, 5841–5844 (1995).
[2] J. J. Holland, J. C. de la Torre, D. K. Clarke, and E. Duarte, J. Virol. **65**, 2960 (1991).
[3] We want to avoid confusion between the notion of fitness accepted in virological literature, and the replication rate used here (we sometimes will be calling it fitness also). Fitness F is usually measured as a rate of the population increase over some finite time interval between passages T (typically, $T = 1$ day). Replication rate r, on the other hand, is an instantaneous rate of exponential growth of the population [see Eq. (1)]. Obviously, $F = \exp(rT)$, so the logarithm of fitness $\log F$ corresponds directly to the replication rate. That is why we argue that "exponential growth of fitness" in [1] is, in fact, a linear growth of the average replication rate $\bar{r}$.
[4] H. J. Muller, Mutat. Res. **1**, 2 (1964).
[5] In experiments, a neutral clone (with fitness approximately equal to the one of the wild-type) was also run through a series of large-population passages (see [1]). In this case, only one phase of the exponential growth of fitness was observed.
[6] R. Lenski and M. Travisano, Proc. Natl. Acad. Sci. U.S.A. **91**, 6808–6814 (1994).
[7] S. Wright, Genetics **16**, 97 (1931).
[8] S. A. Kaufman, *The Origins of Order* (Oxford University Press, Oxford, 1993).
[9] M. Eigen and C. Biebricher, in *RNA Genetics*, edited by E. Domingo, J. J. Holland, and P. Ahlquist (CPC Press, Boca Raton, FL, 1988).
[10] W. Fontana, W. Schnabl, and P. Schuster, Phys. Rev. A **40**, 3301 (1989).
[11] This is reminiscent of the work by T. B. Kepler and A. S. Perelson [J. Theor. Biol. **164**, 37 (1993)] who assigned different B-cell clones with similar affinities to a single category (affinity class) for modeling somatic hypermutation in the vertebrate immune system.
[12] Indeed, among many mutations that change genotype, only a few may actually change phenotype (and fitness); see M. A. Huymen, P. F. Stadler, and W. Fontana, Proc. Natl. Acad. Sci. U.S.A. **93**, 397 (1996).
[13] T. A. Witten and L. M. Sander, Phys. Rev. B **27**, 5686 (1983).
[14] This is a limiting case of Kaufman's NK model [8] with $K = 0$.
[15] The idea that motion along strictly *neutral* fitness landscapes takes place by a diffusive process is due originally to M. Kimura [*The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, England, (1983)].
[16] A. Arneodo *et al.*, Phys. Rev. Lett. **63**, 984 (1989); E. Brener, H. Levine, and Y. Tu, Phys. Rev. Lett. **66**, 1978 (1991).
[17] A similar idea was put forth by T. B. Kepler and A. S. Perelson [Proc. Natl. Acad. Sci. U.S.A. **92**(18), 8219 (1995)].
[18] J. C. de la Torre and J. J. Holland, J. Virol. **64**, 6278 (1990).
[19] A. H. Nayfeh, *Introduction to Perturbation Techniques* (J.Wiley and Sons, New York, 1981).