Quasispecies evolution of finite populations

Yi-Cheng Zhang
Institut de Physique Théorique, Université de Fribourg, CH-1700, Switzerland
(Received 9 January 1997)

We analyze a differential stochastic model to study quasispecies evolution. Using a variational method we show that the apparently smooth equation naturally gives rise to intermittent behavior—punctuated evolution. We also show that a finite population puts severe constraints on the evolution modes, and that the role played by stochastic noise is emphasized by the random fitness landscape. [S1063-651X(97)51204-9]

PACS number(s): 05.40.+j, 05.70.Ln

Punctuated evolution is a conjecture proposed by paleontologists Gould and Eldredge [1] more than two decades ago, to explain the fossil records of life forms. The evolution of the species is argued to follow an intermittent pattern with long stasis interrupted by short activity bursts. It would be desirable to put this on an analytic basis. Bak and Sneppen took the first steps in this direction [2]. In their prototype model they showed that under very general conditions of interspecies interaction, evolution can indeed be punctuated, and with a rich fractal spatial temporal structure. A more traditional approach is based on studying the evolution of species on a fitness landscape. We shall see that the evolution problem is equivalent to the diffusion in a random potential, albeit in genotype space rather the physical one. The nonperturbative method [3] developed previously can be readily generalized here.

In this work we shall show that the evolution of a single ''quasispecies'' in a random fitness landscape is also punctuated. Our starting point is closely related to the quasispecies approach of Eigen and Schuster [4]. We generalize to take into account stochastic noise and finite populations, which will play an important role. We shall show that, though the underlying differential equation is apparently smooth, it gives rise naturally to *punctuated* solutions. Following Eigen and Schuster's notation, we consider a genotype represented by a binary sequence of length N, which can have 2^N distinct configurations. The total population is assumed to be very large but finite (M). We denote by x_i the number of individuals having the ith genotype ($i=1,\ldots,2^N$), and by $d(i,j)$ the Hamming distance between two genotypes i and j. The random fitness landscape is taken to be independent and Gaussian with a unity variance denoted by V_i. Though our analysis can generalize straightforwardly to other distributions, we shall limit ourselves here to the simplest case to serve the purpose of a prototype model. Our evolution equation is

$$\dot{x}_i=\sum_{j=1}^{2^N} W_{ij}x_j+[V_i-E(t)]x_i+\eta_i(t)\sqrt{x_i}, \qquad (1)$$

where $W_{ij}=\mu^{d(i,j)}(1-\mu)^{N-d(i,j)}$ is the mutation rate between the sequences i and j, μ is the mutation rate per bit and per generation, and is assumed to be very small (typically 10^{-8}). Since we are interested in the relative population only, the above equation is to be normalized. For normalization the factor $E(t)$ is introduced to impose that the total population is always constant, $\Sigma_i x_i=M$. In the following we shall simply ignore $E(t)$, whose freedom is used to ensure the normalization. $\eta_i(t)$, the stochastic noise, assumes ± 1 with equal probability, for each i and t independently. Note that $x_i \geqslant 1$, as there cannot be less than one individual for a given sequence.

The evolution equation resembles the original Eigen-Schuster equation for quasispecies [4]. The first term on the right-hand side of Eq. (1) is purely due to mutation, and the second to reproductive advantages (disadvantages). We find the separation convenient, and it should not affect essential features of the Eigen-Schuster equation. The stochastic contribution η is new, and it can be derived following the reasoning of the Kimura evolution equation on a flat landscape [5]. The Kimura equation applies for two species only. It can be shown that the generalization to many species takes a particularly simple form if we work in the Langevin framework (the Kimura equation corresponds to the Focker-Planck framework). A detailed discussion will be presented elsewhere.

The role played by the stochastic term is best illustrated by the limiting case, where the landscape is flat ($V_k=$const). Equation (1) without the η term would imply unlimited diffusion, and this is not correct. With it, Eq. (1) predicts that the population is nevertheless confined in a limited region in the genotype space. It can be shown that the typical Hamming distance of such a population is $\sqrt{M\mu}$. This has already been discussed using slightly different language elsewhere [6] for the flat physical space, and its generalization for the *flat* genotype space [7] has also been worked out. In this work we consider another limit, where the fluctuations of a random landscape dominate. It can be shown that when both stochastic and landscape fluctuations are present, the latter are always dominant. Therefore in the following we first consider the evolution without η noise.

First let us consider what happens if the population is *infinite*. On the random landscape, there is a global optimum that typically (using extreme statistics for *Gaussian* variables of the sample size 2^N) has a fitness value $V_{\text{global}}=\sqrt{N\ \ln 2}$. We see that the global optimum increases rather quickly as a function of the sequence length. Let us assume that the population is initially concentrated around a master sequence at the configuration $(00\ldots0)$. It decays exponentially with the Hamming distance d, on average [3]. Within the region lim-

ited by a given distance d, the best fitness peak can be easily estimated for Gaussian variables, $V_d=\sqrt{d\ \ln N}$. We denote by V_0 the fitness at the origin, which is of order unity. We would like to know the location of the current master sequence—in other words how large d is—after a given time t. We denote the number having the current master sequence by x, which is a function of d and t. We make the following estimate for Eq. (1), in the strong localization limit [3]. The number of individuals x should be maximal for the master sequence at d:

$$x\sim M\mu^d\exp t(V_d-V_0)=\max. \tag{2}$$

The factor $M\mu^d$ represents the exponential decay away as a function of d from the master sequence. There is an exponential increase for a sequence d distance away with an advantage fitness V_d. It should compete with the old sequence of fitness V_0. For t fixed, d is larger, the fitness advantage is better (i.e., larger V_d), but the term μ^d is smaller. There is an optimal d which maximizes the above expression as a function of t, hence we can identify the current master sequence. Variation with respect to d leads to the relation

$$d(t)=t^2\ln N/(\ln\mu)^2. \tag{3}$$

Note that the constraint $d(t)\leqslant N$ implies that after time $t\approx|\ln\mu|\sqrt{N/\ln N}$, the global optimum will be reached.

The next question is whether this movement is smooth or intermittent in time. Suppose that the current configuration of the master sequence has fitness V, and the next configuration has $V'=\sqrt{d\ \ln N}$; d is the Hamming distance between the two configurations. In order for the master sequence to move away from the current configuration, the next fitness peak must be better. This already requires that $V'>V$. Let us estimate better. If the competing peak is on the verge of winning, its contribution must be comparable to that of the current peak. We thus have the estimate

$$M\mu^d\exp tV'\approx\exp tV. \tag{4}$$

The left-hand side denotes the exponential advantage as well as the cost for reaching out. At the same time the current peak also grows exponentially, the right-hand side. This gives us a relation for t,

$$t\approx d|\ln\mu|/(V'-V). \tag{5}$$

In principle all better peaks can win: either one with an infinitesimal advantage which is near, or one with a very large advantage that is far away. However, both need very long times to be realized. In reality only the peak requiring the *least* time can win. As a consequence all candidate peaks whose fitness lies between V and the winning peak will be skipped, while those with *higher* fitness will still remain candidates, to be examined for the next evolutionary step. We minimize the above t with respect to varying d. We find that the winning peak has the fitness $V'\approx 2V$, to the leading order approximation ($d>1,|\ln\mu|>1$). Therefore we conclude that the master sequence moves in an intermittent fashion, so that each new peak has about twice the fitness of that of the current peak. It is easy to verify that the global optimum will be reached after a finite number of jumps $n\approx\ln N\ln 2/(2\ln 2)$.

The above result cannot, however, be correct for a finite population. As a matter of fact, for any realistic population, no matter how large, the above result is grossly wrong. The short answer to this puzzle is that while M can be very large, $\ln M$ can hardly be; the latter is the only factor that appears in the calculation. Another way to see this is that normally we have $M\ll 2^N$; i.e., the number of individuals is much smaller than that of the total possible genotypes, for any realistic population.

For a finite population M the constraint $x_i\geqslant 1$ or $M\mu^d\geqslant 1$ imposes that the distance from the master sequence is limited. The population can only explore a region limited by $d_M\approx\ln M/|\ln\mu|$, and the best peak within the region $V_M\approx\sqrt{\ln M\ln N/|\ln\mu|}$ is much smaller than the global optimum. The above senario is nevertheless obtained here: each time the fitness value of a new peak is about double of that of the old peak, the population makes a few jumps $[n\approx\ln(\ln M\ln N/|\ln\mu|)/(2\ln 2)]$, and ends up in the local optimum. After that the population quickly settles down around the local optimum, and there are still interesting movements in search of better peaks, albeit at a much slower pace.

Let us consider a population already settled around a local optimal peak, the deterministic part of Eq. (1) cannot make further moves. The stochastic noise is in to help. The effective contribution from noise is to make the population drift away from its equilibrium position. From now on we want to relax the contraint $x_i\geqslant 1$ to include all positive real values. The interpretation for x_i is slightly changed: it should be proportional to the *probability* of finding an individual. For instance, for a highly unlikely sequence i $1\gg x_i>0$, it will appear in the population if we wait long enough. In the following we just interpret x_i as a probability; noise makes this possible but we shall not use it explicitly.

From the above reasoning we know that far away from the current optimal peak, $M\mu^d$ can be smaller than unity. However, since it is proportional to the probability of finding an individual at distance d, we can always have $tM\mu^d\geqslant 1$; t is the waiting time. From this estimate we have a relation for d, the drift distance from the local optimum as a function of time,

$$d=\frac{\ln M+\ln t}{|\ln\mu|}. \tag{6}$$

This is much slower than that for an infinite population, Eq. (3). Let us call this mode of motion noise assisted. We need an extremely long time to reach the global optimum, and we cannot know for certain that it will ever be achieved in the present framework.

In the noise-assisted mode the motion of the master sequence is also intermittent. In place of Eq. (4), for a finite population we now have

$$t_d^{-1}\exp(t-t_d)V'\approx\exp tV, \tag{7}$$

where $t_d=1/(M\mu^d)$ is the time for the population to find a better peak V_d, d distance away from the present peak V, $1/t_d$ is the probability for this to happen. The above relation

is expressed in exponential form. The only difference with respect to Eq. (4) is the retarded time $t-t_d$ in place of t. This is because of the fact that before a better peak can be found there is no reproduction. A similar variational analysis gives rise to the fitness improvement relation

$$V' \approx V + \frac{1}{2}\frac{\ln N}{|\ln \mu|} \tag{8}$$

to leading order, where condition $V \gg \ln N/|\ln \mu|$ is assumed. Compare the above with $V' \approx 2V$. The improvement for each jump is much smaller, and for this jump to be realized a long waiting time is needed [$t \approx (1/M)\mu^{-\ln N/4|\ln \mu|^2}$].

In this work we studied a prototype evolution model, generalized from the standard Eigen-Schuster model to include stochastic noise. Using variational methods developed in other branches of statistical physics, we show that punctuated solutions are inevitable, for noninteracting quasispecies evolving on a uncorrelated landscape. Finite populations introduce severe constraints on the evolution pace, and in reality all population should be considered finite. Evolution follows two distinct modes, both of which are punctuated: In the initial mode the population evolves faster and jumps are also larger, as if the population were infinite; this mode ends when the finite population limit is felt, and evolution enters a slow, so-called noise-assisted mode which is characterized by rare fluatuations around a finite region centered at a master sequence. Whereas the stochastic noise is negligible in the initial evolution mode because the pace is fast, it is the only driving force for later evolution.

[1] S. J. Gould and N. Eldredge, Paleobiology **3**, 114 (1977); Nature **366**, 223 (1993).
[2] P. Bak and K. Sneppen, Phys. Rev. Lett. **71**, 4083 (1993).
[3] Y.-C. Zhang, Phys. Rev. Lett. **56**, 2113 (1986); T. Halpin-Healy and Y.-C. Zhang, Phys. Rep. **254**, 215 (1995).
[4] M. Eigen and P. Schuster, *The Hypercycle-a Principle of Natural Self-Organization* (Springer-Verlag, Berlin, 1979), and references therein.
[5] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983).
[6] Y.-C. Zhang, M. Serva, and M. Polikarpov, J. Stat. Phys. **58**, 849 (1990).
[7] B. Derrida and L. Peliti, Bull. Math. Biol. **53**, 355 (1991).