Gene-influx-driven evolution

David B. Saakian[1,*] and Eugene V. Koonin[2]

[1]*A.I. Alikhanyan National Science Laboratory (Yerevan Physics Institute) Foundation, 2 Alikhanian Brothers St., Yerevan 375036, Armenia*

[2]*National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA*

(Received 13 April 2022; accepted 31 May 2022; published 12 July 2022)

Here we analyze the evolutionary process in the presence of continuous influx of genotypes with submaximum fitness from the outside to the given habitat with finite resources. We show that strong influx from the outside allows the low-fitness genotype to win the competition with the higher fitness genotype, and in a finite population, drive the latter to extinction. We analyze a mathematical model of this phenomenon and obtain the conditions for the transition from the high-fitness to the low-fitness genotype caused by the influx of the latter. We calculate the time to extinction of the high-fitness genotype in a finite population with two alleles and find the exact analytical dynamics of extinction for the case of many genes with epistasis. We solve a related quasispecies model for a single peak (random) fitness landscape as well as for a symmetric fitness landscape. In the symmetric landscape, a nonperturbative effect is observed such that even an extremely low influx of the low-fitness genotype drastically changes the steady state fitness distribution. A similar nonperturbative phenomenon is observed for the allele fixation time as well. The identified regime of influx-driven evolution appears to be relevant for a broad class of biological systems and could be central to the evolution of prokaryotes and viruses.

DOI: 10.1103/PhysRevE.106.014403

I. INTRODUCTION

Darwinian evolution is based on competition between individuals of the same species and/or between species for common, limiting resources. Under population genetic theory, the outcome of this competition depends on fitness distribution and the population size. In popular evolutionary models, the evolving population either maintains a constant size [1–3], or population growth saturates with the carrying capacity. In the absence of mutation, only the fittest genotype survives in the evolving population. With a nonzero mutation rate, if there are two fitness peaks $\exp(r_1)$, $\exp(r_2)$, and a passable (i.e., containing no strongly deleterious mutations) mutational path between the peaks, with the total transition probability U, then the population of the peaks is comparable if $r_1 - r_2 \sim U$, and otherwise, the type with low fitness goes extinct.

Here, we investigate the case when there is an influx of the low fitness type, which is a common situation in biology, particularly in the microbial world. A pertinent case, for example, is a pandemic, such as COVID-19, during which an influx of viruses from one location to another is common [4,5]

More generally, the phenomena modeled here relates to gene drive, the natural or engineered process where a certain suite of alleles is propagated across a population replacing other alleles [6,7]. Gene drive is a powerful process to manipulate fates of populations, both in sexually reproducing organisms, such as insects [8], and in viruses [9].

Recently, related phenomena have been thoroughly investigated in the context of social evolutionary dynamics [10,11]. Here we explore a distinct form of gene drive where a less fit genotype supplants the fittest one due to continuous influx of the former. To model gene-influx-driven evolution, we first consider the Moran model for a finite population [1], and then, the Crow-Kimura model for an infinite population [12–23], with migration between two habitats [24], an external habitat with a single, fixed genotype, and the main habitat with many genotypes that are subject to selection and mutation. We construct and analyze a model for the many genes case with epistasis. The critical influx of the low-fitness genotype that is required for the elimination of the high-fitness genotype is calculated analytically and validated by numerical simulations. We further solve the Crow-Kimura model with invasion for the case of large genomes, versus two genotype models with invasion considered in [25].

II. TWO-ALLELE MODEL WITH AN INFLUX OF THE LOW-FITNESS MUTANT IN A FINITE POPULATION

Consider a finite population with i mutants at the time t and $(N - i)$ wild-type replicators, so the total population size is N.

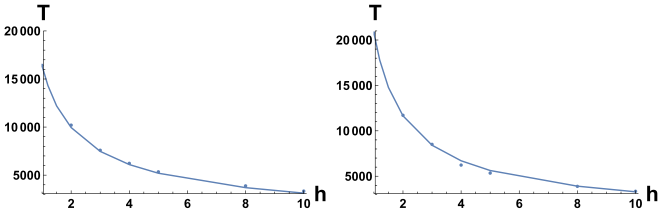
Consider the following random processes.

(a) With the probability $\frac{h+e^{-s}i/N}{h+e^{-s}i/N+1-i/N}$ there is a growth of the mutant number by 1. Here the parameter h defines the influx of the mutants.

(b) With a probability of $\frac{1-i/N}{h+e^{-s}i/N+1-i/N}$ there is a birth of wild type.

(c) In case there is a birth of a new replicator, there is a uniform dilution of the system to hold a constant population size: the mutant number decreases by 1 with a probability i/N and the wild-type number decreases by 1 with a probability $(N - i)/N$.

[*] saakian@yerphi.am



(a) (b)

FIG. 1. (a) The mean fixation time T versus h, the neutral case $N = 100, s = 0, i = 1$. The smooth line is our analytical result by Eq. (5), and the solid dots are the numerical results. Here s is the selection coefficient, while h describes the rate of the influx. (b) The mean fixation time T versus h, the case $N = 100, s = 1, i = 1$. The smooth line is our analytical result by Eq. (5), and the solid dots are the numerical results.

Our Markov model is described via the following transition probabilities to the ith state:

$$P_{i,i+1} = \frac{(h + e^{-s/N} i/N)(1 - i/N)}{h + e^{-s/N} i/N + 1 - i/N} = f(x),$$
$$P_{i,i-1} = \frac{(1 - i/N)i/N}{h + e^{-s/N} i/N + 1 - i/N} = b(x),$$
$$P_{ii} = 1 - f(x) - b(x), \tag{1}$$

where we denoted $x = i/N$. The latter expression coincides with Eq. 3.65 from [1], if we put $h = 0$.

Then we can calculate the average time to the irradiation of the wild type. As all the time there is influx of the mutants, they will be fixed with probability 1. Let us make an ansatz $T(x) = N^2, t_i, x = i/N$.

Then we apply the diffusion approximation [26–28]

$$\begin{aligned} &N^2 T(x + 1/N) f(x) + N^2 T(x - 1/N) b(x) \\ &\quad - N^2[f(x) + g(x)]T(x) \\ &= 1. \end{aligned} \tag{2}$$

After an expansion via $1/N$ degrees, we get

$$T'' x(1 - x) + (h - xs)(1 - x)T' = 1. \tag{3}$$

We set the conditions

$$T(1) = 0, \quad T(0) < \infty. \tag{4}$$

Eventually we derive the following solution:

$$T(x) = \int_x^1 dx \frac{\exp[sx]}{x^h} \int_0^x \frac{\exp[-sy] y^{h-1}}{(1 - y)} dy. \tag{5}$$

We performed numerics for $0.01 < h < 10$, and then numerics confirmed the $O(1/N)$ accuracy of our analytical result (see Fig. 1).

III. THE DRIVEN EVOLUTION IN THE CASE OF CROW-KIMURA MODEL WITH SINGLE PEAK FITNESS LANDSCAPE

We consider a model with N genes; two types for any one. For the N genes we have 2^N types. Following Refs. [5,6], for 2^N frequencies P_i we write the following set of equations:

$$\frac{dP_i}{d\tau} = \left(r_j - \frac{P}{\kappa} - \mu\right)P_j + h\delta_{i,0} + \frac{\mu}{N} \sum_{d_{ij}=1} P_j,$$
$$P = \sum_{j=0}^{(2^N-1)} P_j, \tag{6}$$

where d_{ij} is the Hamming distance, the number of point mutations to convert the ith sequence to the jth sequence, and μ is the total mutation rate per genome, κ is the carrying capacity in the system, and h is the influx rate. P_i is a total number of the ith type, and P is the population size. We can rescale r_i, μ, h to get $\kappa = 1$. Further we will take $\kappa = 1$, then define P as a population size and P_i as a relative frequency.

A. Driven evolution with influx at the peak

We consider the model by Eq. (6) for the case when

$$r_0 = J, \quad r_i = 0, i \geqslant 1. \tag{7}$$

We define as an lth Hamming class as the sequences with the same number of mutations l from the reference sequence. There are $N_l = N!/(l!(N - l)!)P_i \approx N^l/$ such sequences. In the steady state all of them have equal relative frequencies P_i. We define $p_l = N_l P_i$.

Ignoring the back mutations with the accuracy $O(1/N)$ in Eq. (11), we derive

$$p_0(J - 1 - P) + h = 0,$$
$$p_n = \frac{p_{n-1}}{P + 1}, \quad n > 0. \tag{8}$$

We get immediately the expression for p_0:

$$p_0 = \frac{h}{(P + 1 - J)},$$
$$\frac{\sum_l p_l}{p_0} = \frac{P + 1}{P}. \tag{9}$$

Eventually, we derive a third-order equation for the total population size:

$$\frac{h}{(P + 1 - J)} \frac{P + 1}{P} = P. \tag{10}$$

We take the single real solution of Eq. (10). Figure 2(a) illustrates the accuracy of our analytical results.

B. Driven evolution in the case of an alternative peak

Consider now the case when there are two peak sequences: a sequence with the fitness J_1 related to the influx, and another sequence with a high peak J_2; all the other sequences have a zero fitness. In this situation, two solutions are possible. In the first case, the population of the second peak is exponentially smaller compared to the population around the first peak. In the second case, there are peaks of population distributions around both sequences, with a ratio of total populations ~1.

We already gave the solution of the first case in the statics, Eq. (10). For the first peak with the fitness peak height J_1, we take as P the real value solution of Eq. (9). The population around the second peak is negligible.

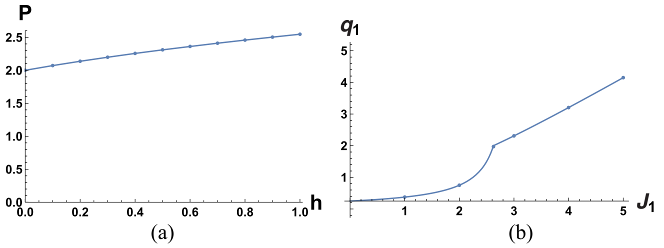


FIG. 2. (a) The population size of the model P versus the influx parameter h. The smooth line is our analytical result by Eq. (10), and the solid dots are the numerical results for $J = 3, N = 1000$. In the case of nonzero influx it is not reasonable to consider a finite population size; the latter is derived from the evolutionary dynamics with a saturation. We used the simple version of saturation from [24]. (b) The population size around the first peak q_1 versus J_1. The smooth line is our analytical result by Eq. (13), and the solid dots are the numerical results for $h = 0.5, J_2 = 3, N = 1000$. Now the state of the model is defined by the f.

Consider now the second solution, focusing on the sequences near the second peak. We have

$$PP_j = (r_j - \mu)P_j + \frac{\mu}{N}\sum_{d_{ij}=1} P_j. \tag{11}$$

We obtain that, for nonzero P_i, P should be the maximal eigenvalue of the model correspondingto the Crow-Kimura model with the fitness landscape having peak value J_2 and zero fitness for all other sequences, so [18]

$$P = J_2 - 1. \tag{12}$$

Equation (9) gives for the population size q_1 near the first peak

$$\begin{aligned} q_1 &= \frac{h}{(P+1-J_1)}\frac{P+1}{P} \\ &= \frac{h}{(J_2-J_1)}\frac{J_2}{J_2-1}. \end{aligned} \tag{13}$$

Then, for the population size q_2 near the second peak, we have

$$q_2 = J_2 - 1 - q_1. \tag{14}$$

Figure 2(b) illustrates the accuracy of our analytical result for the population around the first peak. Figure 3(a) illustrates the accuracy of our analytical result for the population around the second peak.

C. The dynamics

We are interested in how the population around the high peak disappears due to the influx of the low-fitness genotypes. We derived the following equations for the relative frequencies of the first peak sequence P_0 and the second peak sequence $\hat{P}_0$ in Eq. (6):

$$\begin{aligned} \frac{dP_0}{dt} &= P_0(J_1 - 1 - P) + h, \\ \frac{d\bar{P}_0}{dt} &= \bar{P}_0(J_2 - 1 - P). \end{aligned} \tag{15}$$

For the size of population around the first peak (without the peak population) q_3 and around the second peak (without the peak population) q_4 we obtain, summing Eq. (6),

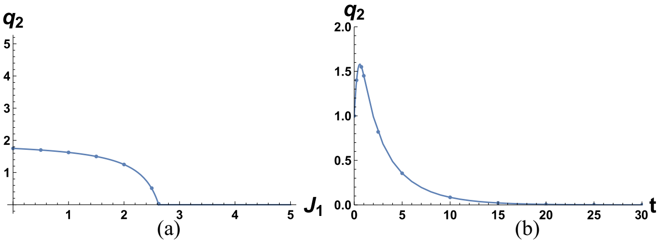


FIG. 3. The figure illustrates how the population around the high peak disappears due to the influx of the low fitness genotypes. (a) The population size around the second peak q_2 versus J_1. The smooth line is our analytical result by Eq. (14), and the solid dots are the numerical results for $J_2 = 3, N = 1000$. After the critical value of $J_1 > J_c$ the population of the second peak disappears. (b) The population size around the second peak P_2 versus t. The smooth line is our analytical result by Eqs. (16) and (17), and the solid dots are the numerical results for $J_1 = 2.625, J_2 = 3, h = 1, N = 1000$.

$$\begin{aligned} \frac{dq_3}{dt} &= -q_3P + P_0, \\ \frac{dq_4}{dt} &= -q_4P + \hat{P}_0 \end{aligned} \tag{16}$$

Then, we have by the definition

$$P = p_0 + \bar{P}_0 + q_3 + q_4. \tag{17}$$

Thus we have a system of four nonlinear equations, Eqs. (15)–(17).

In Fig. 3(b), we compare the analytical dynamics (smooth line) with the direct numerics (solid dots).

The single peak fitness landscapes are not generally relevant in biology, whereas random fitness landscapes are far more realistic [29]. However, the key features of evolution are identical for both these cases, as shown in [15,23] using the relation of the random fitness evolution model with the random energy model by Derrida [30]. The steady state properties of the model with a single peak landscape are equivalent to the properties of the model with a random fitness distribution when the gap between the maximum fitness and the vast majority of sequence fitness values is the same as in the Crow-Kimura model.

IV. SYMMETRIC FITNESS LANDSCAPE

Consider Eq. (6) for the case when the fitness depends on the total number of mutations from the reference zeroth sequence.

$$r_i = f(x), \quad x = 1 - 2d_{i,0}/l, \tag{18}$$

where $f(x)$ has a maximum at $x = 1$, so the zeroth sequence has a maximal fitness.

A. The statics in the case of the influx at the peak

We assume that the population concentrates around the peak (zeroth) sequence. Then, summing Eq. (6), we obtain

$$P^2 = Pf(1) + h. \tag{19}$$

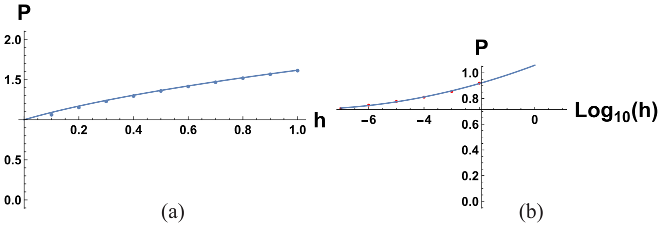


FIG. 4. The symmetric fitness landscape around one peak. (a) The population size of the symmetric fitness landscape case versus h, $f(x) = x^2, N = 1000$. The smooth line is our analytical result by Eq. (22), and the solid dots are the numerical results. After the critical value of $J_1 > J_c$ the population of the second peak disappears. (b) The population size of the symmetric fitness landscape case versus $\log_{10}(h)$, $f(x) = x^2, N = 1000$. The solid dots are the numerical results, and the smooth line is the quadratic fit. There is a nonperturbative phenomenon via the influx rate h.

While deriving Eq. (19), we assumed that the distribution around the peak is narrow, which can be proved rigorously, so we replaced r_i by r_{i_0} in the expression of $\langle r_i P_i \rangle$.

Taking the large value solution of the latter equation, we find that even a very small influx can drastically change the mean fitness. The numeric results strongly support this conclusion [Fig. 4(a)]. Instead of the solution

$$P = \max[f(x) + \sqrt{1 - x^2} - 1]|_x,$$

we obtain

$$P = f(1). \tag{20}$$

Thus, $h = 0$ is the singularity point. Figure 4(b) illustrates the behavior of the model at the limit of very small h.

B. The influx at the low peak and extinction of the population at high peak

Consider the case of two peaks with the sequences 1 and 2. There are finite population sizes around both the peaks. Near the first peak, we have the fitness function $f(x)$,

$$r_i = f(1 - 2d_{1i}/N),$$

and near the second peak, the fitness function $g(x)$:

$$r_i = g(1 - 2d_{2i}/N).$$

The total population size is defined by the steady state mean fitness of the corresponding Crow-Kimura model, which has the corresponding surplus s_2,

$$\begin{aligned} P &= \max[g(x) + \sqrt{1 - x^2} - 1], \\ s_2 &= \langle 1 - 2d_{2i}/N \rangle. \end{aligned} \tag{21}$$

Summing Eq. (6), we obtain

$$\begin{aligned} R &= \max[g(x) + \sqrt{1 - x^2} - 1]|_x, \\ q_1 + q_2 &\equiv P = R, \\ P^2 &= q_1 f(1) + h + q_2 P, \end{aligned} \tag{22}$$

where q_1, q_2 are the population *sizes* around the first and the second peak, respectively. While deriving the latter equations,

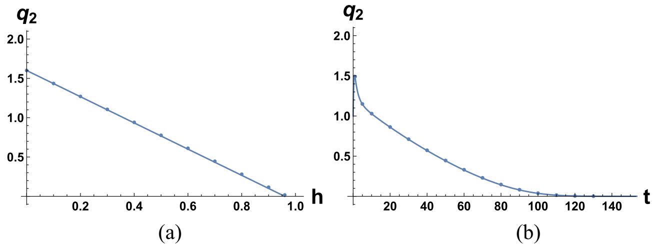


FIG. 5. There are two peaks, with symmetric fitness landscapes in their vicinity. The extinction of the high peak population in the case of the symmetric fitness landscape. (a) The population size around the second peak of the symmetric fitness landscape versus h, $f(x) = x^2$, $g(x) = 2.5x^2, N = 1000$. The smooth line is the analytical result by Eq. (22), and the solid dots are the numerical results. Beyond the critical value of $h > h_c$, the population of the second peak goes extinct. (b) The dynamics of q_2 for the case when originally the population is focused at the high peak. The smooth line is our analytical result by Eq. (23), and the solid dots are the result of numerics.

we again assume narrow distributions around the peaks (the width of distribution is $\sim\sqrt{N}$).

We calculated q_1, q_2 from this equation and compared with the direct numerics see Fig. 5(a).

C. The dynamics

Consider now the dynamics in the case of a symmetric fitness landscape. Summing via the index i in Eq. (6), we obtain

$$\begin{aligned} dq_1/dt &= q_1[f(1) - (q_1 + q_2)] + h, \\ dq_2/dt &= q_2[g(s(t)] - (q_1 + q_2)), \end{aligned} \tag{23}$$

where $s(t) = 1 - 2d/N$, and d is the mean Hamming distance for the sequences around the second peak. $s(t)$ is the surplus around the second peak. It is identical to the corresponding surplus of the standard Crow-Kimura model with the fitness function $g(x)$. The latter has been already calculated in [22] and the following expression has been derived for its inverse function $T(s)$:

$$T(s) = -\frac{1}{2}\int_s^1 \frac{d\xi}{\sqrt{[f(s) + 1 - f(\xi)]^2 - 1 + \xi^2}}. \tag{24}$$

Thus, we derived an analytical expression for the extinction dynamics of the high peak sequences. Our analytical results coincide with those of numerical simulations, with the accuracy $O(1/N)$ [see Fig. 5(b)].

V. CONCLUSIONS

In this work, we analyze the phenomenon of driven evolution, when there is a permanent influx of a single genotype into an evolving population. We explore this phenomenon for a finite population case, calculating the time to fixation, as well as in the infinite population case of the many genes model with epistasis. We investigated several models. Our main goal was to calculate both the threshold values of the influx, which are required to eliminate the high peak sequences from the population, and the dynamics of the process. We analyzed the case of a symmetric fitness landscape as well as a single peak landscape, the latter being mathematically equivalent to the realistic case of a random fitness landscape [15,23]. In the smooth fitness case, the driven evolution phenomenon shows nonperturbative behavior such that even a very small influx of the low-fitness genotype changes the mean fitness. Our analytical results have a relative accuracy $O(1/N)$ (N is the number of genes in the genome). This nonperturbative phenomenon is expected to be manifest in complex adaptive systems as well, when the agents competing for finite resources could either change their strategies or have a rigid strategy. Such could be the situation in learning [31] or opinion dynamics [32].

We further explored a simple evolutionary dynamics model that includes interaction between different genotypes via competition for the resources. It can be expected that the mechanism demonstrated for this model also applies to more complex, strongly nonlinear evolution phenomena with adaptation. Close finite population models were previously analyzed in detail [25] for the birth-death model, close to the one considered here. Novozhilov and colleagues also examined the deterministic invasion model in the case of two genotypes. Our current finite population model is simpler, and we derive only the mean time to fixation. The main difference is in the deterministic model. We solve models with many genotypes, obtaining several phases, including nonperturbative phenomena, which is not addressed in the work of Novozhilov and colleagues [25].

The results of this work will inform management of gene drive interventions by defining the specific requirements for the gene influx to eliminate the dominant genotype, which is crucial for the control of ecological consequences [33]. Among possible applications are control of viral infections and cancer therapy [34,35], where invasion of genetically distinct viruses or tumor cells into habitats occupied by a dominant variant is a key phenomenon. From the biological perspective, our main finding is that, if there is mutation-selection balance in a population, even an extremely low but continued influx of a new variant, not necessarily a high-fitness one, can drastically change that balance. The relevance of this finding for biological phenomena of major impact, such as epidemics, including the ongoing COVID-19 pandemic, is obvious.

ACKNOWLEDGMENT

This work was supported by SCS of Armenia, Grant No. 21T-1C037.

[1] W. J. Ewens, *Mathematical Population Genetics* (Springer-Verlag, New York, 2004).
[2] E. Baake and W. Gabriel, Annu. Rev. Comput. Phys. **7**, 203 (2000).
[3] B. Drossel, Adv. Phys. **50**, 209 (2001).
[4] V. P. Chavda, A. B. Patel, and D. D. Vaghasiya, J. Med. Virol. **94**, 2986 (2022).
[5] Y. Chen, S. Li, W. Wu, S. Geng, and M. Mao, J. Med. Virol. **94**, 2035 (2022).
[6] L. S. Alphey, A. Crisanti, F. Randazzo, and O. S. Akbari, Proc. Natl. Acad. Sci. USA **117**, 30864 (2020).
[7] J. Champer, A. Buchman, and O. S. Akbari, Nat. Rev. Genet. **17**, 146 (2016).
[8] P. T. Leftwich, M. P. Edgington, T. Harvey-Samuel, L. Z. Carabajal Paladino, V. C. Norman, and L. Alphey, Biochem. Soc. Trans. **46**, 1203 (2018).
[9] M. Walter and E. Verdin, Nat. Commun. **11**, 4884 (2020).
[10] M. Perc, J. J. Jordanc, D. G. Rand, Z. Wang, S. Boccaletti, and A. Szolnoki, Phys. Rep. **687**, 1 (2017).
[11] M. Jusup, P. Holme, K. Kanazawa, M. Takayasu, I. Romić, Z. Wang, S. Geček, T. Lipić, B. Podobnik, L. Wang, W. Luo, T. Klanjšček, J. Fan, S. Boccaletti, and M. Perc, Phys. Rep. **948**, 1 (2022).
[12] M. Eigen, J. Mc Caskill, and P. Schuster, Adv. Chem. Phys. **75**, 149 (1989).
[13] J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory* (Harper Row, New York, 1970).
[14] E. Baake, M. Baake, and H. Wagner, Phys. Rev. Lett. **78**, 559 (1997).
[15] S. Franz and L. Peliti, J. Phys. A: Math. Gen. **30**, 4481 (1997).
[16] J. Hermisson, O. Redner, H. Wagner, and E. Baake, Theor Popul. Biol. **62**, 9 (2002).
[17] E. Baake and H. Wagner, Genet. Res. **78**, 93 (2001).
[18] D. B. Saakian and C. K. Hu, Phys. Rev. E **69**, 046121 (2004).
[19] J. M. Park and M. W. Deem, J. Stat. Phys. **125**, 971 (2006).
[20] D. B. Saakian, J. Stat. Phys. **128**, 781 (2007).
[21] K. Sato and K. Kaneko, Phys. Rev. E **75**, 061909 (2007).
[22] D. B. Saakian, O. Rozanova, and A. Akmetzhanov, Phys. Rev. E **78**, 041908 (2008).
[23] D. B. Saakian and J. F. Fontanari, Phys. Rev. E **80**, 041903 (2009).
[24] B. J. Waclaw, R. J. Allen, and M. R. Evans, Phys. Rev. Lett. **105**, 268101 (2010).
[25] A. S. Novozhilov, G. P. Karev, and E. V. Koonin, Mol. Biol. Evol. **22**, 1721 (2005).
[26] S. Wright, in *Proceedings of the Sixth International Congress on Genetics* (Brooklyn Botanic Garden, New York, 1932), Vol. 1, p. 356.
[27] S. Wright, Proc. Natl. Acad. Sci. USA **31**, 382 (1945).
[28] R. A. Fisher, *The Genetical Theory of Natural Selection* (Clarendon, Oxford, 1930).
[29] R. Sanjuan, A. Moya, and S. F. Elena, Proc. Natl. Acad. Sci. USA **101**, 8396 (2004).
[30] B. Derrida, Phys. Rev. B **24**, 2613 (1981).
[31] V. G. Red'ko, Biol. Inspired Cognit. Archit. **22**, 95 (2017).
[32] E. Ben-Naim, L. Frachebourg, and P. L. Krapivsky, Phys. Rev. E **53**, 3078 (1996).
[33] K. M. Esvelt, A. L. Smidler, F. Catteruccia, and G. M. Church, eLife **3**, e03401 (2014).
[34] K. Danesh, R. Durrett, L. J. Havrilesky, and E. Myers, J. Theor. Biol. **314**, 10 (2012).
[35] B. Waclaw, I. Bozic, M. E. Pittman, R. H. Hruban, B. Vogelstein, and M. A. Nowak, Nature (London) **525**, 261 (2015).