Selection-mutation process of RNA viruses

Amal Aafif and Juan Lin
Department of Physics, Washington College, Chestertown, Maryland 21620
(Received 1 August 1997)

RNA viruses mutate at a rate 10^5-10^6 times faster than their DNA counterparts. This process can be simulated by a continuous stochastic model on a smooth one-dimensional fitness landscape where selection forces the viral quasispecies to climb uphill to higher fitness values. Theoretical results of the model with drift velocity proportional to fitness are fitted to the experimental observations made by Novella *et al.* [Proc. Natl. Acad. Sci. U.S.A. **92**, 5841 (1995)]. [S1063-651X(98)11802-0]

PACS number(s): 87.10.+e

RNA viruses such as influenza A, the human immunodeficiency virus (HIV), and the vesicular stomatitis virus (VSV) are known to have unusually high mutation rates [1]. Rapid evolution and high replication rates allow these viruses to evade a host's acquired immunity to previous strains and form populations with high levels of polymorphism and varying degrees of fitness [2]. In viral replication, point mutations accumulate over time, creating new subtypes of the same virus that can reinfect previously immune hosts. This stochastic mutation process, known as genetic drift, along with the natural forces of selection, generates a distribution of mutants moving in a swarm with the attributes of a quasispecies [3,4]. Instead of a homogeneous population dominated by the fastest replicating mutant, a quasispecies is characterized by an entire group of clones upon which selection operates [5]. This group is adapted to evolve toward local or global peaks on a multidimensional rugged fitness landscape.

A typical RNA virus has a genome size consisting of over 10^4 nucleotide bases mutating at rates of about $10^{-4}-10^{-5}$ mutations per nucleotide per replication. The sequence space for mutation, defined as the total volume in which all possible variations of the genome are represented, is extremely large (about $4^{10\,000}$ states) although only a small portion of this space is available for drift evolution [6]. In recent experimental studies of virus growth in cell cultures [7], relative fitness values of VSV clones were measured through a series of replication passages. Fitness was then quantified as the slope of the logarithm of the relative concentration of the mutating clone to that of the wild-type virus. Results obtained from these experiments were fitted to the equation

$$r=A+Bt+Ce^{-Dt}, \tag{1}$$

where r is the average fitness parameter (relative growth rate) and A, B, C, and D are empirical parameters. The variable t measures the number of replication passages usually lasting 1 day. Relative fitness values were observed to increase by as much as 4 units in 50 passages.

A ''mean field'' model [8] for the evolution of a population along a one-dimensional fitness space was recently proposed to describe these experimental observations. Instead of assigning a fitness value to each genome sequence, the model clusters different sequences with similar replication rates into a probability density per unit fitness. This clustering of sequences averages the local peaks and valleys in genome space into a smooth fitness landscape. The loss of local detail in the model (existence of metastable states) is compensated by the possibility of predicting robust trends.

The evolution in fitness space is represented as a Markov process on a binary string of 1's and 0's of total length N (the genome size). The sum of 1's is directly proportional to the total fitness of the virus with no distinction made as to the location where the flips occur. It is found after neglecting a drift velocity that, in the continuous limit, the fitness variable r satisfies the equation

$$\frac{\partial P(r,t)}{\partial t}=\theta(P-P_C)(r-\langle r\rangle)P(r,t)+D\,\frac{\partial^2 P(r,t)}{\partial r^2}, \tag{2}$$

where D is a diffusion constant proportional to the mutation rate, θ the Heaviside step function, $\langle r\rangle$ the average fitness, and P_c an arbitrary lower threshold value that accounts for the discreteness of the replication process. Numerical simulation of this model shows a pulselike distribution with two stages of linear growth in approximate agreement with experiment. To prevent divergence in $\langle r\rangle$, the threshold value P_c is introduced. However, this assumption does not ensure the proper convergence of the asymptotic state.

In this paper, we study a continuous model derived from an almost identical discrete formulation as in Ref. [8]. The only three differences are the following: (a) we introduce a scale factor Δ to connect flips in the binary string to changes in fitness (phenotype space), (b) we keep the drift velocity independent of the mutation rate as a meaningful variable to ensure the convergence of statistical moments, and (c) we measure fitness as deviations from the most probable number of 1's ($\frac{1}{2}N$) in a string of genome size N. We assign the most probable state (zero fitness) to clones that grow as fast as the wild-type virus [7].

The continuous approximation is a partial differential equation with constant diffusion coefficient and variable drift velocity linearly dependent on fitness. The drift velocity can be interpreted as the gradient of a quadratic potential fitness function. This smooth fitness landscape seems to be a relatively good approximation to the experiments on RNA evolution. The continuous limit can be solved to obtain equations for the mean fitness and its variance. We show that if one starts with a symmetrical distribution of mutants, all cu-

mulants beyond the second vanish. Finally, we fit the theoretical predictions of the model to the experimental observations of Novella *et al.* [7].

The evolution of mutants can be visualized as a one-dimensional random walk in fitness space x with step size Δ and time between steps ζ. The probability to be at x at time t, $P(x=n\Delta,t)$, obeys the following equation:

$$P(x,t+\zeta)=(x-\langle x\rangle)\zeta P(x,t)+\left(\frac{1}{2}-\frac{x-\Delta}{N\Delta}\right)P(x-\Delta,t)+\left(\frac{1}{2}+\frac{x+\Delta}{N\Delta}\right)P(x+\Delta,t). \tag{3}$$

Selection is incorporated in the term $(x-\langle x\rangle)\zeta$, which only allows mutants with fitness values above the average to reproduce successfully. The constant transition probabilities in the last two terms account for mutation in a flat landscape while the x-dependent terms describe the asymmetric hill-climbing process along a fitness potential. The uphill climb of mutants forced by selection is counterbalanced by a higher probability of falling back to lower fitness levels (deleterious mutations) especially at high values of x, thereby preventing the average fitness from growing indefinitely.

The constant N constrains the random walk to the limits $-N\Delta\leqslant x\leqslant N\Delta$. To obtain a continuous approximation we subtract $P(x,t)$ from Eq. (3) and divide the whole expression by ζ. In the limit $\zeta\to 0$, $\Delta\to 0$, and $N\to\infty$, we find the continuous approximation

$$\frac{\partial P(x,t)}{\partial t}=(x-\langle x\rangle)P(x,t)+\lambda\,\partial_x(xP)+\frac{\mu}{2}\,\partial_x^2 P \tag{4}$$

with the mutation μ and drift λ coefficients defined by the limits [9]

$$\frac{\Delta^2}{\zeta}\to\mu \quad\text{and}\quad \frac{2}{N\zeta}\to\lambda. \tag{5}$$

The above partial differential equation represents a diffusion process with a drift velocity derived from a fitness potential $\frac{1}{2}\lambda x^2$.

We proceed to determine the average fitness and variance as functions of time. Equation (4) is used to evaluate the time derivative of the cumulant generating function [10]

$$Z(k,t)=\ln\left(\int e^{-ikx}P(x,t)dx\right). \tag{6}$$

After integration by parts we verify that Z satisfies the equation

$$\partial_t Z=i\,\partial_k Z-i\,\partial_k Z|_{k=0}-k\lambda\,\partial_k Z-\frac{\mu}{2}\,k^2, \tag{7}$$

which can be solved by the method of characteristics. The general solution is given as

$$Z(k,t)=Z_0(k,t)-Z_0(k=0,t), \tag{8}$$

where $Z_0(k,t)$ is the solution of Eq. (7) without the $k=0$ term. To determine this function we solve the system of equations

$$\frac{-dt}{1}=\frac{dk}{i-k\lambda}=\frac{-dz}{-\mu k^2/2}. \tag{9}$$

The general solution can be written as

$$Z_0(k,t)=\frac{\mu}{2}\left[\frac{-ik}{\lambda^2}-\frac{k^2}{2\lambda}+\frac{1}{2\lambda^3}\ln(1+k^2\lambda^2)+\frac{1}{\lambda^3}\arctan(k\lambda)\right]+\Omega[\lambda t-\ln(i-k\lambda)], \tag{10}$$

where Ω can be found by assuming the initial distribution P to be a Gaussian function centered at x_0 with standard deviation σ_0. Expressed in the coordinates of Z_0 we find

$$Z_0(k,0)=\tfrac{1}{2}\ln(2\pi\sigma_0^2)-ikx_0-\tfrac{1}{2}k^2\sigma_0^2. \tag{11}$$

The final expression for $Z_0(k,t)$ is rather long and we will not write it. But once this function is found we can derive the first and second cumulants by taking derivatives of $Z(k,t)$ with respect to k:

$$\langle x(t)\rangle=i\,\partial_k Z(k,t)|_{k=0} \quad\text{and}\quad \sigma^2(t)=i^2\partial_k^2 Z(k,t)|_{k=0}. \tag{12}$$

We quote the results. The average fitness

$$\langle x(t)\rangle=\frac{\mu+e^{-\lambda t}(2\sigma_0^2\lambda-2\mu+2x_0\lambda^2)+e^{-2\lambda t}(\mu-2\sigma_0^2\lambda)}{2\lambda^2} \tag{13}$$

reaches an asymptotic value $\langle x\rangle_\infty=\mu/2\lambda^2$ as $t\to\infty$. Similarly, the variance

$$\sigma^2(t)=\frac{\mu+e^{-2\lambda t}(2\sigma_0^2\lambda-\mu)}{2\lambda} \tag{14}$$

also reaches a constant value $\sigma_\infty^2=\mu/2\lambda$ as $t\to\infty$. By taking additional derivatives it is possible to show that all cumulants C_n of order higher than two vanish. This result justifies setting $C_3=0$ when the first two cumulants are calculated by direct integration of Eq. (4),

$$\frac{d\langle x\rangle}{dt}=\sigma^2-\lambda\langle x\rangle,$$

$$\frac{d\sigma^2}{dt}=C_3+\mu-2\lambda\sigma^2. \tag{15}$$

System (15) yields the same solutions as Eqs. (13) and (14).

We now fit Eq. (13) to the experimental observations on clones of monoclonal antibody-resistant mutants (MARM) of VSV [7]. These clones are progenies of virus particles with low initial fitness. Values for the initial conditions, x_0 and σ_0^2, as well as the drift and mutation coefficients are determined for each fit (Figs. 3(A), 3(B), and 3(C) of Ref. [7]). In all cases the mutation rate μ is found to be several

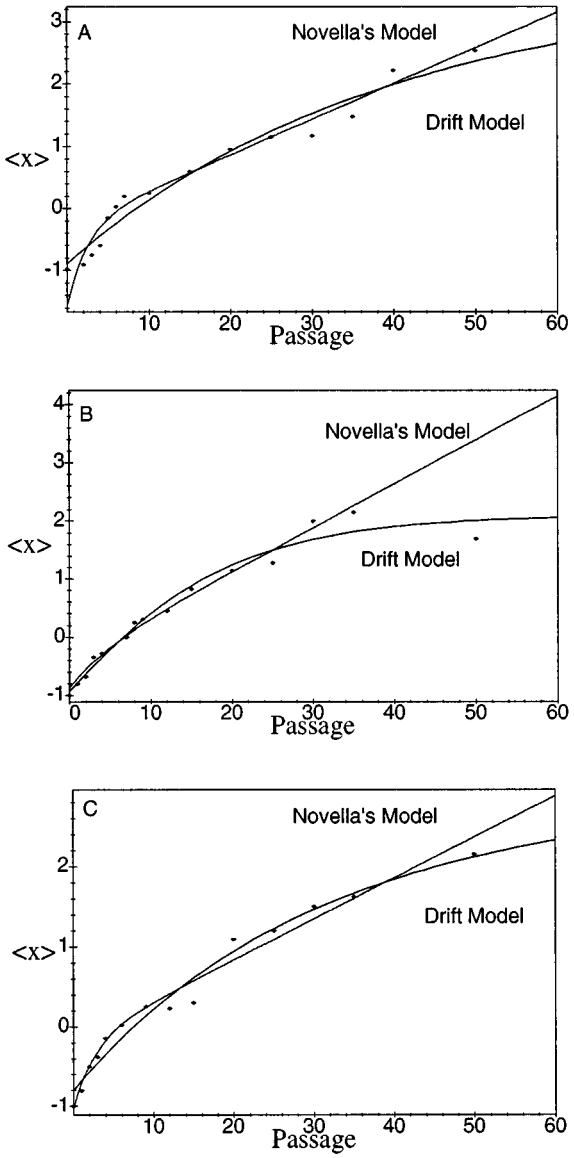

FIG. 1. Data points taken from Fig. 3 of Ref. [7]. Curve fit using Eq. (13) yields the parameter values for (A) $x_0 = -0.892$, $\lambda = 0.0196$, $\mu = 0.00294$, $\sigma_0^2 = 0.100$, and mean squared error (MSE) of 0.0446, (B) $x_0 = -0.930$, $\lambda = 0.0754$, $\mu = 0.0239$, $\sigma_0^2 = 0.0750$, and MSE of 0.0291, and (C) $x_0 = -0.803$, $\lambda = 0.0253$, $\mu = 0.00381$, $\sigma_0^2 = 0.101$, and MSE of 0.0209.

times smaller than the drift coefficient λ. By considering a smooth one-dimensional landscape we are confining ourselves to a hill-climbing process of selection-mutation without metastable states. More data for longer times would be required to verify if this picture is valid in the experiments on VSV. Figure 1 displays the data of Novella *et al.*, their fit (1) and formula (13). Our curves slightly underestimate the initial growth rate (for small samples, the initial Gaussian distributions are only an approximation) but they seem to work well for the latter portion of the data. In Fig. 2 we plot the three variances using the best fit parameters for the three graphs above. In cases (A) and (C) the variances decay to a smaller steady state but in case (B) we observe a higher asymptotic value. A higher scatter of $\langle x \rangle$ values may occur at

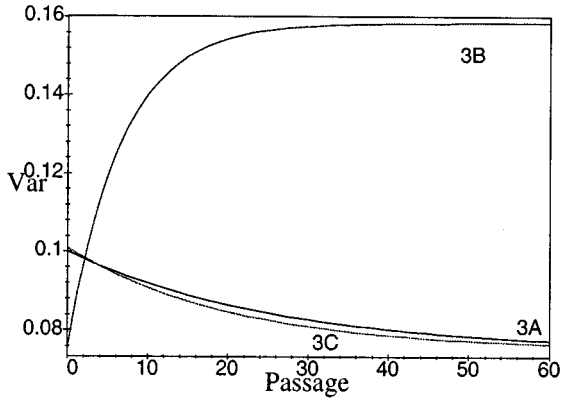

FIG. 2. Using the best fit parameter values of Fig. 1 we plot the variance for each of the curves above. The one corresponding to Fig. 1(B) has a higher scatter at large fitness values.

large fitness levels when the ratio μ/λ is high.

The exponential fit (1) proposed for the experimental findings of VSV seems to work well for low to medium fitness values. However, the fit does not seem to correctly represent the virus fitness evolution at long times (fitness defined as a rate cannot grow indefinitely), although more data points are needed to make this point clear. The mean field theory [8] previously presented also has similar behavior with two stages of linear growth in fitness. Their constant diffusion model with no drift term requires the inclusion of P_c, a threshold value below which selection vanishes. Without this constraint, the average fitness will diverge in finite time. A more recent paper [11] redefines this model by including a neighborhood two-point correlation between parent and offspring. The more refined model correctly predicts the asymptotic state for $\langle x \rangle$ without introducing P_c, but this asymptotic state also grows linearly in time.

The one-dimensional mean field model used here and in Ref. [8] simulates in broad strokes the distribution of a quasispecies as selection moves the population toward higher fitness levels. This picture may require modifications at high fitness values. The high scatter of data [12] in this regime may signal the presence of variable sharp peaks and valleys superimposed to a generally robust smoother landscape. This noisy background is one of the consequences of contracting the high-dimensional sequence space into a one-dimensional fitness space.

The presence of drift suggests that deleterious mutations expressed at the phenotypic level are not only unavoidable but also necessary to stabilize virus populations. The reason why a smooth fitness landscape may work at all can be traced to the knowledge that although the variability in sequence space is enormously large, only a limited number of mutations induces an adaptive advantage to the population [13]. Further experimental study on virus evolution at longer times would be necessary to determine how significant the drift velocity is.

We wish to thank H. Levine and J. J. Holland for several helpful comments. Partial support for this project was provided by the Clayton Fund, Inc.

[1] A. J. Levine, *Viruses* (W. H. Freeman, New York, 1992).
[2] W. M. Fitch, Mol. Phylog. Evol. **5**, 247 (1996).
[3] M. Eigen, Naturwissenschaften **58**, 465 (1971).
[4] M. Eigen and P. Schuster, *The Hypercycle—A Principle of Natural Self-Organization* (Springer, Heidelberg, 1979).
[5] M. Eigen, J. McCaskill, and P. Schuster, J. Phys. Chem. **92**, 6881 (1988).
[6] P. Schuster, Complexity **2**, 22 (1996).
[7] I. S. Novella, E. A. Duarte, S. F. Elena, A. Moya, E. Domingo, and J. J. Holland, Proc. Natl. Acad. Sci. USA **92**, 5841 (1995).
[8] L. Tsimring, H. Levine, and D. Kessler, Phys. Rev. Lett. **76**, 4440 (1996).
[9] M. Kac, in *Selected Papers on Noise and Stochastic Processes*, edited by N. Wax (Dover Publications, New York, 1954), p. 295.
[10] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 1981).
[11] D. Kessler, H. Levine, D. Ridgway, and L. Tsimring, J. Stat. Phys. **87**, 519 (1997).
[12] E. A. Duarte, I. S. Novella, S. Ledesma, D. K. Clarke, A. Moya, S. F. Elena, E. Domingo, and J. J. Holland, J. Virol. **67**, 4295 (1994).
[13] M. A. Huynen, P. F. Stadler, and W. Fontana, Proc. Natl. Acad. Sci. USA **93**, 397 (1996).