Analysis of Evolution through Competitive Selection

Morten Kloster

Department of Physics, Princeton University, Princeton, New Jersey 08544, USA
NEC Laboratories America, 4 Independence Way, Princeton, New Jersey 08540, USA
(Received 15 February 2005; published 12 October 2005)

Recent studies of *in vitro* evolution of DNA via protein binding indicate that the evolution behavior is qualitatively different in different parameter regimes. I here present a general theory that is valid for a wide range of parameters, and which reproduces and extends previous results. Specifically, the mean-field theory of a general translation-invariant model can be reduced to the basic diffusion equation with a dynamic boundary condition. The simple analytical form yields both quantitatively accurate predictions and valuable insight into the principles involved.

DOI: 10.1103/PhysRevLett.95.168701 PACS numbers: 87.23.Kg, 87.10.+e

In modern biotechnology, evolutionary techniques have found many new applications; in particular, *in vitro* evolution has been widely used to evolve DNA [1], RNA [2], and proteins [3]. In order to improve the efficiency of such evolution, a quantitative understanding of the process would be helpful. Key questions include how the rate of evolution and the equilibrium genotype distribution depend on the main parameters of the process, such as the mutation rate, population size, and selection strength. Recently, evolution through competitive selection, or just "competitive evolution," has been introduced as a model of *in vitro* molecular evolution [4–6]. Unlike traditional "fixed fitness" evolution [e.g., [7–11]], competitive evolution has a well-behaved mean-field theory which allows accurate predictions of the evolution dynamics [4,5]. However, these analyses are valid only for limited parameter regions—either high mutation rate and weak selection [4], or low mutation rate and strong selection [5]. In this Letter, I introduce a translationally invariant model of competitive evolution, which through appropriate transformations is reduced to the basic diffusion equation with a dynamic boundary condition, and ultimately to a universal set of equations. This formulation is valid for a broad parameter range, and its simplicity makes it a powerful tool both for quantitative predictions and for qualitative understanding of the evolution process. This Letter includes some key results; additional results and details are given elsewhere [12].

The dynamics of an evolution process depend crucially on the fitness (or phenotype) landscape. Analytical results typically rely on a simple landscape, which leads many authors to consider *specific* simple models [4,7,8,13]. I instead use the general assumption that the rate of mutations that change the phenotype (binding energy) by some amount ϵ is the same for all molecules; i.e., the model is translationally invariant in phenotype space—this corresponds to the limit of infinitely long genomes/DNA molecules [12]. While this is a strong assumption, the resulting formulation gives accurate results also for finite systems [see below; [12]].

The model.—The present model is inspired by the evolution experiment performed by Dubertret *et al.* [14]. An initial population of N DNA molecules is subjected to consecutive cycles of evolution, each of which consists of amplification and mutation followed by selection through protein binding. First, the population is amplified by a factor K. Each base of each DNA molecule is then mutated with some probability which may depend on the position and type of the base only; i.e., mutations are independent. Finally, the strongest protein binders are selected: each molecule is kept (bound) with probability $P[E(S)] = \frac{1}{1+e^{\beta[E(S)-\mu]}}$, where $E(S)$ is the binding energy of DNA sequence S, $\beta = \frac{1}{k_B T}$, and the chemical potential of unbound protein, μ, is tuned such that the expected number of selected molecules is N: $\langle P[E(S)]\rangle = 1/K$. This Letter only covers the zero temperature limit $T = 0$, for which the selection function becomes the step function $P(E) = \Theta(\mu - E)$; this is a good approximation in many cases [4,15,16]. Typically, the average binding energy improves through the evolution process; i.e., the evolution rate $\frac{d}{dt}\langle E(S)\rangle$ is negative.

For sufficiently large populations, a mean-field approach is valid [4,5]; i.e., we can use deterministic equations for the population density $n_t(E)$, where $\int_{-\infty}^{E} n_t(E')dE'$ is the fraction of the population that has binding energy $\leq E$ at time t. While the actual evolution proceeds in discrete time steps, the following continuum differential equation is a good approximation [12]:

$$\partial_t n_t(E) = \int_{-\infty}^{\infty} p(\epsilon)[n_t(E-\epsilon) - n_t(E)]d\epsilon + k n_t(E) \quad (1)$$

with the boundary condition $n_t[\mu(t)] = 0$. The integral term in Eq. (1) describes the mutation process, where $p(\epsilon)$ is the rate of mutations that change the binding energy by ϵ, while the second term gives exponential growth and corresponds to the amplification process, with $k = \ln(K)$. The boundary condition enforces perfect selection by removing all molecules with binding energy above the

chemical potential $\mu(t)$, which is determined by the normalization condition for the population, $\int_{-\infty}^{\mu(t)} n_t(E)dE = 1$.

A simple formulation.—Assume there exists a mutation that can improve the binding energy, i.e., $p(\epsilon) > 0$ for some $\epsilon < 0$. It is then easily shown that, for any $k > 0$, there is a solution

$$n_t^0(E) \propto (E - c_0 t)e^{\alpha_0(E-c_0 t)}, \tag{2}$$

with α_0 and c_0 given by the conditions

$$-c_0\alpha_0 = \int_{-\infty}^{\infty} (e^{-\alpha_0\epsilon} - 1)p(\epsilon)d\epsilon + k \tag{3}$$

$$c_0 = \int_{-\infty}^{\infty} \epsilon e^{-\alpha_0\epsilon} p(\epsilon)d\epsilon. \tag{4}$$

For a general solution with the same exponential factor, $n_t(E) = \tilde{n}_t(E)e^{\alpha_0(E-c_0 t)}$, Eq. (1) simplifies to

$$\partial_t \tilde{n}_t(E) = \int_{-\infty}^{\infty} e^{-\alpha_0\epsilon} p(\epsilon)[\tilde{n}_t(E - \epsilon) - \tilde{n}_t(E)]d\epsilon \tag{5}$$

$$\approx -c_0\partial_E \tilde{n}_t(E) + \gamma\partial_E^2 \tilde{n}_t(E), \tag{6}$$

where $\gamma = \int \frac{\epsilon^2}{2} e^{-\alpha_0\epsilon} p(\epsilon)d\epsilon$. Equation (5) describes the mean field of a random walk with an exponentially rescaled mutation/transition rate, $\tilde{p}(\epsilon) = e^{-\alpha_0\epsilon}p(\epsilon)$, and c_0 and γ are the corresponding drift and diffusion coefficients, respectively. If ϵ_0 is the characteristic energy scale for the mutations, then the condition $\alpha_0\epsilon_0 \ll 1$ ($\alpha_0\epsilon_0 \gg 1$) specifies the regime of high (low) mutation rate/weak (strong) amplification for which the results in [4] ([5]) apply (see below).

Changing coordinates to $\tilde{E} = E - c_0 t$ now yields

$$\partial_t \tilde{n}_t(\tilde{E}) \approx \gamma\partial_{\tilde{E}}^2 \tilde{n}_t(\tilde{E}), \qquad \tilde{n}_t[\tilde{\mu}(t)] = 0, \tag{7}$$

where $\tilde{\mu}(t) = \mu(t) - c_0 t$. The only nontrivial part left is the normalization condition

$$\int_{-\infty}^{\tilde{\mu}(t)} \tilde{n}_t(\tilde{E})e^{\alpha_0\tilde{E}}d\tilde{E} = 1. \tag{8}$$

The remaining parameters in Eqs. (7) and (8) are α_0 and γ, and appropriate rescaling of energy and time yields a *universal* set of equations, which in turn gives a universal shape for the population distribution (see below). Additionally, the simplicity of the new formulation—the basic diffusion equation with a dynamic boundary condition—allows us to apply all our knowledge and experience about the diffusion equation to this evolution process. As shown below and in [12] we can, through simple arguments, use this formulation to find accurate corrections to the various approximations/assumptions we have made. The analytical predictions are confirmed with simulations, using appropriate simple models [12] with a wide range of parameters.

The propagating pulse.—As $\tilde{n}_t$ obeys a simple diffusion equation, it will be very smooth, except possibly at very early times. The population density $n_t(E) = \tilde{n}_t(\tilde{E})e^{\alpha_0\tilde{E}}$ contains a factor that varies exponentially with $\tilde{E}$, thus a linear expansion of $\tilde{n}_t$ around the boundary gives a good estimate of the population size:

$$\text{Population size}\,(\tilde{n}_t) \approx \frac{Ne^{-\alpha_0\tilde{\mu}(t)}}{\alpha_0^2}\partial_{\tilde{E}}\tilde{n}_t(\tilde{E})|_{\tilde{E}=\tilde{\mu}(\tilde{t})}. \tag{9}$$

The normalization condition (8) then relates the position of the boundary to the slope of $\tilde{n}_t(\tilde{E})$ at the boundary:

$$\partial_{\tilde{E}}\tilde{n}_t(\tilde{E})|_{\tilde{E}=\tilde{\mu}(\tilde{t})} \approx -\alpha_0^2 e^{-\alpha_0\tilde{\mu}(t)}. \tag{10}$$

If $\tilde{n}_t$ is reasonably smooth on the characteristic energy scale $1/\alpha_0$, then the linear expansion around the boundary will in fact give a good description of the full population at time t. Applying Eq. (10), we find

$$n_t(E) \approx \alpha_0 f_0(\alpha_0[E - \mu(t)]), \tag{11}$$

where

$$f_0(x) = \begin{cases} -xe^x & x \le 0 \\ 0 & x > 0 \end{cases} \tag{12}$$

is the universal pulse shape (Fig. 1). The dynamics of the system can now be described as the motion $\mu(t)$ of a pulse of almost constant shape. For $t \to \infty$ (and mean field), the pulse propagates at the constant speed c_0.

Note that Eq. (2) describes a population that extends to arbitrarily low energies. If we impose a second boundary condition $n_t[\mu(t) - \Delta] = 0$ and require $n_t(E) > 0$ for $\mu(t) - \Delta < E < \mu(t)$, then there is a *bounded* solution $n_t(E) \propto \sin\{b[\mu(t) - E]\}e^{\alpha_b[\mu(t)-E]}$ that moves at constant speed c_b, where $b = b(\Delta) = \pi/\Delta$ and

$$c_b \approx c_0 + \gamma b^2/\alpha_0. \tag{13}$$

In the limit $b \to 0$ we recover the unbounded solution. Similarly, a population that at $t = 0$ is located in a small energy range will initially improve slower than the maximal rate c_0—the lowest order correction is $\propto t^{-1}$ [12].

Evolution rate.—While we cannot in general solve Eqs. (3) and (4) for α_0 and c_0 analytically, they are manageable for certain models. For a simple discrete infinite-length model [12], the mutation dynamics can be fully described by the diffusion coefficient D and the drift speed v [4]:

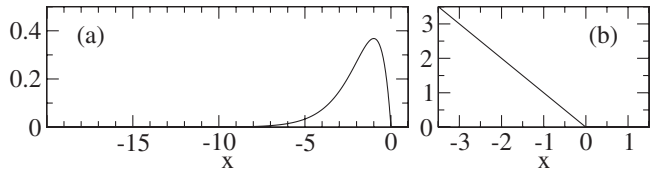

FIG. 1. (a) The universal shape $f_0(x)$ of the propagating pulse. (b) The rescaled formulation $\tilde{f}_0(x) = f_0(x)e^{-x}$ for which the mean-field dynamics are purely diffusive.

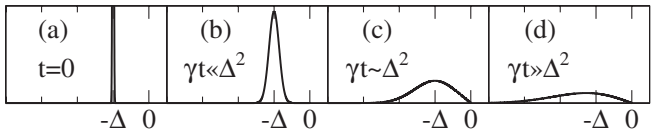

FIG. 2. An initial perturbation given by a delta function (a) will spread as a Gaussian (b) until it encounters the boundary. At some time $\sim\Delta^2$ the slope at the boundary will be maximal $\sim 1/\Delta^2$ (c), after which it decays as $\sim t^{-3/2}$ for $t \to \infty$ (d).

$$p(\epsilon) = \left(D + \frac{v}{2}\right)\delta(\epsilon - 1) + \left(D - \frac{v}{2}\right)\delta(\epsilon + 1), \quad (14)$$

where $\delta(\cdot)$ is Dirac's delta function. D, v, and k can be written as simple functions of c_0, α_0, and γ [12], and for sufficiently small $\alpha_0 \approx \sqrt{k/D}$ we find the evolution rate c_0 as a power series in k/D and v/D:

$$c_0 = v - 2\sqrt{kD} + \frac{vk}{6D} - \frac{k^{3/2}}{12\sqrt{D}} + \frac{v^2k^{3/2}}{36D^{5/2}} + \cdots. \quad (15)$$

The first two terms are the solution in [4] ($k = \frac{\phi^{-1}-1}{\tau}$). For low mutation rate/strong selection, however, let us write $p(\epsilon) = p_+\delta(\epsilon - 1) + p_-\delta(\epsilon + 1)$. For $\alpha_0 \gg 1$ the exponential rescaling $\tilde{p}(\epsilon) = e^{-\alpha_0\epsilon}p(\epsilon)$ makes p_+ irrelevant:

$$c_0 \approx \frac{-k}{\alpha_0 - 1} \approx \frac{-k}{\ln(-c_0/p_-) - 1}, \quad (16)$$

which is a more accurate, general version of Eq. (6) in [5]. These results confirm qualitative differences between different parameter regimes: for $\alpha_0\epsilon_0 \ll 1$ the evolution rate depends strongly on the mutation rate, while for $\alpha_0\epsilon_0 \gg 1$ the dependence is only logarithmic.

Finite population size.—For *in vitro* evolution, an important question is how large the evolving population must be to achieve good results. Peng *et al.* [4] estimate the effect of the population size on the evolution rate through a cutoff procedure [17] that ignores amplification wherever $n(E) < \frac{1}{N}$. While the result matches simulations fairly well, this might not be the correct cutoff [17]. A better approach is to explicitly apply the key criterion: when will mean-field theory fail? Consider the addition of a single DNA molecule with binding energy Δ below the selection threshold μ_0. In mean-field theory, the transformed initial perturbation $\delta\tilde{n}_0(\tilde{E}) = \frac{1}{N}e^{\alpha_0(\Delta-\tilde{\mu}_0)}\delta(\tilde{E} - [\tilde{\mu}_0 - \Delta])$ will evolve according to Eq. (7) with a fixed boundary $\tilde{\mu}(t) = \tilde{\mu}_0$. Simple scaling yields a maximal slope at the boundary $\sim \frac{1}{N}e^{\alpha_0(\Delta-\tilde{\mu}_0)}/\Delta^2$ after a time $t_\Delta^{\max} \sim \Delta^2/\gamma$ (Fig. 2). From Eq. (9), the *population size perturbation* is then $\delta N_\Delta^{\max} \sim e^{\alpha_0\Delta}/(\alpha_0\Delta)^2$ (for $\alpha_0\Delta \gg 1$). For Δ large enough, $\delta N_\Delta^{\max} > N$; i.e., the descendants of *one* molecule will eventually replace the whole population, and the selection threshold must "jump" to keep the population size constant. This clearly violates mean field, and $\delta N_{\Delta_N}^{\max} = N$ yields a lower boundary $\mu_2(t) = \mu(t) - \Delta_N$ beyond which mean field fails. Our estimate of the evolution rate is then the propagation speed $c_{b(\bar{\Delta}_N)}$ of a pulse of length $\bar{\Delta}_N = \Delta_N - t_{\Delta_N}^{\max}(c_{b(\bar{\Delta}_N)} - c_0)$, which corrects for the resulting motion of the boundary $\tilde{\mu}(t)$: [18]

$$c_N \approx c_0 + \frac{\gamma\alpha_0\pi^2}{\{\ln[N\ln^2(N)] - \pi^2/6\}^2}. \quad (17)$$

This estimate is very accurate (Fig. 3, "Theory")—it ignores the contribution from mean-field-violating events, but is far better than the estimate in [4] ("$1/N$ cutoff"). Figure 4 shows the difference between the actual value of $\mu(t)$ and our estimate $\mu_0 + c_N t$ as a function of time for two simulations with different populations sizes. These plots support our discussion above—the difference is almost constant for long periods but occasionally makes a large jump; i.e., there is a mean-field-violating event.

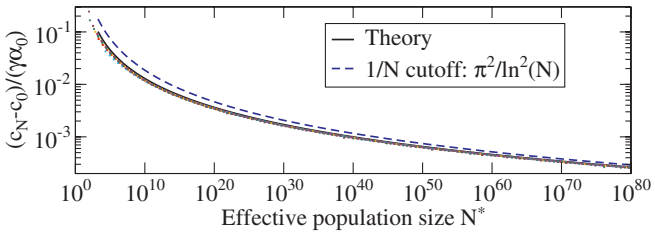

FIG. 3 (color online). Deviation from mean-field propagation speed as a function of effective population size N^* [see text, [12]]; theory results and scaling collapse of simulations (small circles).

Finite size.—In a real, finite system, the mutation rate distribution $p(\epsilon)$ will not be the same for all molecules. However, as long as $p(\epsilon)$ depends primarily on the binding energy E of the molecule and varies reasonably slowly with E on the characteristic energy scale $1/\alpha_0$, we expect the results found above to be good local approximations to the actual dynamics. Indeed, in [4], the authors argued that a population initially far from equilibrium will form a pulse of almost constant shape that exponentially approaches its equilibrium position, in excellent qualitative agreement with simulations.

Using the tools developed here, we can find good *quantitative* predictions also for finite systems [more in [12]]. For the discrete, finite model described in [4,19], $v = v(r)$ and $D = D(r)$ in Eq. (14) depend on the mismatch number $r = E/\epsilon_0$. For $t \to \infty$, the propagating pulse and its bound-

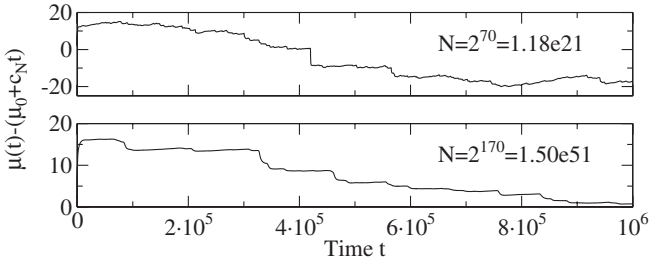

FIG. 4. Motion of boundary $\mu(t)$ relative to estimate [Eq. (17)] for simulations with different population sizes.

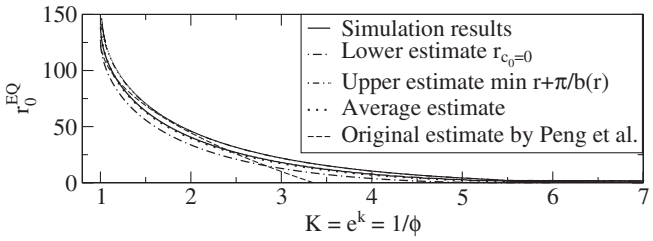

FIG. 5. Equilibrium position for the boundary as a function of amplification K; estimates and simulation results, as discussed in the text. Number of bases $L = 170$; mutation rate/base $\nu = 0.01$.

ary r_0 will reach an equilibrium position [4]. A simple estimate of r_0^{EQ} is then the value of r for which the propagation speed $c_0[v(r), D(r)]$ is 0 [from Eqs. (3) and (4)]:

$$k = 2D(r) - \sqrt{4D(r)^2 - v(r)^2} \tag{18}$$

—this equals the ''second region formula'' in [19]. As the entire population has mismatch number below r_0, the solution to Eq. (18) is actually a lower bound for r_0^{EQ}.

The equilibrium condition $c_b = 0$ for a *bounded* solution yields

$$k = 2D(r) - \sqrt{4D(r)^2 - v(r)^2}\cos(b). \tag{19}$$

By solving the above equation for b, we find a population that lies between r and $r + \pi/b(r)$; i.e., the entire population has *higher* mismatch number than the position at which $c_b[v(r), D(r)] = 0$. Its upper bound $r + \pi/b(r)$ is thus an upper bound for r_0^{EQ} [20], and minimizing over r yields the best upper bound. The average of the lower and upper bound is our final estimate:

$$r_0^{\mathrm{EQ}} \approx \frac{1}{2}(r_{c_0=0} + \min_{r > r_{c_0=0}}[r + \pi/b_{c_b=0}(r)]). \tag{20}$$

As shown in Fig. 5, this estimate is very accurate for most values of $\phi^{-1} = K = e^k$; there are significant deviations only for very low amplifications K [12]. Figure 5 also includes the estimate found in [4] for comparison.

In this Letter, we have seen how the assumption of a translationally invariant mutation rate leads to a powerful formulation of competitive evolution that can give very accurate predictions in many different situations. Most of these results will be locally accurate for any system for which the mutation rates $p(\epsilon)$ vary reasonably slowly with E on the characteristic length scale $1/\alpha_0$. The effective dynamics of the evolution process are to a good approximation captured by the rescaled mutation rate $\tilde{p}(\epsilon) = e^{-\alpha_0 \epsilon} p(\epsilon)$: far from equilibrium, or for small populations, this is the rate at which mutations are fixed in the population, while for a sufficiently large population close to equilibrium, this gives the relative occurrences of different bases.

One significant result is the accurate prediction of the limits of mean-field theory for a finite population and the way in which it fails—when a molecule by chance gets sufficiently far ahead of the general population, the population is quickly replaced by the descendants of that single molecule, evidenced by a jump in the selection threshold. This limits the genetic variation that can build up during competitive evolution: whenever such a jump occurs, essentially all prior genetic variation is wiped out. For instance, for an organism that reproduces mostly by asexual growth but with periodic sexual reproduction stages, this sets a lower limit for the frequency of sexual reproduction necessary to maintain genetic diversity.

In summary, I have presented a simple model that accurately captures the key qualitative and quantitative features of asexual competitive evolution on a smooth landscape. This allows prediction of the performance of competitive evolution for a wide range of parameters for any given, reasonably smooth energy landscape. Additionally, this forms a baseline against which the performance of more complex evolutionary procedures, e.g., using recombination [6], can be compared.

The author would like to thank Rahul Kulkarni for helpful comments.

[1] J. A. Bittker, K. J. Phillips, and D. R. Liu, Curr. Opin. Chem. Biol. **6**, 367 (2002).
[2] L. F. Landweber, Trends Ecol. Evol. **14**, 353 (1999).
[3] *Evolutionary Protein Design*, edited by F. H. Arnold (Academic, New York, 2001).
[4] W. Peng, U. Gerland, T. Hwa, and H. Levine, Phys. Rev. Lett. **90**, 088103 (2003).
[5] M. Kloster and C. Tang, Phys. Rev. Lett. **92**, 038101 (2004); M. Kloster and C. Tang, cond-mat/0301372.
[6] W. Peng, H. Levine, T. Hwa, and D. A. Kessler, Phys. Rev. E **69**, 051911 (2004).
[7] J. Felsenstein, Genetics **78**, 737 (1974).
[8] G. Woodcock and P. G. Higgs, J. Theor. Biol. **179**, 61 (1996).
[9] S. P. Otto and N. H. Barton, Evolution (Lawrence, Kansas) **55**, 1921 (2001).
[10] D. A. Kessler, H. Levine, D. Ridgway, and L. Tsimring, J. Stat. Phys. **87**, 519 (1997).
[11] I. M. Rouzine, J. Wakeley, and J. M. Coffin, Proc. Natl. Acad. Sci. U.S.A. **100**, 587 (2003).
[12] M. Kloster, q-bio/0408027.
[13] D. Ridgway, H. Levine, and D. A. Kessler, J. Stat. Phys. **90**, 191 (1998).
[14] B. Dubertret, S. Liu, Q. Ouyang, and A. Libchaber, Phys. Rev. Lett. **86**, 6022 (2001).
[15] M. Kimura and J. F. Crow, Proc. Natl. Acad. Sci. U.S.A. **75**, 6168 (1978); J. F. Crow and M. Kimura, Proc. Natl. Acad. Sci. U.S.A. **76**, 396 (1979).
[16] U. Gerland and T. Hwa, J. Mol. Evol. **55**, 386 (2002).
[17] E. Brunet and B. Derrida, Phys. Rev. E **56**, 2597 (1997).
[18] Numerically, $t_\Delta^{\max} \approx \Delta^2/(6\gamma)$.
[19] E. Cohen and D. A. Kessler, cond-mat/0301503.
[20] Using a bounded population can only increase the bound.