Rare Events in Population Genetics: Stochastic Tunneling in a Two-Locus Model with Recombination

Alexander Altland, Andrej Fischer, Joachim Krug, and Ivan G. Szendro
Institut für Theoretische Physik, Universität zu Köln, D-50973 Köln, Germany
(Received 8 December 2010; published 22 February 2011)

We study the evolution of a population in a two-locus genotype space, in which the negative effects of two single mutations are overcompensated in a high-fitness double mutant. We discuss how the interplay of finite population size N and sexual recombination at rate r affects the escape times t_{esc} to the double mutant. For small populations demographic noise generates massive fluctuations in t_{esc}. The mean escape time varies nonmonotonically with r, and grows exponentially as $\ln t_{\mathrm{esc}} \sim N(r - r^*)^{3/2}$ beyond a critical value r^*.

DOI: 10.1103/PhysRevLett.106.088101 PACS numbers: 87.23.Kg, 03.65.Sq, 05.45.−a, 87.10.Mn

Point mutations at different loci of the genome affect the fitness of living organisms via complex intragenomic correlations. These interactions, known as epistasis in population genetics, range from elementary pair interactions to extended patterns reflecting the structure of genetic networks [1]. An important form of epistasis termed reciprocal sign epistasis occurs when the deleterious effects of a single point mutation get (over)compensated by the beneficial effects of a secondary mutation [2,3]. The abundance of such elementary motifs is expected from the nature of complementary base pair binding in RNA [4,5], but it also stands to reason that they act as building blocks of evolutionary processes in more complex ''fitness landscapes'' [3,6].

The effects of epistatic pair correlation can be conveniently studied in a two-locus two-allele prototype system, i.e., a projection of the full genomic structure onto just two loci with alleles a, A and b, B, respectively. Here, the term ''locus'' refers to a specific genomic position, and the ''alleles'' $a, \ldots, B$ stand for the nucleotides present at the loci. Throughout this Letter we assume a fitness assignment such that the double mutant AB is superior to the wild-type ab and both have higher fitness than the single mutants Ab, aB, whose fitnesses are, moreover, taken to be equal. When evolving under the joint influence of mutation, selection, sexual recombination, and demographic noise, this system exhibits dynamical phenomena on a variety of time scales. Of particular interest here is the ''switching'' process occurring at very large time scales, when the effects of a deleterious single mutation at one of the two loci $ab \to Ab$ or $ab \to aB$ need to be overcome to reach a high-fitness doubly mutated configuration $(Ab/aB) \to AB$.

Several decades of research on the two-locus system notwithstanding [4,5,7–16], important aspects of the above compensatory mutation mechanism remain insufficiently understood. While the switching process, termed stochastic tunneling in the literature [9], is relatively well understood in asexuals [11,12], the behavior induced by sexual recombination is surprisingly complex. This is because of the dual role played by recombination in this system. On the one hand, the recombination of the unfavorable single mutants $Ab + aB \to AB$ provides a channel of AB generation that does not rely on mutation. On the other hand, this mechanism competes with back-recombination $AB + ab \to Ab, aB$ which breaks up AB genotypes once they have formed.

Detailed studies of the deterministic, infinite population dynamics show that the latter mechanism categorically wins in large populations. Beyond a critical value r^* of the recombination rate r a stable stationary solution localized at the wild-type genotype ab emerges [14], and the escape to the double mutant AB is completely suppressed [13]. However, computer simulations of finite populations also display a parameter regime for $r < r^*$ where the escape is aided by recombination, such that an initial decrease of the escape time t_{esc} eventually gives way to an increase at larger values of r [10]. The effect of recombination on the speed of adaptation is thus seen to depend on the population parameters in a complex way, as has also been observed in studies using empirical fitness data [17].

In addition to the recombination rate, the dynamics of the two-locus system is governed by the population size N, the mutation rate μ, and the strength of selection given by the typical scale of fitness differences s. Here we focus on populations that are moderately large, in the sense that $N\mu$, $Ns \gg 1$, and subject to strong selection with $s/\mu \gg 1$ [18]. Using a linearization of the full master equation near the initial wild type population, we derive an approximate expression for the full distribution of escape times for $r < r^*$. This allows us to identify two fundamentally different switching scenarios. In populations that are smaller than a characteristic size to be specified below, the temporal bottleneck of the evolution is the appearance of the first few individuals of the AB populations, which is a rare event with an exponentially distributed waiting time. The most likely escape time is then much smaller than the typical time, and fluctuations in

$t_{\rm esc}$ are of the same order as the mean. In contrast, in large populations the evolution is limited by the growth of the AB population and the distribution of $t_{\rm esc}$ is sharply peaked around the mean. The nonmonotonicity of $t_{\rm esc}$ observed in [10] is a feature of the fluctuation-dominated regime which disappears in large populations.

For $r > r^*$ the emergence of bistability in the deterministic dynamics implies that the problem becomes similar to the noise-driven escape from a metastable state. For this kind of problem path-integral methods akin to semiclassical quantum mechanics have recently been developed [19], and we show that they can be applied in the present context as well.

Model.—We consider the dynamics of the two-locus system, as governed by the interplay of selection, mutation and recombination. The two low fitness genotypes Ab, aB are lumped into a single subpopulation. This is justified when the creation rate of single mutants satisfies $N\mu \gg 1$, such that the number of single mutant individuals is large compared to unity at all times and the recombination process $Ab + aB \to (ab/AB)$ is not limited by the simultaneous presence of both parental types. The fitness of the three types is given by, respectively, $ab \leftrightarrow 1$, $(Ab/aB) \leftrightarrow 1 - s_d$, $AB \leftrightarrow 1 + s_b$, where $s_{d,b}$ denote the selection coefficients of the deleterious single mutants and the beneficial double mutant relative to the wild type ab, respectively. Mutation alters allelic content at a rate $\mu \ll s_d, s_b$. For instance, $ab \to^{2\mu} Ab/aB$, where the factor of 2 accounts for the fact that the change of either allele, a or b generates a single mutant. For simplicity, we will neglect the effect of back mutation $Ab/aB \to ab$ in our analysis of the escape time. In view of the smallness of the single mutant population (see below), this assumption is largely inconsequential. Finally, the random mating of individuals generates offspring by recombination at a rate r; for example, $AB + ab \to^{r} (Ab/aB)$.

The (Moran [20]) master equation for the evolution of the three population sizes, $n_0 \leftrightarrow ab$, $n_1 \leftrightarrow Ab + aB$, and $n_2 \leftrightarrow AB$ at constant total population $N = n_0 + n_1 + n_2$ reads as $\partial_t P(\mathbf{x}, t) = H(\mathbf{x})P(\mathbf{x}, t)$, where $\mathbf{x} = (x_0, x_1, x_2)$,

$$H(\mathbf{x}) = \sum_{i=0}^{2}[(E_i^+ - 1)d_i(\mathbf{x}) + (E_i^- - 1)b_i(\mathbf{x}) + (E_i^+ E_{i+1}^- - 1)m_i(\mathbf{x})] + \sum_{i,j=0}^{2}(E_i^+ E_j^- - 1)r_{ij}(\mathbf{x}), \quad (1)$$

and the notation emphasizes the analogies to an imaginary time "Schrödinger equation." In (1), $x_i = n_i/N$ are the genotype "frequencies," the "operators" $E_i^{\pm} f(x_i) \equiv e^{\pm(1/N)\partial_{x_i}} f(x_i) = f(x_i \pm N^{-1})$ act by translation by one individual, and the rates d_i, b_i, m_i, r_{ij} are the death, birth, mutation, and recombination rates [21]. Equation (1) describes the evolution of the system at one individual death or birth event per time step. Alternatively, one may employ a Wright-Fisher approach where an entire generation update is performed at each step [20]. While this is numerically faster by a factor of $\mathcal{O}(N)$, and has been used in parts of our computer simulations, the long rangedness of the Wright-Fisher equation in the variables (n_0, n_1, n_2) makes it difficult to handle analytically.

Escape time.—Figure 1 shows results for the time dependence of (n_0, n_1, n_2) obtained by simulation of the master equation (1). These profiles depend sensitively on the value of the parameter $N\gamma$, where the meaning of the rate $\gamma = 2\frac{\mu^2}{s_d} + r(\frac{\mu}{s_d})^2$ will be discussed momentarily. For $\gamma N \ll 1$ (upper panel), the limiting factor for the escape process is the stochastic generation of a sufficient number of AB individuals (cf. the inset). Once the AB clone reaches a critical size, the fitness advantage triggers a fast sweep through the entire population. In the opposite case $\gamma N \gg 1$, the quasideterministic increase of the x_2 population determines the escape time.

We aim to describe the dynamical processes characterizing the respective temporal bottlenecks from the master equation. To this end, we first notice that starting from an initially unmutated "wild population" $(x_0, x_1, x_2) = (1, 0, 0)$, a competition of mutation out of the ab population and a counteracting selection pressure s_d generates a population $x_1 \equiv \bar{x}_1 = 2\mu/s_d$ at short time scales $\sim s_d^{-1}$. In the parameter regime of interest here there are many individuals in a singly mutated state at any instance of time, $\bar{x}_1 \gg N^{-1}$. During the evolutionary stages determining the switching time, the frequencies $x_0 \simeq 1$ and $x_1 \simeq \bar{x}_1$ do not change significantly and the assumption of independence of P on (x_0, x_1) does not lead to qualitative errors. We thus substitute $(x_0, x_1) \simeq (1, \bar{x}_1)$ into (1) and set $x_2 \equiv x \equiv n/N$ for notational simplicity to obtain the linear master equation $\partial_t P(n, t) = H(n)P(n, t)$, where

$$H(n) = (E^- - 1)(R_+ n + \gamma) + (E^+ - 1)R_- n \quad (2)$$

and $R_+ = \frac{1}{N}(1 + s_b + r/2)$, $R_- = \frac{1}{N}(1 + 3r/2)$. The coefficient γ has been defined above, and can now be

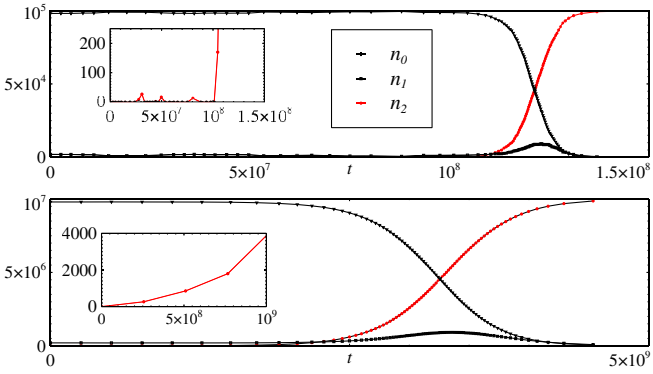

FIG. 1 (color online). Single runs of the escape process. Top ($\gamma N = 0.2$): creation of the first AB mutant determines escape time, bottom ($\gamma N = 20$): sweep through bulk of the population determines escape time. In the insets the population size axes have been enlarged to highlight the initial stages of the process.

understood as the effective rate of generation of double mutants, including contributions from both the mutational ($\sim \mu\bar{x}_1$) and recombinational channel ($\sim r\bar{x}_1^2$). The usage of the discrete variable $n \in \mathbb{N}$ in (2) accounts for the importance of the "quantization" of individuals to our further analysis. Starting from zero AB individuals, $P(n, t = 0) = \delta_{n,0}$, the time scale for the generation of an AB clone is set by γ^{-1}.

To obtain $P(n, t)$ we adapt the general solution [22] for the generating function $G(z, t) = \sum_{n=0}^{\infty} z^n P(n, t)$ of linear master equations to Eq. (2). This yields $G(z, t) = (1 + \kappa(t)(1 - z))^{-\lambda}$, where $\lambda = \gamma/R_+ \sim N\gamma$, $\kappa(t) = \frac{R_+}{\Delta R}(e^{\Delta R t} - 1)$, and $\Delta R = R_+ - R_- > 0$. Computing the inverse transform, we obtain

$$P(n, t) = \frac{\kappa^n}{(\kappa + 1)^{n+\lambda}} \prod_{j=0}^{n-1} \frac{j + \lambda}{j + 1}. \tag{3}$$

For fixed n, $P(n, t \to \infty)$ scales to zero showing that the distribution flows towards large values of n. Writing the master equation in the form $\partial_t P = J_{n-1} - J_n$, we identify the current $J_n(t) = (R_+ n + \gamma)P(n, t) - R_-(n + 1)P(n + 1, t) \simeq \Delta R n P(n, t)$ at which probability flows through a fixed reference value $n \gg 1$. The flow of the distribution implies $\int dt J_n(t) = 1$, which shows that $J_n(t)$ may be interpreted as the distribution $f_n(t)$ of escape times through n. We here consider $n \simeq N/2$ as a reference value where the AB population is about to take over, while the linearized approximation of the master equation has not yet become fully invalid. (For most parameter values, the ensuing threshold times are comparable to the times of full fixation, $n = N$.) For large n the product in (3) can be expressed in terms of the Γ function, which yields the result

$$f_n(t) \approx \Delta R \frac{\kappa(t)^n}{(\kappa(t) + 1)^{n+\lambda}} n^{\lambda} \Gamma^{-1}(\lambda). \tag{4}$$

Two time scales can be extracted from (4). For $t \to \infty$ the distribution decays exponentially on the time scale $t_{\text{tail}} = \frac{R_+}{\gamma \Delta R}$. The second time scale is the time at which (4) is maximal, given by $t_{\max} = (\Delta R)^{-1} \ln(\Delta R n/\gamma + 1)$. This is also the time when the solution of the rate equation $\dot{n} = \Delta R n + \gamma$ corresponding to (2) reaches the value n. Apart from the logarithmic factor we see that $t_{\max}/t_{\text{tail}} = \lambda$, showing that the parameter λ distinguishes the two dynamic regimes described above. For $\lambda \ll 1$ the distribution (4) becomes purely exponential, and fluctuations in t_{esc} are of the same order as the mean t_{tail}. In contrast, for $\lambda \gg 1$ the distribution is sharply peaked around the most likely value $t_{\max}$, with fluctuations of order Δt with $\Delta t/t_{\max} \sim N^{-1/2}$.

Inserting the explicit expressions for R_+, ΔR, and γ it is straightforward to show that t_{tail} displays a minimum as a function of r provided $s_d < s_b/2$, i.e., in cases where the fitness valley separating the genotypes ab and AB is relatively shallow. The minimum is located at $r_{\min} = s_b/2 - s_d < s_b/2$. The maximal speedup due to recombination can be substantial, and is given by $t_{\text{tail}}(r_{\min})/t_{\text{tail}}(r = 0) \approx \frac{8 s_d}{s_b}$ for $\mu \ll s_d \ll s_b$ [23]. By contrast, $t_{\max}$ increases monotonically with r.

Metastability.—Both time scales diverge when $\Delta R \sim (s_b - r)$ vanishes as $r \to s_b$. Within the simplified model defined by (2), this reflects the emergence of a stable stationary distribution centered around the wild-type genotype ab, which has been found in previous studies of the deterministic model [8,13,14]. Figure 2 shows the mean values $t_{\text{esc}}(r)$ (triangles) along with the escape times obtained from the simulation of the Moran master equation (1) (bars), and from the solution of the corresponding deterministic rate equations (dots). All three curves show a rapid increase of the escape time at the critical value $r^* \simeq s_b$. The rationale behind this behavior is that at strong recombination the "reshuffling" $AB + ab \to^r Ab, bA$ outperforms the fitness advantage of the AB population, thus preventing its growth.

We sketch the analysis of the model near criticality, $r \simeq r^*$. To this end we consider a Fokker-Planck approximation to (1), i.e., a second order expansion in the "momentum" variables $p_i = \partial_{x_i}$. This results in an operator of the generic form $H = -\frac{1}{N} p_i F_i + \frac{1}{2N^2} p_i X_{ij} p_j$, where the F_i denote the components of the deterministic drift term and X_{ij} is the diffusion matrix. Eliminating one frequency by normalization, $x_0 = 1 - x_1 - x_2$, the problem becomes two dimensional with $i = 1, 2$. The quickly equilibrating frequency $\bar{x}_1 \sim \frac{2\mu}{s_d}$ of the single mutant population follows from stationarity under drift, $F_1(x_0, \bar{x}_1) = 0$, which gets us to the effective one-dimensional Hamiltonian, $H = -\frac{1}{N} pF + \frac{1}{2N^2} X p^2$, where we have set $x_2 = x$ and $p_2 = p$ for notational simplicity. Then

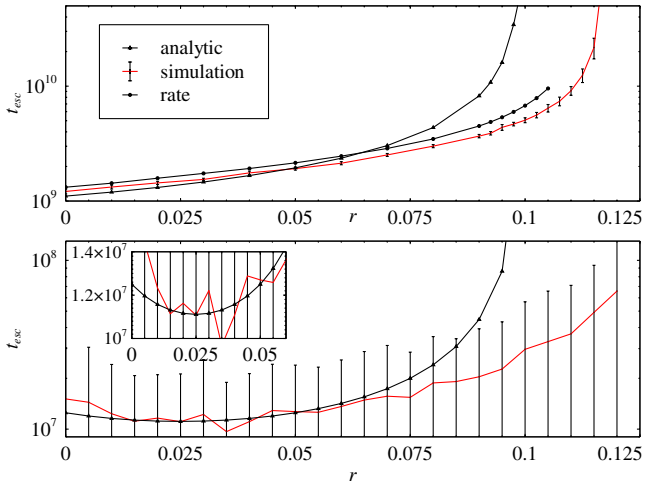

FIG. 2 (color online). Mean escape times as a function of r for $s_b = 0.1$, $s_d = 0.025$, $\mu = 10^{-4}$. Upper panel: $N = 10^7$, lower panel, $N = 10^4$. Error bars in lower panel illustrate the large fluctuations in the escape time, and inset highlights the minimum in t_{esc} at $r = r_{\min}$.

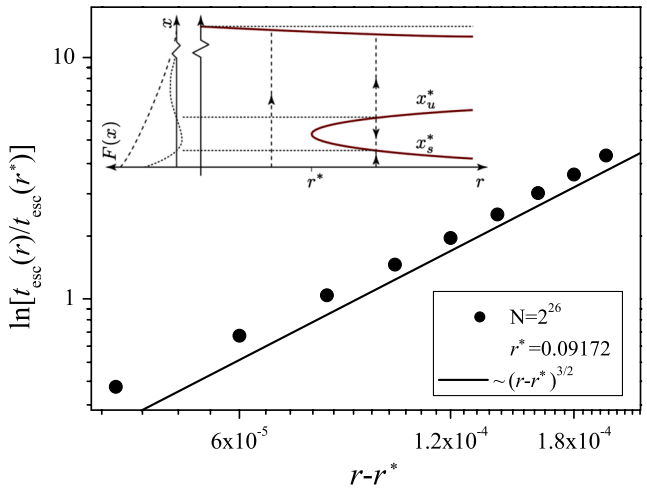

FIG. 3 (color online). Main figure shows the logarithm of the escape time, normalized to its value at $r = r^*$, as a function of $r - r^*$ in double-logarithmic scales. The escape time was defined as the time when the AB clone reaches 80% of the total population. Black dots correspond to simulations of a population of size $N = 2^{26} \approx 6.7 \times 10^7$ with parameters $s_b = 0.1$, $s_d = 0.025$, and $\mu = 10^{-5}$. All data are averaged over 10^5 runs. The straight line has slope 3/2. The inset illustrates the deterministic one-dimensional dynamics of the AB population. The vertical graph on the left shows the shape of the function $F(x)$ for $r < r^*$ (dashed line) and $r > r^*$ (dotted line). The horizontal graph shows the evolution of the fixed point structure with r. For $r < r^*$ the only fixed point is near $x_2 = x \approx 1$; for $r > r^*$ another pair of fixed points (x_s^*, x_u^*) emerges near $x \approx 0$.

$F(x) \equiv F_2(x, \bar{x}_1)$, $X(x) \equiv X_{22}(x, \bar{x}_1) \approx x(1 - x)$ describe the effective dynamics of the AB population.

The singularities in the escape times reflect the emergence of a pair of stable (x_s^*) and unstable (x_u^*) fixed points, $F(x_{s,u}^*) = 0$, $0 < x_s^* < x_u^* \ll 1$, close to the boundary $x = 0$ (inset of Fig. 3). The unstable fixed point x_u^* defines a "recombination barrier" [4] which blocks the fixation of the fitter population out of a small number of initial individuals $x(0) < x_u^*$. Based on general results for the escape from metastable states obtained within a large deviations ("WKB") approach [19], one expects the escape time to grow exponentially with N as $t_{esc} \sim \exp[CN\delta^3]$, where $C > 0$ is a constant and $\delta \sim x_u^* - x_s^*$. The analysis of the deterministic Wright-Fisher version of the problem shows that $\delta \sim (r - r^*)^{1/2}$ [14], and we predict that $t_{esc} \sim \exp[CN(r - r^*)^{3/2}]$. Figure 3 shows data for the escape time obtained by simulating a Wright-Fisher process [17] at values of r slightly above r^*. The value $r^* \simeq 0.09172$ was calculated using expressions derived in [14] for the deterministic $(N \to \infty)$ limit. The results are seen to be in good agreement with the prediction $\ln t_{esc} \sim (r - r^*)^{3/2}$.

Conclusion.—To summarize, we have presented an analysis of a paradigmatic two-locus model of population genetics. We have seen how, depending on the population size, recombination may speed up or delay the evolution towards the high fitness state. A challenge for future work is to develop the WKB-type approximation for $r > r^*$ into a fully quantitative theory, which can be used to predict the size of the recombination barrier in a specific biological setting.

This work was supported by DFG within SFB TR 12 and SFB 680. We thank U. Gerland, B. Meerson, D. B. Weissman, and R. Neher for useful discussions and remarks.

[1] P. C. Phillips, Nat. Rev. Genet. **9**, 855 (2008).
[2] F. J. Poelwijk, D. J. Kiviet, D. M. Weinreich, and S. J. Tans, Nature (London) **445**, 383 (2007).
[3] F. J. Poelwijk, S. Tănase-Nicola, D. J. Kiviet, and S. J. Tans, J. Theor. Biol. **272**, 141 (2011).
[4] W. Stephan, Genetics **144**, 419 (1996).
[5] P. G. Higgs, Genetica **102/103**, 91 (1998).
[6] D. M. Weinreich, R. A.Watson, and L. Chao, Evolution **59**, 1165 (2005).
[7] J. F. Crow and M. Kimura, Am. Nat. **99**, 439 (1965).
[8] I. Eshel and M. W. Feldman, Theor. Popul. Biol. **1**, 88 (1970).
[9] Y. Iwasa, F. Michor, and M. A. Nowak, Genetics **166**, 1571 (2004).
[10] D. M. Weinreich and L. Chao, Evolution **59**, 1175 (2005).
[11] D. B. Weissman, M. M. Desai, D. S.Fisher, and M. W. Feldman, Theor. Popul. Biol. **75**, 286 (2009).
[12] C. S. Gokhale, Y. Iwasa, M. A.Nowak, and A. Traulsen, J. Theor. Biol. **259**, 613 (2009).
[13] K. Jain, Theor. Popul. Biol. **77**, 23 (2010).
[14] S.-C. Park and J. Krug, J. Math. Biol. (in press).
[15] M. Lynch, Proc. Natl. Acad. Sci. U.S.A. **107**, 16 577 (2010).
[16] D. B. Weissman, M. W.Feldman, and D. S. Fisher, Genetics **186**, 1389 (2010).
[17] J. A. G. M. de Visser, S.-C. Park, and J. Krug, Am. Nat. **174**, S15 (2009).
[18] Results for other parameter regimes are found in [15,16].
[19] M. Assaf and B. Meerson, Phys. Rev. E **81**, 021116 (2010).
[20] R. Durrett, *Probability Models for DNA Sequence Evolution* (Springer, New York, 2002).
[21] The rates in (1) are defined as follows: Death rates $d_i = x_i$; birth rates $b_i = (w_i/\bar{w})x_i$, where w_i are the genotype fitnesses given in the text and $\bar{w} = \sum_{i=0}^{2} w_i x_i$ is the population mean fitness; mutation rates $m_0 = 2\mu x_0$, $m_1 = \mu x_1$; and recombination rates $r_{01} = rx_0(x_1 + x_2)$, $r_{02} = r_{20} = (r/2)x_0x_2$, $r_{10} = rx_0x_1 + (r/4)x_1^2$, $r_{12} = rx_1x_2 + (r/4)x_1^2$, $r_{21} = rx_2(x_0 + x_1)$.
[22] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (North-Holland, Amsterdam, 2007).
[23] See [15] for an analysis of the neutral limit $s_d \to 0$.