Evolutionary advantage via common action of recombination and neutrality

David B. Saakian[1,2,3,*] and Chin-Kun Hu[1,†]

[1]*Institute of Physics, Academia Sinica, Nankang, Taipei 11529, Taiwan*

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

[3]*Physics Division of National Center for Theoretical Sciences Taipei Branch, National Taiwan University, Taipei, Taiwan*

(Received 3 December 2012; revised manuscript received 9 July 2013; published 25 November 2013)

We investigate evolution models with recombination and neutrality. We consider the Crow-Kimura (parallel) mutation-selection model with the neutral fitness landscape, in which there is a central peak with high fitness A, and some of 1-point mutants have the same high fitness A, while the fitness of other sequences is 0. We find that the effect of recombination and neutrality depends on the concrete version of both neutrality and recombination. We consider three versions of neutrality: (a) all the nearest neighbor sequences of the peak sequence have the same high fitness A; (b) all the l-point mutations in a piece of genome of length $l \geqslant 1$ are neutral; (c) the neutral sequences are randomly distributed among the nearest neighbors of the peak sequences. We also consider three versions of recombination: (I) the simple horizontal gene transfer (HGT) of one nucleotide; (II) the exchange of a piece of genome of length l, HGT-l; (III) two-point crossover recombination (2CR). For the case of (a), the 2CR gives a rather strong contribution to the mean fitness, much stronger than that of HGT for a large genome length L. For the random distribution of neutral sequences there is a critical degree of neutrality ν_c, and for $\mu < \mu_c$ and $(\mu_c - \mu)$ is not large, the 2CR suppresses the mean fitness while HGT increases it; for ν much larger than ν_c, the 2CR and HGT-l increase the mean fitness larger than that of the HGT. We also consider the recombination in the case of smooth fitness landscapes. The recombination gives some advantage in the evolutionary dynamics, where recombination distinguishes clearly the mean-field-like evolutionary factors from the fluctuation-like ones. By contrast, mutations affect the mean-field-like and fluctuation-like factors similarly. Consequently, recombination can accelerate the non-mean-field (fluctuation) type dynamics without considerably affecting the mean-field-like factors.

DOI: 10.1103/PhysRevE.88.052717 PACS number(s): 87.23.Kg, 87.19.xd

I. INTRODUCTION

Applications of statistical physics to molecular models of biological evolution [1–8] and the origin of life [2,9,10] have attracted much attention in recent years. A still unsolved and interesting problem in evolution theory is the origin of sex. It has been well recognized that recombination of genomes from sexual organisms gives some evolutionary advantages [11–14], which have been confirmed by many experimental and theoretical results [15–18].

An important concept in modern molecular theory of genetics and biological evolution is epistasis, which means that different genes or mutations are not independent. The epistasis is negative (positive) when the second derivative of the fitness with respect to the number of mutations is negative (positive). Feldman, Christiansen, and Brooks [13] proposed that negative epistasis is needed for recombination to be beneficial, but de Visser and Elena [18] reported that this is not the case at least for viruses. To address the advantage of sex, one needs to find some features, which are valid for rather general situations. Here we solve the evolution models with both neutrality (mutational robustness: the fitness does not change for some mutations) [9] and recombination, and identify clearly the evolutionary advantage of combining both factors.

According to Ref. [19], for intermediate-size populations, genetic drift (fluctuations due to finite population size) and selection together give a stronger benefit to recombination than an appropriate epistasis. In [20] the advantage of recombination has been assumed in case of neutrality. A further advance was in [21], where the numerical calculations reveal some increase of mean fitness in the case of recombination and neutrality; however, there was no analytical theory for the effect. Here, we solve rigorously an evolution model with mutationally robust (neutral) fitness landscape and recombination, and give analytical theory for the phenomenon. Our analytical solutions are consistent with Refs. [20,21]. According to our calculations, the increase of mean fitness due to neutrality and recombination could be stronger than the change due to epistasis [13], found in the HIV virus [22–26]. The effect does not depend much on population size and gives a more general background to the recombination advantage than the mechanism suggested in Ref. [19]. We give analytical theory of the increase of mean fitness both for the HGT and recombination with two-point crossing.

Here we consider the genome of length L as a collection of L nucleotides of two types: $+1$ and -1 [3], and there are 2^L different sequences, labeled by S_i with $0 \leqslant i \leqslant 2^L - 1$. S_i has the probability $p_i(t)$ to appear at time t, and the reproduction rate r_i which is independent of time. In the Crow-Kimura (CK) model [1,4], $p_i(t)$ satisfy coupled differential equations in which the mutation and the reproduction appear in different terms, and the CK model belongs to the parallel mutation-selection scheme. One can subtract a constant from r_i without changing the coupled differential equations for the CK model [4]. In the Eigen model [2], $p_i(t)$ satisfy coupled differential equations in which the mutation and the reproduction appear in the same term, and the Eigen model

*saakian@phys.sinica.edu.tw

†huck@phys.sinica.edu.tw

 ©2013 American Physical Society

belongs to the coupled mutation-selection scheme. The CK model [1,4] and the Eigen model [2] were first studied with the single-peak fitness function (also called landscape), in which one sequence, say S_0, has higher reproduction rate, and other sequences have small reproduction rate. For the CK model with the single peak fitness function, one can simply choose r_0 of S_0 to be $A > 0$, and r_i of other S_i with $i \neq 0$ to be 0.

This paper is organized as follows. In Sec. II, we define three models from the CK model [1,4] with the neutral fitness landscape. The first one is a simple microscopic model of infinite population, with neutral fitness landscape and one-nucleotide exchange during recombination. Models of horizontal gene transfer (HGT) without neutrality were suggested recently [5,6] and some exact results were derived [6,8]. In the second model, there is an exchange of l alleles during one recombination event. The third model corresponds to two-point crossover. For the first and second models, we derive rigorous analytical solutions which are exact in the limit of large genome length L. For the third model we derive a reasonable analytical approximation for the increase of mean fitness due to recombination. We perform numerical calculations for the finite- and infinite-population version of the first model, and finite-population version of the third model to check the reliability of our analytical results. We find that our analytic results are consistent with numerical results. In Sec. III, we consider the dynamics of recombination. In Sec. IV, we summarize and discuss our results. In Appendixes A, B, and C, we give detail derivations for some results of Sec. III.

II. MODEL SYSTEMS

A. HGT model with neutrality

The Crow-Kimura (CK) model with single-peak fitness landscape has been solved in [4]. In the single-peak fitness landscape the 0th sequence has a fitness A and the rest have a fitness 0. Different versions of neutral fitness landscapes have been considered in [9]. Here we consider the simple recombination in an infinite population in the case of the specific neutral fitness landscape: the peak (0th) sequence with only $+1$ nucleotides and its $d \equiv \nu L$ neighbors ($0 < \nu \leqslant 1$) with single mutations have a Malthusian fitness A; other sequences have a fitness 0. Due to neutrality, even without recombination, the population is grouped around the sequence with a large number of neutral neighbors (0th sequence in our case) [27].

Now we first consider the case $\nu = 1$. The infinite-population version of the HGT model could be described via the following system of equations for p_n: total probabilities of sequences in the nth Hamming class (collection of genome types having the same number n of mutations) [5,6,8] with $0 \leqslant n \leqslant L$,

$$\begin{aligned}\frac{dp_n}{dt} &= p_n(r_n - R) + \frac{\mu}{L}[(L-n+1)p_{n-1} + (n+1)p_{n+1}] \\ &\quad - \mu p_n + c\left[\left(1-\frac{\bar{n}}{L}\right)\left(1-\frac{n}{L}\right) + \frac{\bar{n}}{L}\frac{n}{L}\right]p_n - cp_n \\ &\quad + c\left[\left(1-\frac{\bar{n}}{L}\right)\frac{n+1}{L}p_{n+1} + \frac{\bar{n}}{L}\left(1-\frac{n-1}{L}\right)p_{n-1}\right],\end{aligned} \tag{1}$$

where c and μ are the per genome recombination and mutation rates, r_n is the fitness of the sequences from the nth Hamming class, $\bar{n} = \sum_{n=0}^{L} n p_n$, and $R = \sum_n r_n p_n$ is the mean fitness.

This system of equations is well known and was explained in details in [6,8]. The probability of choosing a -1 spin in the given sequence from the $(n+1)$th class is $(n+1)/L$. This spin could be replaced by a $+1$ spin from the sequence pool; the probability of such a choice is $(1 - \frac{\bar{n}}{L})$. Thus we obtain the term $c(1 - \frac{\bar{n}}{L})\frac{n+1}{L}p_{n+1}$ in Eq. (1). The other c proportional terms in Eq. (1) are derived in a similar way.

In the limit of infinite genome length, the mean fitness of the HGT model with the single-peak fitness [6] coincides with the mean fitness R_{sp} of the Crow-Kimura model without recombination [4]

$$R_{sp} = A - \mu. \tag{2}$$

Finite genome length corrections for the single-peak model and the general fitness landscapes with the HGT have been calculated in [8]. In Appendix B we investigate Eq. (1) for the $\nu = 1$ neutral case, $r_0 = A, r_1 = A$, and calculate the mean fitness.

Now we consider the more general case: $0 < \nu \leqslant 1$. For both single-peak and neutral-fitness-like fitness landscape with νL neutral neighbors, the population is concentrated near the 0th Hamming class, and thus $\bar{n}/L \ll 1$. The peak sequence and neutral neighbor sequences have a fitness A, while other sequences have a fitness 0. Neglecting $\bar{n}/L$ terms and the terms from the higher classes, in Appendix A we get following equations for p_0 of the peak sequence and p_1 of νL neutral neighbors,

$$\begin{aligned}\frac{dp_0}{dt} &= p_0((A-\mu) - R) + \frac{(\mu + c)}{L}p_1, \\ \frac{dp_1}{dt} &= p_1((A-\mu) - R) + p_0\mu\nu,\end{aligned} \tag{3}$$

up to $O \sim 1/\sqrt{L}$ relative accuracy.

B. The model of recombination with exchange of l alleles

In the second model (HGT-l), l spins (a genome part) are exchanged during each recombination event. This is supported by experimental data [23]. If we assume that $cl \ll L$, the majority of the population is near the 0th Hamming class; we can again write the equations for class probabilities. In the steady state, extending Eq. (3) for l spin exchange, we have

$$\begin{aligned}p_0(R - (A-\mu)) &= \frac{(\mu + cl)}{L}p_1, \\ p_1(R - (A-\mu)) &= p_0\mu\nu.\end{aligned} \tag{4}$$

During the recombination event, there is an exchange of l nucleotides. The backflow to the wild sequence (first equation) due to recombination is $cl\,p_1/L$. There is also a backflow due to mutations $\mu p_1/L$. The neutral neighbor sequence can be obtained from the 0th sequence only via point mutations, each one with a probability μ/L. As only νL mutations are neutral, we get the multiply factor $\mu\nu$ in the second equation of Eq. (4). The details of the derivation of Eq. (4) are given in Appendix B.

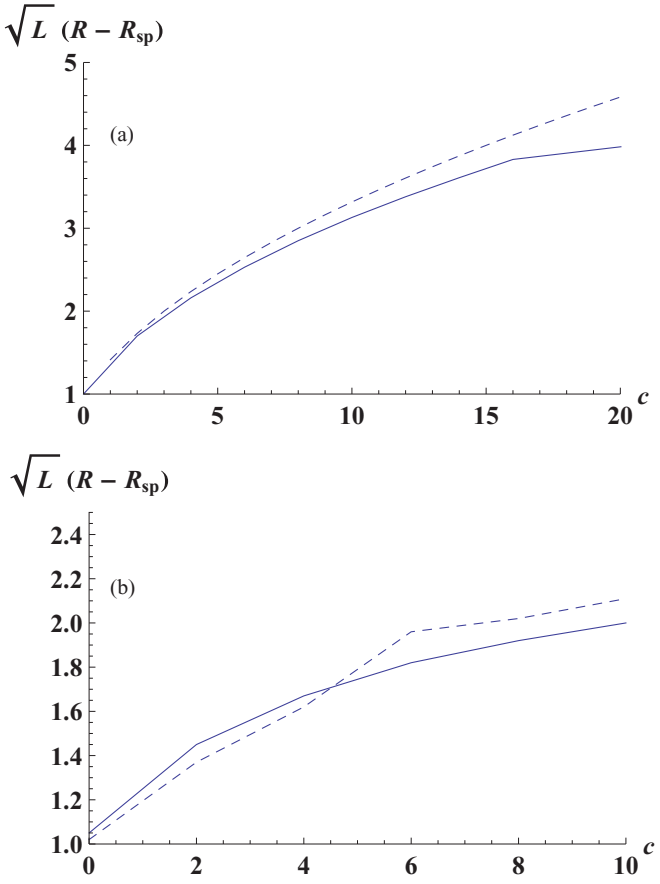

FIG. 1. (Color online) $\sqrt{L}(R - R_{sp}) \equiv \sqrt{L}\Delta R$ as a function of the per genome recombination rate c for an infinite population HGT model with $l = 1, \nu = 1, \mu = 1$, $r_0 = r_1 = 2$, and $r_i = 0$ for $i \geqslant 2$. $R_{sp} = r_0 - \mu$ is the mean fitness of the single-peak fitness model; see Eq. (2) and [4]. (a) The dashed line is the theoretical result for $L = 10\,000$ by Eq. (5), and the solid line is the numerical result. (b) Infinite-population result (solid line) versus finite-population result for $L = 100$ and population size $N = 10\,000$ (dashed line); the solid line was obtained by solving Eq. (1) with a numerical method.

We denote $\Delta R = (R - R_{sp})$ the increase of the mean fitness due to finite genome length, neutrality, and recombination. Equation (4) implies that

$$\Delta R = \mu \frac{\sqrt{\nu(1 + \frac{cl}{\mu})}}{\sqrt{L}}. \tag{5}$$

This is the main result of our work. It demonstrates the *collective* character of the common action of recombination and neutrality.

For the fitness landscape with isolated peaks ($\nu = 0$) with high fitness A, the mean fitness is $R_{sp} + O(1/L)$. The $1/\sqrt{L}$ corrections in our model by Eq. (5) arises due to neutrality, which we define as a collective phenomenon (a result of an interaction of neutral neighbor sequences with high fitness). Equation (5) shows that the effect could become very strong in case of large l.

Our formula Eq. (5) is an accurate analytical estimate for $cl \ll L$ as shown in Fig. 1(a) for the case $l = 1$, $\nu = 1$, and $L = 10\,000$. Figure 1(b) shows that the infinite population case (solid line) obtained by solving Eq. (1) with a numerical method is consistent very well with the finite-population case when $L = 100$ and population size is $N = 10\,000$ (dashed line).

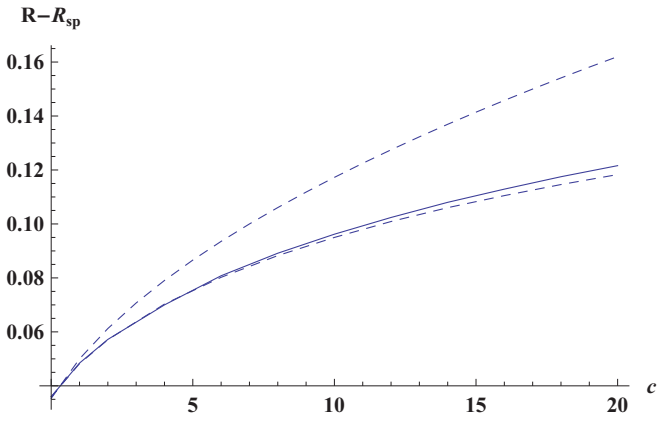

FIG. 2. (Color online) $\Delta R \equiv R - R_{sp}$ versus the recombination rate c for the infinite-population HGT model with $L = 800, r_0 = 2, r_1 = 2$, $r_i = 0, i > 1, \nu = 1, \mu = 1$. The solid line corresponds to the direct numerics. The upper dashed line is by Eq. (5); the lower dashed line is given by higher accuracy formula Eq. (B15).

In Appendix B 3, we calculate ΔR up to the $1/L$ accuracy as Eq. (B15) for the case $l = 1$ and $\nu = 1$. Figure 2 shows that Eq. (B15) is consistent with numerical data better than Eq. (5).

For the HGT model, how the neutral sequences distribute among the nearest neighbors of the peak sequence influences very little the value of ΔR.

C. Two-point crossover recombination

The third model corresponds to the two-point crossovers recombination (2CR): during the recombination event one generates two random points along the genome, and there is an exchange of the piece of genome between such two points.

We first consider the case of full neutrality for the nearest neighbor sequences of the peak sequence S_0: $\nu = 1$. Assuming that the crossover points are randomly chosen along the genome, for the two-point crossover case we get as an effective length $l = L/6$, which is derived in Appendix C. From Eq. (5), at $\nu = 1$ we get an estimate

$$\frac{\Delta R}{\mu} \sim \sqrt{\frac{c}{6\mu}}. \tag{6}$$

Let us assume that both the total mutation rate μ and the recombination rate c are proportional to the genome length L: $\mu = \mu_0 L$, $c = c_0 L$. According to the HIV data in [22] and its analysis [26], we take $\mu_0 = 0.0001$ and $c_0 = 0.002$ according to [24]. Thus we have for $L = 10\,000$ the total mutation rate

$$\mu \sim 1. \tag{7}$$

While Eq. (5) has been confirmed by Fig. 1 for the case $l = 1$, Eq. (6) is a qualitative estimates for the 2CR model. We can write a more general version of Eq. (6) for the mean fitness increase due to recombination in the 2CR model at the limit of the large L:

$$\frac{\Delta R}{\mu} = \phi\left(\frac{c}{\mu}\right), \tag{8}$$

where ϕ is some function.

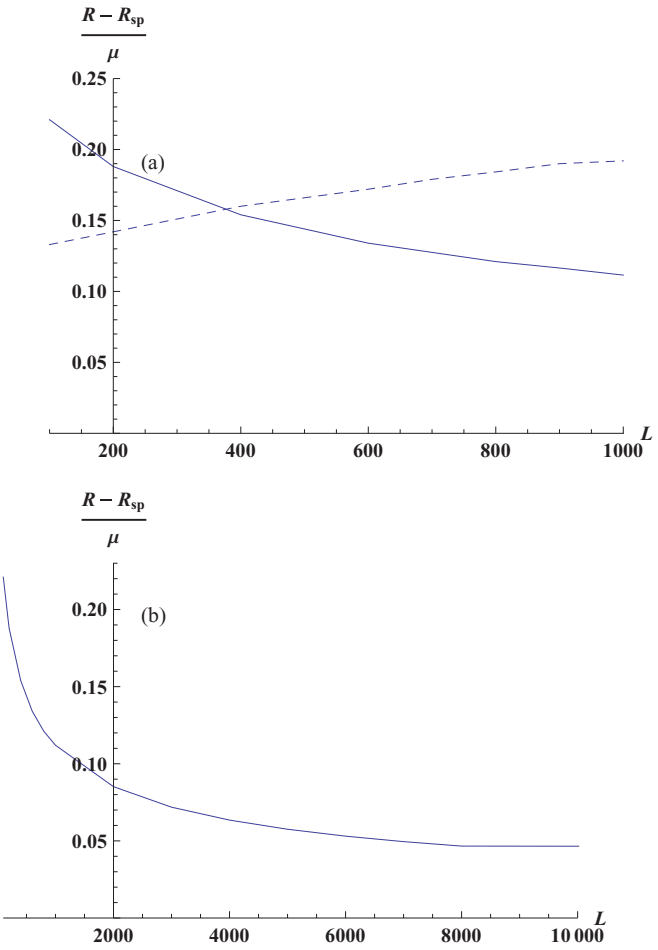

FIG. 3. (Color online) $(\Delta R)/\mu$ versus the genome length L for the 2CR model and the infinite population HGT. (a) For the model with parameters $c/\mu = 20, r_0 = 2, r_1 = 2$, and $r_i = 0$ for $i > 1$, $\nu = 1$, and $L \leqslant 1000$. The solid line corresponds to the infinite population HGT. The dashed line corresponds to the 2CR model with the population size 10^4. (b) For the HGT model with the infinite population and same parameters as (a), but with L up to 10 000.

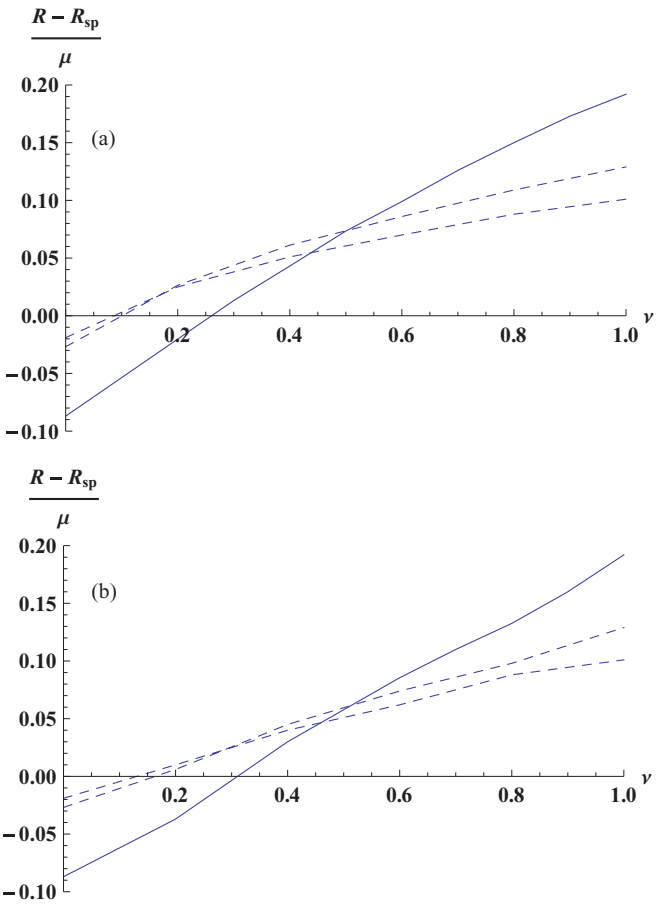

FIG. 4. (Color online) $\Delta R/\mu \equiv (R - R_{sp})/\mu$ versus the degree of neutrality ν for the finite-population 2CR model with population size $N = 10^4, L = 1000, c/\mu = 20$, and $A = 2$. The solid line corresponds to the 2CR model. At $\nu = 1$, the upper dashed line corresponds to the HGT-2 model and the lower dashed line corresponds to the HGT model. (a) All the neutral sequences correspond to the 1 point mutations in the adjacent sites in one part of the genome; (b) the neutral sequences are randomly distributed in the first Hamming class.

Figure 3 shows $\Delta R/\mu \equiv (R - R_{sp})/\mu$ as a function of the genome length L for the infinite population HGT (solid line) and the finite population $N = 10^4$ 2CR model obtained by numerical calculations. Figure 3(a) shows that the dashed line is larger than the solid line for $L > 400$. Our computer facilities do not allow us to perform direct numerics for recombination with $L = 10\,000$. We calculate the $\Delta R/\mu$ for different values of L, and extrapolate our results for the realistic case with $L = 10\,000$. The maximal genome length of our numerics is $L = 2000$, where $\Delta R/\mu \approx 0.2$ for 2CR and 0.084 for infinite population HGT. Our results indicate that for $L = 10^4$ and $c = 20$, we have

$$\Delta R/\mu = 0.2, \tag{9}$$

while the numerics of HGT gives $\Delta R/\mu = 0.042$. Thus $\Delta R/\mu$ of the 2CR model is about 5 times larger than that due to HGT for $L = 10\,000$. For $c = 1$, HGT gives only 0.0145, which is about 10 times smaller than that in the 2CR model. Peck and Waxman [28] claimed that the recombination can change the error threshold.

In the next step, we consider the partial neutrality for the nearest neighbors of the peak sequences S_0; i.e., $\nu < 1$. According to the experimental data, $\nu = 0.27$ for some RNA viruses [29]. Contrary to the result of Fig. 3(a) for $\nu = 1$, now ΔR decreases with the L for the 2CR model.

Some results of our numerical calculations are given in Fig. 4(a) for the case, when all the neutral sequences correspond to the mutations at the adjacent sites. Below some critical value of neutrality μ_c [$\nu_c \approx 0.25$ in Fig. 4(a)], $\Delta R/\mu \equiv (R - R_{sp})/\mu < 0$. Let us try to explain qualitatively this phenomenon.

Equation (24) in [8] implies that for the single peak fitness landscape, there is a decrease of fitness due to the recombination,

$$\Delta R_1 = -\frac{c}{A(A-1)L}. \tag{10}$$

To calculate the change due to neutrality, we should consider both Eq. (5) and (10). For $\nu < \nu_c$, the decrease of Eq. (10) is

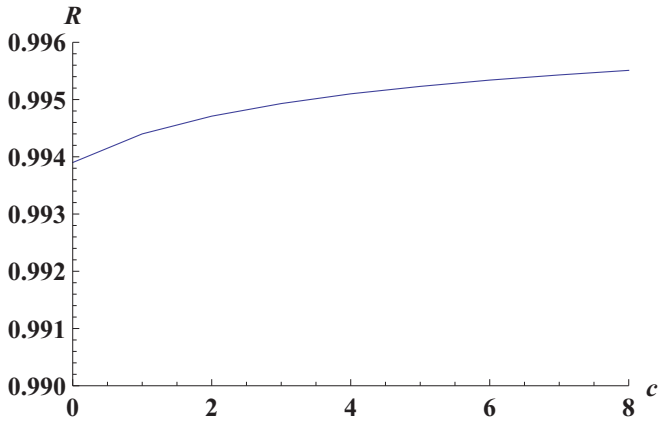

FIG. 5. (Color online) The mean fitness R versus recombination rate c for the model of [26] with $L = 100$ and $\nu = 1$.

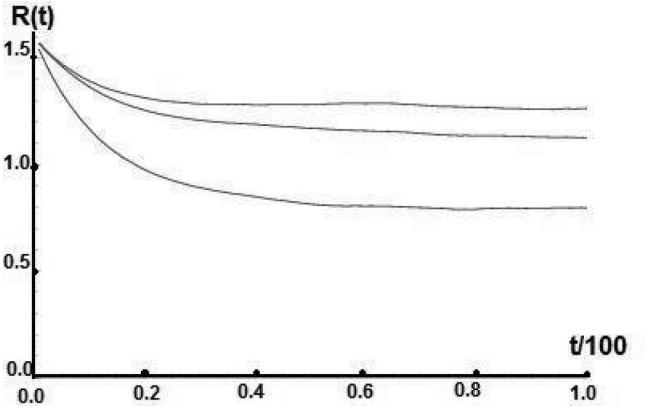

FIG. 6. The dynamics for the model of [26] with the biological motivated values of parameters from [19] with $L = 100$, $N = 10\,000$, $f(m) = 2m + 0.04(1 - m)^2$. Originally the population is located at $L/10$ Hamming distance from the reference sequence. The time scale is chosen to have L mutations during the unit period of time. The mean fitness as a function of $t/100$, t is the time, for $\mu = 1, c = 0$, upper line; $\mu = 1, c = 1$, middle line; $\mu = 2, c = 0$, lower line.

larger than the increase due to neutrality given by Eq. (5) and $\Delta R/\mu$ is negative; for $\nu > \nu_c$, the contribution of Eq. (5) to $\Delta R/\mu$ is larger than that of (10) and $\Delta R/\mu$ is positive.

The $|\Delta R_1|$ for the 2CR model is larger than that for the HGT model because the effective c for the 2CR model is larger. Thus the critical value ν, ν_c, of the 2CR model is larger than those of the HGT models. We also study the case that the neutral sequences are randomly distributed in the first Hamming class and show our calculated results in Fig. 4(b), which is similar to Fig. 4(a).

We perform the numerics for the 2CR model with different values of L and c. We find that ν_c slightly changes with c, but it is strongly affected by L. At $L = 500$, $\nu_c = 0.2$; at $L = 1000$, $\nu_c = 0.25$; and at $L = 2000$, ν_c is about 0.36.

D. The HIV case with fitness from [26]

The realistic fitness landscape of HIV is too complicated, and simple microscopic models have been introduced recently [24,26]. Based on the experimental data of [22], Vijay *et al.* [26] suggested the following fitness landscape:

$$r_n = 1 - 0.731\frac{n^3}{n^3 + (L/2)^3}, \quad n > 1. \tag{11}$$

Although in [22] it has been reported that there is a positive epistasis, the fitness in Eq. (11) does not have a definite epistasis [the sign of the curvature of the function $r(n) \equiv r_n$]: the epistasis is negative for $n < n_c$ and positive for $n > n_c$, where $n_c = L/2^{4/3}$.

Figure 5 shows the results for the model of [26] with fitness of Eq. (11). We see that recombination gives a slight advantage in the infinite population limit. Thus, contrary to that reported in Ref. [22] there is no discrepancy between theory and experimental data.

III. THE DYNAMICS OF RECOMBINATION

Consider now the dynamics in case of HGT with a narrow and symmetric original distribution [26]. For the real biological applications, one needs to investigate how the population moves on a Hamming distance much smaller than the genome length. The virologists measure the diversity and the divergence. Consider P_i as the fraction of population with the ith sequence, $0 \leqslant i \leqslant 2^L - 1$. The divergence is defined as $\alpha = \sum_i P_i d_{i0}/L$, where d_{i0} is the number of mutations (Hamming distance) between sequences 0 and i. We define $m = 1 - 2\alpha$. The diversity is defined as $\pi = \sum_{i,j} P_i P_j d_{ij}/L$, where d_{ij} is the Hamming distance between sequences i and j. For the HIV case the nucleotide diversity after the evolution in the patient is $\pi \approx 0.03$–0.04 [30].

In Ref. [8] the dynamics of recombination has been investigated using the Hamilton-Jacobi-Equation (HJE) [31,32]. In the selection-free case it has been found that $m(t)$ is independent of the recombination rate. We performed a similar analysis and found that for the symmetric fitness function and original distribution, the recombination does not change the dynamics for a rather large periods of time.

We performed numerical simulations for both positive and negative epistasis for the biological motivated values of parameters from [19] with a finite population. Figure 6 illustrates that involving recombination with the same rate as the mutation only slightly affects the divergence [α is defined through the equation $f(1 - 2\alpha(t)) = R$], while the doubling of the mutation rate affects the divergence strongly. On the other hand, we have found that the variance does not distinguish the mutation and recombination; rather they affect the variance similarly. For the population size $N = 10\,000$, the dynamics becomes smoother than in the case when $N = 1000$. In [26] it has also been realized from the numerical simulation that the recombination strongly influences the diversity and less on divergence.

One should distinguish the mean-field-like divergence, mean fitness, haplotype frequency, and fluctuation-like (diversity, mean-fitness variance, diploid genotype frequency) quantities in evolution; see Table I. In this article we found the results (different behavior under mutation and recombination) for the divergence and diversity; the character of the rest of the evolution factors was already known.

TABLE I. Mean-field-like and fluctuation-like evolution quantities.

Mean-field-like evolution quantities	Fluctuation-like quantities
Divergence	Diversity
Mean fitness	Mean fitness variance
Haplotype frequency	Diploid genotype frequency
	Linkage disequilibrium

IV. DISCUSSION

The robustness of fitness landscapes is very common for many viruses [33]. We derived some analytical results for infinite population models with neutral fitness landscape and recombination. Our numerics confirms that our analytical results work well even for the finite populations. The result of common action of neutrality and recombination depends on the concrete version of neutrality and recombination. We considered a simple HGT model with an exchange of a single nucleotide during the recombination event, the HGT with the exchange of l-adjacent nucleotides (HGT-l), and recombination model with two crossover points 2CR. We considered several version of of neutrality, considering neutral networks: the case (a) with complete neutrality at the first Hamming class; the case (b) when all the neutral sequences at the first Hamming class correspond to the mutations at the adjacent sites; in case (c), the neutral sequences are randomly distributed in the first Hamming class.

We found that for the case (a) of complete neutrality in the first Hamming class, the recombination with two crossover points can have much stronger affect on the mean fitness than the horizontal gene transfer. Two point recombination strongly (10 times for $L = 1000$ and certainly more for the real virus genome lengths) increases the contribution of neutrality to the mean fitness.

If only ν fraction of 1-point mutations are neutral, then for $\nu < \nu_c$ the two point recombination suppresses the mean fitness, while HGT increases it. The ν_c slightly depend on the recombination rate and strongly depends on the genome length, for $L = 1000$ it is about 0.25, and for $L = 2000$, $\nu_c \approx 0.36$. These results are derived for the $J = 2$ and it is better to perform a numerics for a concrete version of neutral fitness landscape.

In the dynamics the recombination distinguishes the mean-field-like and fluctuation-like variables. The recombination just acts heavily on the second part, while it only slightly affects the mean-field-like characteristics. In contrast, the mutation is active in both cases. From the evolutionary perspective, it is a serious advantage to have both robust (mainly mean-field-like) and fragile (connected with fluctuations) features.

ACKNOWLEDGMENTS

We thank I. Derenyi for a critical reading and R. Andino for discussion. This work was supported by NSC 100-2112-M-001-003-MY2, NSC 98-2811-M-001-080, NSC 101-2923-M-001-003-MY3, NCTS (North), and Academia Sinica.

APPENDIX A: RECOMBINATION IN A TWO-DIMENSIONAL FITNESS LANDSCAPE

To describe the recombination with the νL neutral neighbors, we consider 2-dimensional (two-block) evolution model with recombination [8,34]. In the two-block evolution model with the block lengths $L_1 \equiv \nu L, L_2 = L(1-\nu)$ and total length $L = L_1 + L_2$, we identify L_1 spins with neutral mutations and L_2 spins with non neutral mutations. We consider the following system of equations [8] for the probabilities $p_{n,m}, 0 \leqslant n \leqslant L_1, 0 \leqslant m \leqslant L_2$:

$$\begin{aligned}\frac{dp_{n,m}}{dt} &= (r_{n,m} - R)p_{n,m} - \mu p_{n,m} \\ &+ \frac{\mu}{L}[(L_1 - n + 1)p_{n-1,m} + (n+1)p_{n+1,m} \\ &+ (L_2 - m + 1)p_{n,m-1} + (m+1)p_{n,m+1}] - cp_{n,m} \\ &+ \frac{c}{L}\left[\left(1 - \frac{\bar{n}}{L_1}\right)(L_1 - n) + \frac{\bar{n}}{L_1}n\right]p_{n,m} \\ &+ \frac{c}{L}\left[\left(1 - \frac{\bar{m}}{L_2}\right)(L_2 - m) + \frac{\bar{m}}{L_2}m\right]p_{n,m} \\ &+ \frac{c}{L}\left[\left(1 - \frac{\bar{n}}{L_1}\right)(n+1)p_{n+1,m}\right. \\ &\left. + \frac{\bar{n}}{L_1}(L_1 - n + 1)p_{n-1,m}\right] \\ &+ \frac{c}{L}\left[\left(1 - \frac{\bar{m}}{L_2}\right)(m+1)p_{n,m+1}\right. \\ &\left. + \frac{\bar{m}}{L_2}(L_2 - m + 1)p_{n,m-1}\right], \end{aligned} \tag{A1}$$

where $\bar{n}, \bar{m}$ are defined as

$$\begin{aligned}\bar{n} &= \sum_{n,m} p_{n,m}n, \\ \bar{m} &= \sum_{n,m} p_{n,m}m. \end{aligned} \tag{A2}$$

We consider the following fitness landscape,

$$r_{0,0} = r_{1,0} = A, \tag{A3}$$

and for all other sequences $r_{n,m} = 0$. Denoting $p_{n,0} = p_n, p_{0,1} = P_1$, we derive

$$\begin{aligned}\frac{dp_0}{dt} &= p_0((A - \mu) - R) + \frac{(\mu + c)}{L}p_1 + \frac{(\mu + c)}{L}P_1, \\ \frac{dp_1}{dt} &= p_1((A - \mu) - R) + p_0\mu\nu, \\ R &= A(p_0 + p_1), \end{aligned} \tag{A4}$$

and

$$\frac{dP_1}{dt} = P_1(-A) + p_0\mu(1 - \nu). \tag{A5}$$

Let as first assume that

$$P_1 \ll p_1, p_0 \ll p_1. \tag{A6}$$

We are interested in the steady state solutions. Dropping P_1 term in the first equation of Eq. (A4), we derive from the first

and second equations of Eq. (A4):

$$\Delta_R \equiv R-(A-\mu)=\sqrt{\mu\nu\frac{(\mu+c)}{L}}\ll 1. \tag{A7}$$

We obtain from the second equation of Eq. (A4) and Eq. (A5):

$$\begin{aligned} p_0 &= p_1\frac{\Delta_R}{\mu\nu}=p_1\sqrt{\frac{(\mu+c)}{\mu\nu L}},\\ P_1 &= \frac{\mu(1-\nu)}{A}p_0. \end{aligned} \tag{A8}$$

Since $\nu L \gg 1$, we have $p_0 \ll p_1$ thus Eq. (A6) is valid.

The third equation in Eq. (A4) gives an equation to define p_1

$$p_1\left(1+\frac{\Delta R}{\nu\mu}\right)=\frac{(A-\mu)}{A}\left(1+\frac{\Delta R}{A-\mu}\right). \tag{A9}$$

Thus we derive up to order ΔR:

$$p_1=\frac{(A-\mu)}{A}\left[1+\Delta R\left(\frac{1}{A-\mu}-\frac{1}{\mu\nu}\right)\right]. \tag{A10}$$

From Eq. (A1) we derive

$$\begin{aligned} p_{n,0} &= p_1\left(\frac{\nu\mu}{A+\Delta_R}\right)^{n-1},\\ p_{1,1} &= \frac{\mu}{A+\Delta_R}[\nu P_1+(1-\nu)p_1]. \end{aligned} \tag{A11}$$

The probabilities for the mixed classes $p_{n,m}$ are calculated recursively,

$$p_{n,m}=\frac{\mu}{A+\Delta_R}[\nu p_{n-1,m}(1-\nu)p_{n,m-1}]. \tag{A12}$$

From what we have discussed above, we can neglect the P_1 term in Eq. (A4) to get Eq. (3) of the main text:

$$\begin{aligned} \frac{dp_0}{dt} &= p_0((A-\mu)-R)+\frac{(\mu+c)}{L}p_1,\\ \frac{dp_1}{dt} &= p_1((A-\mu)-R)+p_0\mu\nu. \end{aligned} \tag{A13}$$

APPENDIX B: THE DERIVATION OF EQ. (4)

1. The case $\nu = 1$

Consider the case $\nu = 1$, the proof can be easily generalized to the case $\nu < 1$. During the recombination there is an exchange of l neighbor spins in the genome. The distribution of p_n for neutral fitness landscape has been investigated well in [9].

We consider the equations near the equilibrium. If the Hamming classes have high fitness A till the maximal distance m, and zero fitness for the higher classes, then the highest population is at the mth Hamming class, $p_m \approx (1-\mu/A)$, and decreases for the higher classes via degrees of μ/A; see Eq. (B9) below. There is much smaller population for the classes $n < m$, $p_{m-1} \sim 1/\sqrt{L}$. In our case $m = 1$. Thus we assume the following scaling for the solutions:

$$p_0 \sim 1/\sqrt{L}, \tag{B1}$$

$p_1 \sim L^0$, p_n for $n > 1$ decreases quickly with n, and therefore

$$\bar{n}\equiv\sum_{n=0} np_n \ll L. \tag{B2}$$

Since p_n for $n \geqslant 3$ are much smaller, we consider below only p_0, p_1, and p_2 and obtain the following system of equations for them:

$$\begin{aligned} \frac{dp_0}{dt} &= \left[p_0((A-\mu)-R)+\frac{(\mu+cl)}{L}p_1\right]+\frac{cl(l-1)}{L^2}p_2,\\ \frac{dp_1}{dt} &= [p_1((A-\mu)-R)+p_0\mu]+ck_1lp_0+\frac{2cl}{L}p_2,\\ \frac{dp_2}{dt} &= [-(\mu+R)p_2+p_1\mu]+ck_2lp_1+\frac{3cl}{L}p_3,\\ R &= A(p_0+p_1), \end{aligned} \tag{B3}$$

where k_1,k_2 are ~ 1 to be derived below.

The terms in the first two equations outside the brackets [. . .] are suppressed via a coefficient $1/L$. Ignoring them we obtain Eq. (3) of the main text for the case $\nu = 1$.

Neglecting high-order terms in Eq. (B3) and considering the steady-state solution, one can get from the first two equations in Eq. (B3)

$$\Delta R \equiv R-A+\mu=\frac{\sqrt{\mu(\mu+cl)}}{\sqrt{L}}\ll 1. \tag{B4}$$

First, ignoring the high-order terms in the second equation in Eq. (B3), we get

$$\frac{p_0}{p_1}=\frac{\Delta R}{\mu}. \tag{B5}$$

Then, using the last equation in Eq. (B3) and Eq. (B5), we get

$$Ap_1\left(1+\frac{\Delta R}{\mu}\right)=A-\mu+\Delta R. \tag{B6}$$

Since $\Delta R \ll 1$, one can neglect the term of order Δ_R^2 and obtain from Eqs. (B4)–(B6) the following equations:

$$\begin{aligned} p_1 &= \frac{A-\mu}{A}\left[1+\frac{\sqrt{\mu(\mu+cl)}}{\sqrt{L}\mu}\left(\frac{\mu}{A-\mu}-1\right)\right],\\ p_0 &= \frac{A-\mu}{A}\frac{\sqrt{\mu(\mu+cl)}}{\mu\sqrt{L}}. \end{aligned} \tag{B7}$$

Thus p_0 is smaller than p_1 by a factor of order $1/\sqrt{L}$.

Consider the equations for p_n with $n \geqslant 2$. The recombination does not change the bulk expression for p_n for $n \geqslant 2$. Thus we have the same expression for p_n/p_1 as before. For $n \geqslant 2$

$$\frac{dp_n}{dt}=-(\mu+R)p_n+p_{n-1}\mu. \tag{B8}$$

One can derive equations for p_n with $n \geqslant 2$ and $\bar{n}$ as follows:

$$\begin{aligned} p_n &= p_1\left(\frac{\mu}{A+\Delta R}\right)^{n-1}[1+O(1/\sqrt{L})],\\ \bar{n} &= \frac{A}{A-\mu}. \end{aligned} \tag{B9}$$

Let us give the details of derivation for higher order terms in Eq. (B3). As the distribution of p_n coincides with the case

$c = 0$ with the accuracy $1/\sqrt{L}$, we write simple, HGT-like expressions for the recombination terms on the right-hand side of Eq. (B3).

Now we first consider the second term in the right-hand side of the first equation in Eq. (B3). There are L sequences with only one -1 spin. For each such sequence, the probability that a recombination of length l contains one -1 spin is l/L. Thus we get a factor $L \times l/L = l$. Multiplying this factor by the rate of such events, c/L, we obtain factor cl/L. Adding to this factor the rate of mutation μ/L, we get the factor $(\mu + cl)/L$.

Consider the last term in the first equation of Eq. (B3). There are $L(L-1)/2$ sequences in the second Hamming class. Among $L(L-1)/2$ sequences in the second Hamming class, there are $(L-1)$ sequences with the adjacent two -1 spins. During one recombination event with the conversion of l adjacent spins, there are $(l-1)$ possibilities to covert two -1 spins into $+1$ spins. There are $(L-2)$ sequences with two -1 spins, at distance 2 from each other. We can convert them to the $+1$ spins in $(l-2)$ ways. There are $(L-l)$ sequences with two -1 spins at the distance l from each other. We can convert them to the $+1$ spins with one piece of l spins. Collection all of these terms among $L(L-1)/2$ sequences in the second Hamming class, we obtain the factor:

$$\begin{aligned}&\frac{2}{L(L-1)}\sum_{n=1}^{l}(L-n)(l-n)\\&\quad=\frac{2}{L(L-1)}\left[Ll^2-\frac{(L+l)l(l+1)}{2}+\frac{l(l+1)(l+2)}{6}\right]\\&\quad\approx\frac{l(l-1)}{L}.\end{aligned}\tag{B10}$$

on the right-hand side of the first equation of Eq. (B3). That is why a small factor $1/L$ arises (p_l is the population of the whole Hamming class). Another c/L coefficient arises as the probability of exchange l allele and eventually we get the factor $cl(l-1)/L^2$ as the coefficient of p_2.

The recombination from the higher Hamming classes to the lower is accompanied with the small coefficient c/L; see the second and third equations in Eq. (B3). The origin of this small coefficient is implicit: the probability of one sequence in the $(n+1)$th Hamming class is L times smaller than for the sequence in the nth Hamming class, and equations are written for the class probabilities in Eq. (B3).

Consider now the recombination terms to the higher classes $\frac{ck_1l}{L}p_0$ and $\frac{ck_2l}{L}p_1$. The first term arises due to recombination event changing one spin in the 0th class. The probability to get a -1 spin due to exchange is $\bar{n}/L = (1-\mu/A)^{-1}/L$ [see the term $\sim \bar{n}/L$ in the last line of Eq. (1) of the main text]; multiplying by l we get $k_1 = \frac{1}{(1-\mu/A)}$. In the same way way we get $k_2 = \frac{1}{(1-\mu/A)}$.

What changes when during the recombination there is an exchange by l alleles which are not adjacent neighbors? Now in the first equation we have on the right-hand side $K\frac{2c}{2L^2}p_2$, $K \sim 1$ instead of $K = l(l-1)/2$; the coefficients k_1, k_2 are also modified. Therefore, again we can get Eq. (4) with accuracy $\sim 1/L$. We proved that the contribution of p_2 terms is $\sim 1/L$ and can be ignored in Eq. (4). In the same way we found that the contribution of $p_n, n \geqslant 3$ also can be ignored in Eq. (4). They are $\sim 1/L^{n-1}$ even smaller than the contribution of p_2.

2. The case $\nu < 1$

We drop the P_1 terms in the equation due to the scaling by Eq. (A8). Now we have for p_0, p_1, p_2:

$$\begin{aligned}\frac{dp_0}{dt}&=\left[p_0((A-\mu)-R)+\frac{(\mu+cl)}{L}p_1\right]+\frac{2cl(l-1)}{L^2}p_2,\\\frac{dp_1}{dt}&=[p_1((A-\mu)-R)+p_0\nu\mu]+\frac{cp_0}{L}\\&\quad+\frac{cl}{L}(2k_1p_2+p_{1,1}),\\\frac{dp_2}{dt}&=[p_2(-\mu-R)+p_1\nu\mu]+\frac{ck_2p_1}{L}+\frac{3cl}{L}p_3,\\R&=A(p_0+p_1).\end{aligned}\tag{B11}$$

We have one new term $clp_{1,1}/L$ in the second equation. As we consider only $\sim 1/\sqrt{L}$ terms in the expression of p_0, this term can be ignored. Considering the steady-state solution of Eq. (B11) without the third equation as in the previous subsection, we derive Eq. (4) in the main text.

3. Higher order expression of ΔR for the case $\nu = 1$ and $l = 1$

In this subsection, we consider the higher order, more accurate expressions for p_0 and p_1 for the case $\nu = 1$ and $l = 1$. In such case, we can have Eq. (1) which implies that in the steady state p_1 and p_0 satify the equations

$$\begin{aligned}\left[(A-R-1)p_0+\frac{1+c}{L}p_1-c\frac{\bar{n}}{L}p_0\right]-\frac{c}{L}\frac{\bar{n}}{L}p_1=0,\\[(A-R-1)p_1+p_0]+2\frac{1+c}{L}p_2-\frac{c}{L}p_1=0,\end{aligned}\tag{B12}$$

where we have set $\mu = 1$. To take into account higher order corrections for p_1 and p_0, we propose the ansatz

$$\begin{aligned}p_1&=\frac{A-1}{A}\left[1+\frac{\sqrt{(1+c)}}{\sqrt{L}}\left(\frac{1}{A-1}-1\right)+\frac{x}{L}\right],\\p_0&=\frac{A-1}{A}\left[\frac{\sqrt{(1+c)}}{\sqrt{L}}+\frac{y}{L}\right],\end{aligned}\tag{B13}$$

and get the following system of equations:

$$\begin{aligned}&\frac{(A-1)^{3/2}}{A}\sqrt{(1+c)}(x+y)+\frac{A-1}{A}\sqrt{(1+c)}x\\&\quad+c\sqrt{(1+c)}\frac{A-1}{A}+(1+c)^{3/2}\frac{A-2}{A-1}=0,\\&\frac{A-2}{A}(1+c)+\frac{(A-1)^2}{A}(x+y)-\frac{(A-1)}{A}x\\&\quad+2c\frac{(A-1)}{A}+2(1+c)\frac{A-1}{A}\left(A+\sqrt{\frac{1+c}{L}}\right)=0,\\&R=(A-1)+\sqrt{\frac{1+c}{L}}+\frac{(A-1)(x+y)}{L}.\end{aligned}\tag{B14}$$

The last equation of Eq. (B14) implies

$$\Delta R=\sqrt{\frac{1+c}{L}}+\frac{(A-1)(x+y)}{L},\tag{B15}$$

The expressions for x and y can be obtained by solving the first two equations of Eq. (B14).

APPENDIX C: FACTOR 1/6 IN THE THIRD MODEL

Consider a genome of length L labeled by integer $i = 1,2,\ldots,L$. One can randomly choose two integers M and N between 1 and L with $1 \leqslant N < M \leqslant L$. With L as the unit of the length and for a large L, the average length $l \equiv M - L$ in unit of L is given by

$$\frac{l}{L} = \int_0^1 dx \int_0^x (x-y)dy = \frac{1}{6}, \tag{C1}$$

where $0 \leqslant y \equiv N/L < x \equiv M/L \leqslant 1$. Thus $l = L/6$.

[1] J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory* (Harper Row, New York, 1970).

[2] M. Eigen, Naturwissenschaften **58**, 465 (1971); M. Eigen, J. J. McCaskill, and P. Schuster, Adv. Chem. Phys. **75**, 149 (1989); D. B. Saakian and C.-K. Hu, Phys. Rev. E **69**, 021913 (2004).

[3] J. Swetina and P. Schuster, Biophys. Chem. **16**, 329 (1982).

[4] D. B. Saakian and C.-K. Hu, Phys. Rev. E **69**, 046121 (2004); for a more general model of the parallel mutation-selection scheme, please read D. B. Saakian, C.-K. Hu, and H. Khachatryan, *ibid.* **70**, 041908 (2004).

[5] E. Cohen, D. A. Kessler, and H. Levine, Phys. Rev. Lett. **94**, 098102 (2005); D. B. Saakian and C.-K. Hu, Proc. Natl. Acad. Sci. USA **103**, 4935 (2006); D. B. Saakian, E. Muñoz, C.-K. Hu, and M. W. Deem, Phys. Rev. E **73**, 041913 (2006).

[6] J.-M. Park and M. W. Deem, Phys. Rev. Lett. **98**, 058101 (2007).

[7] M. S. Boerlijst and S. Bonhoeffer, Proc. R. Soc. London B **263**, 1577 (1996).

[8] Zh. Avetisyan and D. B. Saakian, Phys. Rev. E **81**, 051916 (2010); see also Sec. 4 in Z. Kirakosyan, D. B. Saakian, and C.-K. Hu, J. Stat. Phys. **144**, 198 (2011).

[9] D. B. Saakian, C. K. Biebricher, and C.-K. Hu, PLoS One **6**, e21904 (2011).

[10] Z. Kirakosyan, D. B. Saakian, and C.-K. Hu, J. Phys. Soc. Jpn. **81**, 114801 (2012).

[11] M. Smith, *The Evolution of Sex* (Cambridge University Press, Cambridge, 1978).

[12] J. Felsenstein, Genetics **78**, 737 (1974).

[13] M. W. Feldman, F. B. Christiansen, and L. D. Brooks, Proc. Natl. Acad. Sci. USA **77**, 4838 (1980).

[14] A. S. Kondrashov, Nature (London) **336**, 435 (1988).

[15] N. H. Barton and B. Charlesworth, Science **281**, 1986 (1998).

[16] S. P. Otto and Th. Lenormand, Nat. Rev. Genet. **3**, 252 (2002).

[17] S. Paland and M. Lynch, Science **311**, 990 (2006).

[18] J. A. G. M. de Visser and S. F. Elena, Nat. Rev. Genet. **8**, 139 (2007).

[19] P. D. Keightley and S. P. Otto, Nature (London) **443**, 89 (2006).

[20] J. A. G. M. de Visser, J. Hermisson, G. P. Wagner *et al.*, Evolution **57**, 1959 (2003).

[21] G. J. Szollosi and I. Derenyi, Math. Biosci. **214**, 58 (2008).

[22] S. Bonhoeffer *et al.*, Science **306**, 1547 (2004).

[23] D. N. Levy, G. M. Aldrovandi, O. Kutsch, and G. M. Shaw, Proc. Natl. Acad. Sci. USA **101**, 4204 (2004).

[24] G. Bocharov, N. J. Ford, J. Edwards, T. Breinig, S. Wain-Hobson, and A. Meyerhans, J. Gen. Virol. **86**, 3109 (2005).

[25] G. W. Suryavanshi and N. M. Dixit, PLoS Comput. Biol. **3**, e205 (2007).

[26] N. N. V. Vijay, V. R. Ajmani, A. S. Perelson, and N. M. Dixit, J. Gen. Virol. **89**, 1467 (2008).

[27] E. V. Nimwegen, J. P. Crutchfield, and M. Huynen, Proc. Natl. Acad. Sci. USA **96**, 9716 (1999).

[28] J. R. Peck and D. Waxman, Evolution **64**, 3300 (2010).

[29] R. Sanjuan, A. Moya, and S. F. Elena, Proc. Natl. Acad. Sci. USA **101**, 8396 (2004).

[30] S. D. W. Frost, M. J. Dumaurier, S. Wain-Hobson, and A. J. Leigh Brown, Proc. Natl. Acad. Sci. USA **98**, 6975 (2001).

[31] D. B. Saakian, J. Stat. Phys. **128**, 781 (2007).

[32] K. Sato and K. Kaneko, Phys. Rev. E **75**, 061909 (2007); D. B. Saakian, Z. Kirakosan, and C.-K. Hu, *ibid.* **77**, 061907 (2008).

[33] A. S. Lauring and R. Andino, PloS Pathogens **6**, e1001005 (2010).

[34] D. B. Saakian, Z. Kirakosyan, and C.-K. Hu, Phys. Rev. E **86**, 031920 (2012).