Recombination Dramatically Speeds Up Evolution of Finite Populations

Elisheva Cohen,[1] David A. Kessler,[1] and Herbert Levine[2]

[1]*Department of Physics, Bar-Ilan University, Ramat-Gan, IL52900 Israel*

[2]*Center for Theoretical Biological Physics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0319, USA*

(Received 13 October 2004; published 9 March 2005)

We study the role of recombination, in the form of bacterial transformation, in speeding up Darwinian evolution. This is done by adding a new process to a previously studied Markov model of evolution on a smooth fitness landscape; this new process allows alleles to be exchanged with those in the surrounding medium. Our results, both numerical and analytic, indicate that, for a wide range of intermediate population sizes, recombination dramatically speeds up the rate of evolutionary advance.

DOI: 10.1103/PhysRevLett.94.098102 PACS numbers: 87.23.Kg, 02.50.Ga

Recombination of genetic information is a common evolutionary strategy, both in natural systems [1] as well as *in vitro* molecular breeding [2]. This idea is also employed in genetic programming [3], a branch of computer science which aims to evolve efficient algorithms. Given all this, it is surprising that we still do not have a good understanding of the conditions under which the benefits of recombination outweigh the inevitable costs.

Of course, there is a large literature on recombination, dating back to the ideas of Muller [4] and Kondrashov [5]. One line of recent work focuses on two loci genomes and considers whether or not recombination would be favored; possible mechanisms include (weak) negative epistasis (in which case the reproduction rate is not just the sum of independent contributions from each individual locus) or negative linkage disequilibrium (the lack of independence of the allele distribution in a finite population) or some combination thereof [6]. Others look at how the (static) genetic background in which a mutation arises will affect fixation probabilities (''clonal interference''), comparing these with or without recombination [7]. In both of these methods, only one or two mutations at a time are ''dynamic,'' a situation unlikely to be true for rapidly evolving microorganisms. In contrast, our analysis considers a large number of contributing loci.

In this Letter, we study recombination in the context of a simple fitness landscape model [8–10] which has proven useful in the analysis of laboratory scale evolution of viruses and bacteria [11]. The specific type of recombination we consider is based on the phenomenon of bacterial transformation [12]. Here, so-called genetically competent bacteria can import snippets of DNA from the surrounding medium; presumably these are then homologously recombined so as to replace the corresponding segment in the genome. This behavior is controlled by a cellular signaling system that ensures that recombination only occurs under stress. The details of the DNA importation and the aforementioned control has convinced most biologists [13,14] that transformation is an important survival strategy for many bacterial species.

Our model consists of a population of N individuals, each of which has a genome of L binary genes, $S_i = \{0, 1\}$. An individual's fitness, x, is the sum of the fitness contributions from each gene, S_i, so that $x = \sum_{i=1}^{L} S_i$. Evolution is implemented as a continuous time Markov process in which individuals give birth at rate x and die at random so as to maintain the fixed population size. Every birth allows for the daughter individual to mutate each of its alleles with probability μ_0 giving an overall genomic probability of $\mu = \mu_0 L$.

The last part of our Markov process implements recombination, parameterized by f_s, the recombination rate per unit time per locus. At rate $f_s L$, an individual has one of its genes deleted and substitutes a new allele from the surrounding medium; the probability of getting a specific S is just its proportional representation in the population. This mimics the transformation mechanism as long as the distribution of recently deceased (and lysed) cells is close to that of the current population [15]. In Fig. 1(a), we show simulation results for the ''velocity,'' $v = d\bar{x}/dt$, i.e., the rate of increase of the mean fitness (at one representative point on the landscape), for different order unity values of f_s, as a function of N. At very small population sizes, recombination has little effect since there is no population diversity upon which to act. Each of the curves rises sigmoidally to a saturation value at very large N which is again roughly independent of the recombination rate [see Fig. 1(b) and later]. Because the population scale for this rise is a strongly decreasing function of f_s, recombination at intermediate N can give a dramatic speedup of the evolution. It is worth mentioning that this basic result is qualitatively consistent with recent experiments [16] in microorganism evolution which demonstrate an increase in the efficacy of recombination as the population size is increased (starting from small); we should note, however, that the details of recombination in these experimental systems are different than those underlying bacterial transformation.

Can one understand these simulation results? At small N, we can appeal to previous results for this model [9] that

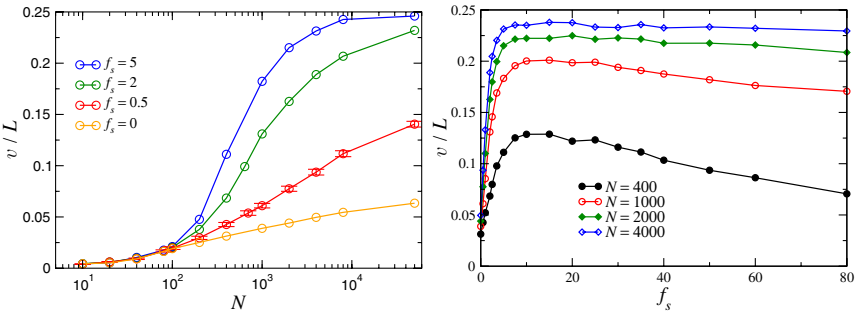

FIG. 1 (color online). Velocity (averaged over 200 samples) measured between $x = 95$ and $x = 105$ starting with an average fitness of 50. $L = 200$, $\mu = 0.1$. (a) v measured as a function of population size, N, for various f_s. Error bars are shown for one value of f_s, and are typical of all the data. (b) v measured as a function of f_s for various N. As N increases, the velocity saturates at an f_s-independent value.

show that the population variance scales as μN. One would therefore expect the small N breakpoint where the curves diverge to be roughly at $N = 1/\mu$; this is consistent with the data in Fig. 1(a) and we have checked this simple scaling with mutation rate (data not shown). Another small N effect becomes evident if the simulations are extended to much larger f_s values, as shown in Fig. 1(b). Now, the velocity begins a slow decline at too large f_s due to the recombination causing a loss of diversity as specific alleles go to fixation at various loci.

The behavior at larger N, past the inflection point of the velocity curves in Fig. 1(a) and in the rising segments of the curves in Fig. 1(b), is much less trivial. To enable us to write down an equation for the time derivative of $P(x)$, the fitness distribution function, we start by assuming that the subpopulation at some particular fitness x, of size $NP(x)$, has its 0 and 1 alleles uniformly distributed at each site of the genome. This means, of course, that selecting at random an allele at any site gives a chance x/L of getting $S = 1$ and $1 - x/L$ of getting $S = 0$. Then, in the infinite-population [mean-field (MFE)] limit:

$$\frac{dP_x(t)}{dt} = (x - \bar{x})P_x(t) + \mu\left[\frac{(x+1)^2}{L}P_{x+1}(t) + (x-1)\left(1 - \frac{x-1}{L}\right)P_{x-1}(t) - xP_x(t)\right] - f_s L\left[\left(1 - \frac{\bar{x}}{L}\right)\frac{x}{L}P_x(t) + \frac{\bar{x}}{L}\left(1 - \frac{x}{L}\right)P_x(t) - \left(1 - \frac{\bar{x}}{L}\right)\frac{x+1}{L}P_{x+1}(t) - \frac{\bar{x}}{L}\left(1 - \frac{x-1}{L}\right)P_{x-1}(t)\right]. \quad (1)$$

The first two terms are standard and reflect the birth-death process and the genomic mutation, respectively; the explicit form of the mutation term arises from considering the probability of an individual with fitness x giving birth (rate $\sim x$), mutating ($\sim\mu$), and hence going either up ($\sim(1 - x/L)$, the number of currently bad alleles) or down ($\sim x/L$, the number of good alleles). The last term is new and reflects the role of recombination. With the aforementioned assumption, the probability that an individual of fitness x will have its fitness altered is proportional to the recombination rate, $f_s L$, times the probability of either: (a) deleting a bad allele $(1 - x/L)$ and picking up a good one $(\bar{x}/L)$ or (b) deleting a good allele (x/L) and picking up a bad one $(1 - \bar{x}/L)$.

Before using this equation [and its modification for finite N effects, see Eq. (4) below] to analyze the numerical results, we need to test the underlying equi-distribution assumption. To do this, we generated a population of $N = 1000$ with $\bar{x} = 50$ and let it evolve using $f_s = 1$ for $L = 100$. We then measured the respective probabilities for a recombination event to increase or decrease the fitness, based on the fitness x of the chosen individual. As shown in Fig. 2, our theoretical expression has the correct functional dependence, although it overestimates these actual probabilities by roughly a fixed amount. This overestimate is due to the fact that individual sites have less diversity than is predicted, a remnant of the aforementioned loss-of-diversity effect. Notwithstanding the error (which we find

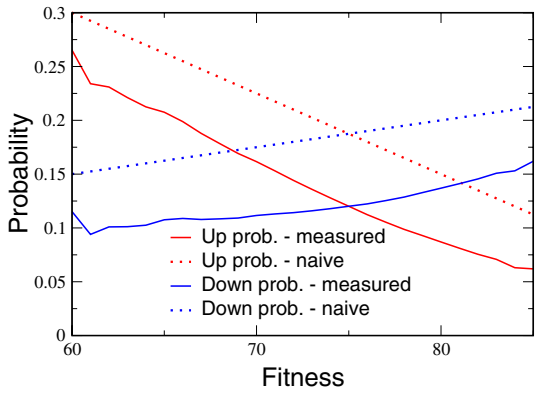

FIG. 2 (color online). The probability of making an up-and-down move due to recombination as a function of individual fitness when the average fitness of the population was 75. For this run, $N = 1000$, $L = 100$, $f_s = 1$, $\mu = 0.1$. The "naive" probabilities are those derived assuming equal distribution of alleles at every locus.

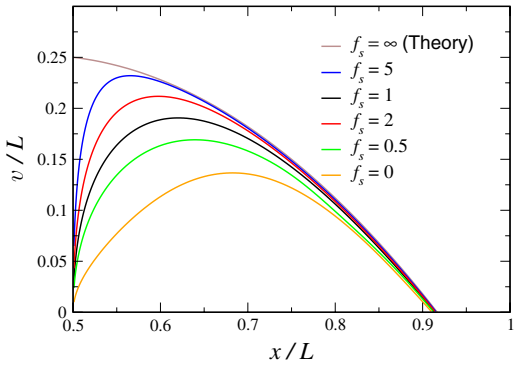

FIG. 3 (color online). Velocity vs average fitness for simulations of the noncutoff MFE [Eq. (1)] for various f_s. Parameters are $\mu = 0.1$, $L = 1000$, initial fitness $x_0 = 500$. The curve for $f_s = \infty$ is taken from Eq. (3).

decreases as N increases), this comparison gives us confidence that the above equation can account semiquantitatively for the recombination process.

In Fig. 3, we show the results of solving the MFE, Eq. (1), numerically for a variety of f_s values. At nonzero f_s, the fitness rapidly approaches a universal trajectory which is f_s independent; only the rate of approach varies. Hence, the amount of recombination is of minor importance if N is large enough for this mean-field theory to apply. We can explain this by noting a binomial fitness distribution, $B(L, \bar{x}/L)$, is preserved under the action of recombination for all values of $\bar{x}$. Thus, for large f_s, the system is quickly driven to a binomial distribution characterized by $p = \bar{x}/L$, the mean fitness per locus. This mean continues to evolve, with a velocity given by

$$v = \mathrm{Var}(x) + \mu\left[\bar{x} - \frac{2}{L}(\bar{x}^2 + \mathrm{Var}\,(x))\right], \qquad (2)$$

which is easily derived by multiplying Eq. (1) by x and summing over x. Notice that recombination has no direct effect in changing the fitness on average. Given the variance of the binomial distribution, $\mathrm{Var}\,(x) = Lp(1-p)$, we find that v grows linearly with L and is thus large:

$$v = L[p(1-p) + \mu p(1-2p)] - 2\mu p(1-p). \qquad (3)$$

The final state is therefore reached in a time essentially independent of L, even for small μ, so this indeed is quite rapid evolution; this analytic curve is included in Fig. 3. Now, the fact that recombination attempts to enforce a binomial distribution but otherwise does not directly change the rate of evolutionary advance explains why it has little consequence in the $N \to \infty$ mean-field limit. Actually this argument is more general. Essentially, the pure mutation-selection problem will, up to small corrections if L is large, also give rise to a binomial distribution which therefore self-consistently solves the entire equation. To see this, we replace the birth rate factors in the mutational part of the MFE by the constant rate $\bar{x}$; this introduces an error of $O[(x-\bar{x})/L]$, which becomes $O(L^{-1/2})$ were we to have a binomial distribution. Then, we can directly check that the same binomial ansatz solves the $f_s = 0$ time-dependent MFE. Hence, the only role for recombination is to cause the system to dynamically *select* this particular solution of the mean-field theory; the value of f_s makes no difference once we are past the transient period.

We have now explained why the large N saturation value in Fig. 1(a) is roughly f_s independent. The remaining issue concerns the critical value of $f_s^*(N)$ at which the system reaches the plateau [see Fig. 1(b)]; the previous argument suggests that as $N \to \infty$, $f_s^* \to 0^+$. This value is of crucial importance, as it represents the amount of recombination needed for a *finite* population to achieve the maximal rate of evolution. Studying this requires inclusion of finite population effects in the evolution equation, for which we employ a heuristic cutoff approach which has been shown to be accurate in a variety of previous investigations [8,17–19]. In detail, we replace the $(x-\bar{x})P(x)$ term in the mean-field equation [Eq. (1)] with the alternate form

$$[x\theta(P_x - P_c) - \lambda]P_x(t) \qquad (4)$$

leaving the rest of the equation unchanged. Here λ is chosen to satisfy population conservation, $\lambda = \int dx x P_x \theta(P_x - P_c)$, and P_c is a cutoff of order $1/N$. Figure 4(a) compares the time evolution of the stochastic system with that predicted by this cutoff MFE, showing reasonable agreement. Finally, Fig. 4(b) shows the desired effect, namely, the fact that the transition point to rapid evolution is a decreasing function of $\ln N$.

Why does finite N matter in this manner? It is easy to check that the cutoff term has no consequential effect as long as the distribution remains binomial. The breakdown in the previous analysis occurs when N becomes so small that the variance (and hence the velocity) saturates at $\ln N$ instead of order L. This transition means that the mutation-selection balance is not consistent with the binomial. We estimate the critical N by comparing the velocity found above, Eq. (3), with that expected when finite N effects are dominant. To estimate the latter, note that the recombination term in Eq. (1) can be written as the sum of a drift piece and a diffusion piece

$$f_s L([VP]' + [DP]''),$$

where prime refers to the finite difference operator and $V_x = (x-\bar{x})/L$, $D_x = (x+\bar{x})/2L - (x\bar{x})/L^2$. The drift term is small, because $x - \bar{x}$ is of order the square root of the variance, and so is always much less than L; hence, the most important effect is that of increased diffusion. This in fact appears to be the secret behind the efficacy of recombination in this model, namely, that it acts to increase variation just like an increased mutation rate but without a mean drift term, aka the "mutational load." The diffusion coefficient is finite as long as we are not very near $\bar{x} = L$. Assuming recombination dominates, we can use the results of previous analyses of the mutation-selection problem with $f_s L$ substituted for the genomic mutation rate $\mu\bar{x}$.

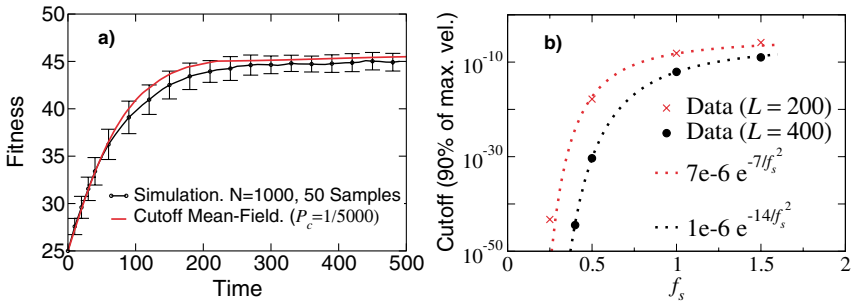

FIG. 4 (color online). (a) Fitness as a function of time averaged over 50 runs, $N = 1000$, compared to a numerical solution of the cutoff MFE with cutoff $P_c = 1/5000$. $\mu = 0.1$, $f_s = 2$, $L = 50$, $x_0 = 25$. (b) Cutoff for which v is 90% of its maximal (cutoff $= 0$) value as a function of f_s for $L = 200$, 400. $\mu = 0.001$. Velocity measured between $x = L/2 \pm 5$, with initial $x = 5$. Dotted line represent fits to Eq. (5).

From Ref. [8], the velocity under this assumption scales as

$$v \sim (f_s L)^{2/3} \ln^{1/3} N.$$

Equating this to our previous result that for large N the velocity scales as L independent of f_s, the predicted value of N (or equivalently the inverse cutoff) beyond which the system exhibits mean-field behavior scales as

$$N^* \sim \exp(CL/f_s^2) \tag{5}$$

for some order one constant C. This is consistent with the data shown in Fig. 4(b) and indeed with the limited direct simulation data in Fig. 1(b).

At this stage of our understanding, it is impossible to make any *quantitative* contact with experimental data. Nonetheless, conceptual insights that emerge from our study seem to offer solutions for some of the mysteries underlying bacterial transformation. Our results show that in the population range of interest for many microorganism colonies, there is a huge potential benefit to be gained from recombination; nevertheless too much recombination can hurt, as the specific genes are too rapidly driven to the most common allele even if it is not the beneficial one. This perhaps explains why recombination is so heavily regulated via intercellular signaling. The mechanism behind this benefit seems to be the increased rate of effective diffusion on the landscape, similar to what would happen with an increased mutation rate except that there is no significant extra load. Finally, we have already mentioned that our results are consistent with recent experiments; these could be extended to check the basic prediction of our approach regarding the scaling of the needed rate versus population size [Eq. (5)]. Even more exciting would be the determination that the signaling system is used for imposition of this result, measuring the effective population by quorum sensing and feeding the information into the competence pathway.

The work of H. L. has been supported in part by the NSF-sponsored Center for Theoretical Biological Physics (Grants No. PHY-0216576 and No. PHY-0225630) and that of D. A. K. and E. C. by the Israel Science Foundation.

[1] For a review, see S. P. Otto and T. Lenormand, Nature Reviews Genetics **3**, 252 (2002).
[2] W. P. C. Stemmer, Nature (London) **370**, 389 (1994).
[3] Hans-Georg Beyer, *The Theory of Evolution Strategies* (Springer-Verlag, Berlin, 2001); for a discussion of the role of recombination, see E. Baum, D. Boneh, and C. Garrett, Evolutionary Computation **9**, 93 (2001).
[4] H. J. Muller, Mutation Research **1**, 2 (1964); J. Felsenstein, Genetics **78**, 737 (1974).
[5] A. S. Kondrashov, J. Hered. **84**, 372 (1993).
[6] N. H. Barton, Genetical Research **65**, 123 (1995); S. P. Otto and N. H. Barton, Evolution (Lawrence, Kansas) **55**, 1921 (2001).
[7] B. Charlesworth, Genetical Research **63**, 213 (1994); W. R. Rice, Nature Reviews Genetics **3**, 241 (2002).
[8] L. S. Tsimring, H. Levine, and D. A. Kessler, Phys. Rev. Lett. **76**, 4440 (1996).
[9] D. A. Kessler, D. Ridgway, H. Levine, and L. Tsimring, J. Stat. Phys. **87**, 519 (1997).
[10] I. M. Rouzine, J. Wakeley, and J. M. Coffin, Proc. Natl. Acad. Sci. U.S.A. **100**, 587 (2003).
[11] I. S. Novella *et al.*, Proc. Nat. Acad. Sci. U.S.A. **92**, 5841 (1995); R. Miralles, A. Moya, and S. F. Elena, J. Virol. **74**, 3566 (2000); R. E. Lenski, M. R. Rose, S. C. Simpson, and S. C. Tadler, Am. Nat. **138**, 1315 (1991).
[12] D. Dubnau, Annu. Rev. Microbiol. **53**, 217 (1999).
[13] B. R. Levin and C. T. Bergstrom, Proc. Nat. Acad. Sci. U.S.A. **97**, 6981 (2000).
[14] For a somewhat more skeptical view, see R. Redfield, Nature Reviews Genetics **2**, 634 (2001).
[15] We intend in future work to explicitly model the distribution of genes in the intercellular medium, enabling us to account for lethal mutations among other effects.
[16] N. Colegrave, Nature (London) **420**, 664 (2002).
[17] E. Brener, H. Levine, and Y. Tu, Phys. Rev. Lett. **66**, 1978 (1991); E. Brunet and B. Derrida, Phys. Rev. E **56**, 2597 (1997); D. A. Kessler, Z. Ner, and L. M. Sander, Phys. Rev. E **58**, 107 (1998); L. Pechenik and H. Levine, Phys. Rev. E **59**, 3893 (1999).
[18] E. Cohen, D. Kessler, and H. Levine, cond-mat/0406336.
[19] D. A. Kessler and H. Levine, Nature (London) **394**, 556 (1998).