Diploid biological evolution models with general smooth fitness landscapes and recombination

David B. Saakian,[1,2,3] Zara Kirakosyan,[2] and Chin-Kun Hu[1,4,*]
[1]*Institute of Physics, Academia Sinica, Nankang, Taipei 11529, Taiwan*
[2]*Yerevan Physics Institute, Alikhanian Brothers Street 2, Yerevan 375036, Armenia*
[3]*National Center for Theoretical Sciences, Physics Division, National Taiwan University, Taipei 10617, Taiwan*
[4]*Center for Nonlinear and Complex Systems and Department of Physics, Chung Yuan Christian University, Chungli 32023, Taiwan*
(Received 1 June 2007; revised manuscript received 2 April 2008; published 11 June 2008)

Using a Hamilton-Jacobi equation approach, we obtain analytic equations for steady-state population distributions and mean fitness functions for Crow-Kimura and Eigen-type diploid biological evolution models with general smooth hypergeometric fitness landscapes. Our numerical solutions of diploid biological evolution models confirm the analytic equations obtained. We also study the parallel diploid model for the simple case of recombination and calculate the variance of distribution, which is consistent with numerical results.

DOI: 10.1103/PhysRevE.77.061907 PACS number(s): 87.23.Kg, 02.50.−r, 87.15.A−

I. INTRODUCTION

After the development of molecular biology, it is of interest to know whether one can use statistical physics to understand biological evolution at the molecular level. The asexual evolution models by Eigen [1,2] and Crow and Kimura (CK) [3] are especially interesting from the statistical physics point of view. The genetic information of a biological system (e.g., a virus) is stored in DNA or RNA sequence. Eigen and CK used models, similar to a one-dimensional Ising model with N spins, to represent a DNA or RNA sequence of N bases and consider the time evolution of the probability distribution p_i, $1 \le i \le M$, of $M=2^N$ spin configurations $S_i \equiv (s_1^{(i)}, \ldots, s_N^{(i)})$ corresponding to M DNA or RNA sequences with +1 representing purines (R) and −1 pyrimidines (Y). Every sequence S_i is assigned a value of the fitness function r_i, which represents the reproduction rate of S_i. In the simplest case of the single-peak fitness function, there is only one peak configuration or sequence—say, S_1—which has the largest value of the fitness function so that $r_1=A \gg 1$ and $r_i \approx 1$ for $i \neq 1$. S_1 can be chosen to be $(1,1,\ldots,1)$—i.e., all spins take +1—without loss of generality. The jth sequence can change into the ith sequence via mutation. Eigen and CK wrote down coupled differential equations for p_i. In the Eigen model [1,2], the reproduction of a new generation of a sequence and the mutation of such a sequence to a new sequence appear in the same term; thus, it is called the coupled or connected mutation-selection model. In the CK model [3], reproduction and mutation appear in different terms of the differential equation for p_i and thus it is called the parallel mutation-selection model.

Eigen introduced the concept of error threshold to the molecular evolution model. When the mutation rate is smaller than the error threshold, any original distribution of p_i will finally concentrate around the peak configuration S_1 after relaxation, while at a high mutation rate the population will distribute evenly in the whole genetic sequence space. Eigen observed the connection of this phenomenon with the ferromagnetic-paramagnetic phase transition in statistical mechanics so that the error threshold and the degree that the population grouped around the peak configuration correspond to the critical point and the magnetization of the magnetic system, respectively [1,2]. Thus the phases below and above the error threshold are called ferromagnetic (FM) and paramagnetic (PM) phases, respectively.

The Eigen and CK models are weakly nonlinear differential equations, which can be mapped onto linear differential equations after simple nonlinear transformations [4] and allow the application of statistical mechanics [5–7]. After the review paper by Peliti [8] was presented, research on biological evolution models has become a popular subject among theoretical physicists and some exact results have been derived [9–16] using either the maximum principle [10–12] or the mapping of evolution models onto quantum mechanical equations [9] and calculating the mean field by quantum statistical mechanics [13–16].

To describe sexual evolution, one should use diploid evolution models [17–20] and a rigorous mathematical approach is important for such studies [17]. For the case of diploid evolution there are only a few exact results [17,18]. Wiehe, Baake, and Schuster (WBS) [19] have investigated the diploid evolution model. However, in such a case the corresponding (Hardy-Weinberg) equations are too strongly nonlinear to apply directly the methods of statistical mechanics. For the simple case of a single-peak fitness landscape, some exact results [19,20] have been obtained.

The hypergeometric model of diploid evolution in population genetics was first introduced by Barton [21] in 1992, then studied further in [22,23]. The hypergeometric model of [22] could be considered as a direct mathematical generalization of the WBS model [19]. The hypergeometric model considers the multiple-locus evolution with two alleles in any loci [22] in a discrete time, while the WBS model considers a multiple-allele one-locus continuous-time evolution. The mutation phenomenon has been considered in [19,20], but not in [22]. The migration- and frequency-dependent selection version of the hypergeometric model was considered in [23].

In the present paper we solve exactly the diploid evolution model with mutation for the case of a general smooth fitness landscape and calculate the mean fitness and steady-state distribution by extending the method of [24], where the

*huck@phys.sinica.edu.tw

 ©2008 The American Physical Society

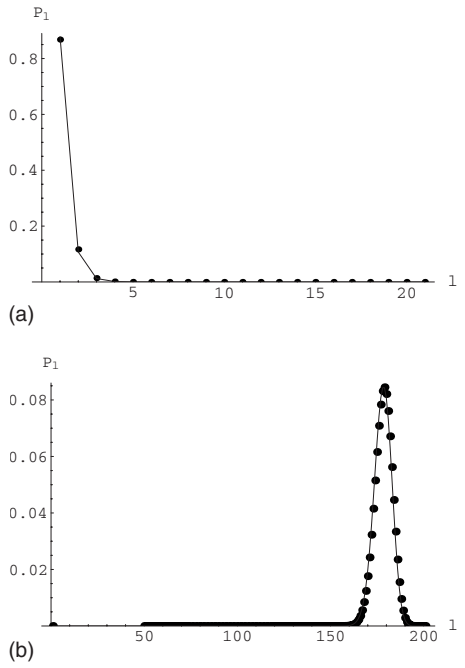

FIG. 1. P_l as a function of l for the fitness function. (a) For Eq. (32) for $N=100$, $s=5.3$, $h=0.1$, and $p=101$. Solid line: analytic solution. Circles: numerical solutions. (b) Equation (34) for $a=4$ $b=0.5$, and $N=200$. Solid line: analytic solution. Circles: numerical solutions.

asexual evolution model has been connected with the Hamilton-Jacobi equation (HJE). Our analytic results are confirmed by numerical solutions and are consistent with results of [19] (the error threshold line BC in Fig. 5, below) for the single-peak fitness function. We investigate also the bistability phase [20] for the coupled diploid model.

Recently there has been progress in the investigation of recombination phenomena in the case of haploid models [25]. It has been formulated as a simple model and investigated numerically [25]. Later, the mean fitness was calculated [26,27]. We consider the [25] version of recombination—i.e., the intragenic conversion [28,29]—and calculate the variance of the distribution for the parallel diploid model, which is consistent with numerical results. We are interested in the steady-state properties of the models. As

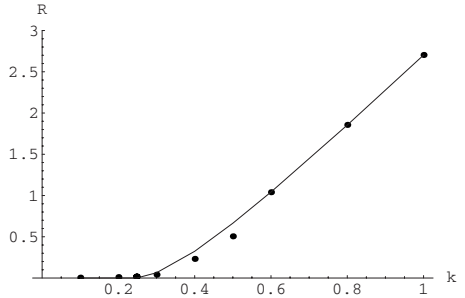

FIG. 2. Mean fitness $R=f(M_0,M_0)$ versus k for the CK model with quadratic fitness $f(m_1,m_2)=k[a(m_1^2+m_2^2)/2+bm_1m_2]$ with $a=4$, $b=0.5$, and $N=100$. Solid line: analytic result obtained with M_0 from Eqs. (34) and (35) including the factor k. Circles: numerical results. Theoretical error threshold is at $k=0.246$.

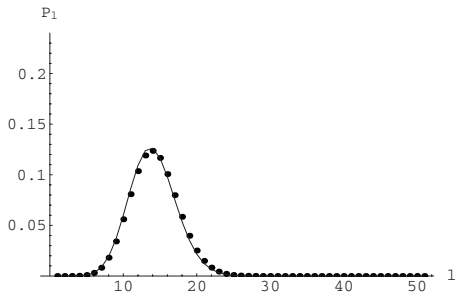

FIG. 3. The theoretical steady-state distribution (solid curve) versus the numerics (circles) for the coupled diploid model with quadratic fitness. The parameters of the model are $N=50$, $a=6$, $b=0.5$, and $\gamma=1$.

the model is symmetric (fitness depends on the Hamming distance from the reference sequence), we have a permutation symmetric steady-state distribution as well.

This paper is organized as follows. In Sec. II, we consider the parallel diploid model, including the single-peak fitness function and general symmetric fitness landscapes. We compare our analytic results with those obtained by numerical calculations and find that they are consistent (Figs. 1 and 2). In Sec. III, we consider the coupled diploid model and find that our analytic results are consistent with numerical calculations (Fig. 3). In Sec. IV, we study the parallel diploid model with recombination and find that our analytic results are consistent with numerical calculations (Fig. 4). Our results are discussed in Sec. V. In the Appendix, we present the stability analysis and the phase diagram for the coupled diploid model with single-peak fitness function (Fig. 5).

II. PARALLEL DIPLOID MODEL

A. Single-peak fitness function

In the WBS parallel diploid model [19], gene probabilities p_i evolve as

$$\frac{dp_i}{dt} = p_i\left[\sum_{j=1}^{M} A_{ij}p_j - \sum_{k=1}^{M}\sum_{j=1}^{M} A_{jk}p_jp_k\right] + \sum_{j=1}^{M} m_{ij}p_j. \qquad (1)$$

Here m_{ij} is a mutation matrix, $m_{ii}=-\gamma$, and $m_{ij}=\gamma/N$ when $d_{ij}=1$ and 0 for other cases with $d_{ij}=(N-\Sigma_{k=1}^{M}s_k^{(i)}s_k^{(j)})/2$ being

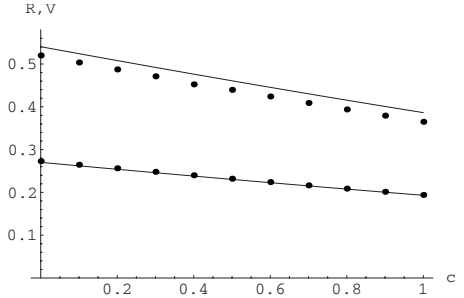

FIG. 4. Comparison of analytic and numerical results for the recombination model of Eq. (53) with the quadratic fitness function of Eq. (34) for $a=1.3$, $b=0.5$, and $N=100$. Lower line: mean fitness R versus c. Upper line: V from Eq. (58) versus c. In both cases, solid lines are analytic solutions and circles are numerical solutions.

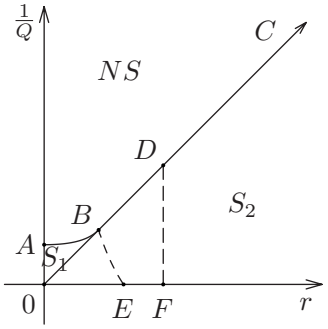

FIG. 5. The phase diagram of the coupled diploid model with single-peak fitness function of Eq. (37) and $R_0=4$. The *OBC* line is the error threshold by WBS, and the *ABC* line is our error threshold. NS is the error catastrophe phase. S_1 is a selective phase with a bistability. In the $(1/Q\text{-}r)$ plane, O has coordinates $(1,1)$, $A=(1,\frac{1+\sqrt{R_0}}{2})$, $B=(\frac{1+\sqrt{1+8R_0}}{4},\frac{1+\sqrt{1+8R_0}}{4})$, $E=(\frac{R_0+2}{3},1)$, $F=(\frac{R_0+1}{2},1)$. The line AB is given by the equation $b^2-4ac=0$.

the Hamming distance between configurations $S_i \equiv (s_1^{(i)}, s_2^{(i)}, \ldots, s_N^{(i)})$ and $S_j \equiv (s_1^{(j)}, s_2^{(j)}, \ldots, s_N^{(j)})$ and representing the number of different symbols in two sequences. We have a balance condition $\Sigma_{i=1}^{M} p_i = 1$. A_{ij} is the fitness of the genotype (S_i, S_j), and $\Sigma_j A_{ij} p_j$ is the marginal fitness for the sequence S_i. WBS have written the single-peak fitness function A_{ij} for the parallel diploid model as [19]

$$A_{11} = 2s, \quad A_{1i} = A_{i1} = 2sh, \quad \text{for } i \neq 1,$$

$$A_{ij} = 0, \quad \text{for } i > 1, \quad j > 1, \tag{2}$$

where $s>0$ and $0 \leq h \leq 1$.

We are looking for the steady-state properties of the model. The sequence consists of the letters ± 1, representing "spins" in physics terminology, or A_1, A_0 alleles in the population genetics hypergeometric model. Due to the symmetry of the fitness function A_{ij} and mutation matrix m_{ij}, there is a corresponding symmetry distribution of p_i so that $p_i = p_j$ when S_i and S_j have the same Hamming distance from the reference sequence S_1. We define the Hamming class l as the collection of all sequences with Hamming distance l from S_1. In each of such sequences, there are l sites with value -1 and the other $N-l$ sites have value $+1$.

To solve the model with $1/N$ accuracy, we assume the following scaling in equilibrium:

$$p_1 \sim 1,$$

$$p_i \sim \frac{1}{N^{(d_{i1}-1)}}, \tag{3}$$

where d_{i1} is the Hamming distance of S_i from the reference sequence S_1. Denoting $x \equiv p_1$, we have the following marginal fitness of the wild sequence $f_1 \equiv \Sigma_{j=1}^{M} A_{1j} p_j$, of other sequences $f_2 \equiv \Sigma_{j=1}^{M} A_{2j} p_j$, and the mean fitness $R \equiv \Sigma_{k=1}^{M} \Sigma_{j=1}^{M} A_{jk} p_j p_k$:

$$f_1 = 2sx + 2sh(1-x), \quad f_2 = 2shx,$$

$$R = xf_1 + (1-x)f_2. \tag{4}$$

In the differential equation for $p_1 \equiv x$, there is a mutational term $p_2 \gamma / N$ proportional to p_2. This term has a small parameter $1/N$. Neglecting this term, we derive with $1/N$ accuracy the equation for x:

$$\frac{dx}{dt} = -x[(2s-4hs)x^2 - 2s(1-3h)x - (2sh-\gamma)]$$

$$\equiv -(1-2h)2sx(x-x_-)(x-x_+). \tag{5}$$

Here $x_\pm$ are solutions of the quadratic equation

$$(2s-4hs)x^2 - 2s(1-3h)x - (2sh-\gamma) = 0$$

on the right-hand side of Eq. (5) and they are given by

$$x_\pm = \frac{(1-3h) \pm \sqrt{(1-h)^2 - 2(1-2h)\gamma/s}}{2(1-2h)}. \tag{6}$$

Applying the stability analysis method used in the Appendix to the present case, we can show that when

$$h > 1/3, \tag{7}$$

the error threshold condition is [19]

$$2sh > \gamma. \tag{8}$$

Having a bulk expression for x, we derive, for R,

$$R = x[2sx + 2sh(1-x)] + 2shx(1-x). \tag{9}$$

At $h<1/3$, there is a bistability [20].

We derive recurrent formulas for p_k, $k>1$, where p_k is the probability of the sequence from the kth Hamming class:

$$p_k = \frac{x\gamma^{k-1}}{N^{k-1}(k-1)!\,(f_1-f_2)^{k-1}}. \tag{10}$$

To derive the last expression, we have neglected in the equations for dp_k/dt mutation terms from higher Hamming classes $\sim p_{k+1}$, as such terms are suppressed by the factor $1/N$. Having explicit expressions for p_1 and p_k, $k \geq 2$, we verify the ansatz of Eq. (3).

B. General symmetric fitness landscapes

To generalize the single-peak fitness, we assume that the fitness of the configuration S_i is a smooth function of the Hamming distance between S_i and the peak configuration S_1. In such a case, it is convenient to work with the overlap $m = (1-2d_{i1}/N)$ instead of the Hamming distance d_{i1}. Consider the following choice of the matrix $\hat{A}$:

$$A_{ii} = f(m,m),$$

$$A_{ij} = f(m_1, m_2), \tag{11}$$

where $m_1 = 1-2d_{i1}/N$ and $m_2 = 1-2d_{j1}/N$. In the hypergeometric model $N(1+m_1)/2$ and $N(1+m_2)/2$ are, respectively, the number of A_1 maternal and paternal alleles. $f(m_1, m_2)$ is a smooth analytical function. If there is a permutation symmetric initial distribution of p_i, the symmetry will be conserved

during the dynamics; see the discussion before Eq. (9) in [22]. We are interested in finding the exact phase structure and the steady state; therefore, we can consider only symmetric solutions of p_i.

The point is that all exact methods for the solution of asexual evolution models—direct mapping to the Schrödinger equation, Suzuki-Trotter, high temperature expansion, functional integral, quadratic form maximum methods (see [24] for a comparison of the different methods)—used the linearity of equations, after the nonlinear transformations [4]. In [24], one of us proposed a method to solve exactly models of asexual evolution by mapping an infinite set of differential equations onto the single HJE. When the original system of equations is linear, we get a simple HJE.

Now we extend the method of [24] to the diploid model. For a symmetric p_i, the total probability of the lth class, P_l, for $0 \le l \le N$ is related to p_i of a typical sequence S_i in that class by

$$p_i = \frac{P_l}{N_l}, \quad N_l = \frac{N!}{l!\,(N-l)!}, \tag{12}$$

where $l \equiv d_{1i}$ is the Hamming distance between S_i and S_1. Choosing $\gamma = 1$, we derive equations for P_l from Eq. (1) for p_i:

$$\frac{dP_l}{dt} = P_l F_l - P_l\left(1 + \sum_k F_k P_k\right) + \frac{1}{N}[(N-l+1)P_{l-1} + (l+1)P_{l+1}]. \tag{13}$$

Here $F_l = \Sigma_{n=0}^{N} f(1-2l/N, 1-2n/N)P_n$. Let us assume that at any moment of time, most sequences concentrate at some Hamming distance $l_0 \equiv N(1-m_0)/2$ and P_n decreases quickly (exponentially) with the distance $(l-l_0)$. On the other hand, $f(1-2l/N, 1-2n/N)$ is a smooth function of l/N. Then, with $1/N$ accuracy, F_l could be replaced by

$$F_l = f(m, m_0). \tag{14}$$

We did not assume any sharpness of the fitness landscape. We just use the sharpness of the distribution P_l and propose the ansatz

$$P_l = \exp[Nu(m,t)], \tag{15}$$

where $m = 1-2l/N$. We find an explicit solution for the function u, confirming our ansatz.

Then we have the following system of equations:

$$\frac{dP_l}{dt} = P_l f\left(1 - \frac{2l}{N}, 1 - \frac{2l_0}{N}\right) + \frac{1}{N}[(N-l+1)P_{l-1} + (l+1)P_{l+1}] - P_l\left[1 + f\left(1 - \frac{2l_0}{N}, 1 - \frac{2l_0}{N}\right)\right]. \tag{16}$$

Our ansatz, Eq. (15), implies that, at $N \to \infty$,

$$P_{l+1} = \exp[Nu(m,t)]\exp[-2u'(m,t)],$$

$$P_{l-1} = \exp[Nu(m,t)]\exp[2u'(m,t)]. \tag{17}$$

Thus we have

$$\frac{Ndu(m,t)}{dt} = f(m,m_0) + \frac{1+m}{2}\exp[2u'(m,t)] + \frac{1-m}{2}\exp[-2u'(m,t)] - 1 - f(m_0,m_0), \tag{18}$$

where the maximum of $\exp[Nu(m,t)]$ is at the point m_0. The large constant N disappeared on the right-hand side of Eq. (18), which is consistent with the ansatz of Eq. (15).

In [24], one of us obtained the equation

$$\frac{Ndu(m,t)}{dt} = f_0(m) + \frac{1+m}{2}\exp[2u'(m,t)] + \frac{1-m}{2}\exp[-2u'(m,t)] - 1, \tag{19}$$

for the CK asexual evolution model [3], after linearization. The solution of Eq. (19) is

$$u(m,t) = \frac{R}{N}t + u_0(m),$$

$$R = \max_m[U(m)], \tag{20}$$

where $u_0(m)$ defines the steady-state distribution (see also [11]). The mean fitness R is defined as the global maximum of

$$U(m) = \sqrt{1-m^2} - 1 + f_0(m), \tag{21}$$

which could be considered as a potential energy. The expression for R was derived in [11] via the maximum principle and in [14] via the Suzuki-Trotter method. It is possible to derive R directly in the HJE approach. The dynamic solution of Eq. (18) is much more involved than the one of Eq. (19), but the steady-state solutions are similar.

Let as assume that at $t \to \infty$ the point of the maximum of $u(m,t)$, m_0, tends to M_0. Then the solution of Eq. (18) should coincide with the solution of the HJE

$$\frac{Ndu(m,t)}{dt} = f(m,M_0) + \frac{1+m}{2}\exp[2u'(m,t)] + \frac{1-m}{2}\exp[-2u'(m,t)] - 1 - f(M_0,M_0). \tag{22}$$

Comparing Eq. (22) with Eq. (19), we see only an extra constant term $-f(M_0,M_0)$ on the right-hand side. Dropping this extra term, we have an equation

$$\frac{Ndv(m,t)}{dt} = f(m,M_0) + \frac{1+m}{2}\exp[2v'(m,t)] + \frac{1-m}{2}\exp[-2v'(m,t)] - 1. \quad (23)$$

The solution of this equation, $v(m,t)$, could be mapped onto the solution of Eq. (22) via a simple formula

$$u(m,t) = v(m,t) - f(M_0,M_0)t/N. \quad (24)$$

Following Eqs. (19) and (20), we have, for the asymptotic solution of Eq. (23),

$$v(m,t) = \frac{R}{N}t + v_0(m),$$

$$R = \max_m[U(m,M_0)]. \quad (25)$$

$v_0(m)$ defines the steady-state solution, and the potential $U(m,M_0)$ is defined as

$$U(m,M_0) = \sqrt{1-m^2} - 1 + f(m,M_0). \quad (26)$$

Having the value of R, we write the solution of u:

$$u(m,t) = [R - f(M_0,M_0)]t/N + v_0(m), \quad (27)$$

where R is given by Eq. (25). On the other hand, we are looking for the steady-state solution. This means that $R = f(M_0,M_0)$. M_0 is the surplus ($M_0 = m_0$ at steady state), defined via the equation

$$M_0 = \sum_l (1-2l/N)P_l. \quad (28)$$

The surplus is the mean value of the overlap. When the population is located at the Hamming distance $N(1-m)/2$ with overlap m, then the surplus equals m.

Thus the mean fitness R and surplus $M_0 \equiv m_0$ are derived via a system of equations

$$R = f(M_0,M_0),$$

$$R = \max_m[U(m,M_0)]. \quad (29)$$

Having the value of M_0, we write the steady-state version of the differential equation (22):

$$f(M_0,M_0) = f(m,M_0) - 1 + \frac{1+m}{2}\exp[2u'(m)] + \frac{1-m}{2}\exp[-2u'(m)]. \quad (30)$$

From Eqs. (15) and (30) we get, for the steady-state solution,

$$P_l = \exp\left[N\frac{1}{2}\int_{-1}^{m} dt \ln\frac{-q+\sqrt{q^2-1+t^2}}{1+t}\right], \quad (31)$$

where $q = f(t,M_0) - 1 - f(M_0,M_0)$ and $m = 1 - 2l/N$.

To check the reliability of the derived equations, let us consider the fitness landscape $A_{ij} \equiv f(m_1,m_2)$ with

$$f(m_1,m_2) = 2(m_1^p + m_2^p)sh + 2s(1-2h)m_1^p m_2^p, \quad (32)$$

which approaches the fitness function of Eq. (2) at $p \to \infty$. We can solve the system of equations (29) for M_0 and the point of maximum $U(m,M_0)$, x, with the following equations:

$$2spx^{p-2}[h + (1-2h)M_0^p] = \frac{1}{\sqrt{1-x^2}},$$

$$2hM_0^p + (1-2h)M_0^{2p} = h(x^p + M_0^p) + (1-2h)x^p M_0^p + \frac{\sqrt{1-x^2}-1}{2s}. \quad (33)$$

After obtaining M_0, we can calculate P_l from Eq. (31) and $R \equiv f(M_0,M_0)$.

With the ansatz $x = 1 - c_1/p^2$, $M_0 = 1 - c_2/p$, we replace $x^p \to 1$ and get for $z = M_0^p$ the stationary version of Eq. (5) for the variable x. Equations (29), (32), and (33) give the expression for R, which coincides with the one by Eq. (4) with x given by Eq. (6). Thus we have verified that our method recovers the result of the single-peak fitness case.

As another example, we consider the fitness landscape $A_{ij} \equiv f(m_1,m_2)$ with the quadratic fitness

$$f(m_1,m_2) = a\frac{(m_1^2+m_2^2)}{2} + bm_1m_2. \quad (34)$$

Equation (29) gives coupled equations for M_0 and the point of maximum $U(m,M_0)$, x:

$$ax + bM_0 = \frac{x}{\sqrt{1-x^2}},$$

$$\frac{a(x^2+M_0^2)}{2} + bxM_0 - 1 + \sqrt{1-x^2} = (a+b)M_0^2. \quad (35)$$

After obtaining M_0, we can calculate the mean fitness R and P_l from Eqs. (29) and (31).

Figure 1 shows that analytic solutions of P_l for the fitness functions of Eqs. (32) and (34) are consistent with numerical solutions. Figure 2 shows that the analytic results for the mean fitness R with the fitness function $f(m_1,m_2) = k[a\frac{(m_1^2+m_2^2)}{2} + bm_1m_2]$ are also consistent with numerical results. There are two pairs of solutions, connected via the transformation $(m,M_0) \to (-m,-M_0)$, but only one pair is stable (the sign of M_0 is chosen from the initial condition).

III. COUPLED DIPLOID MODEL

For the coupled diploid evolution, WBS [19] have proposed the following equation:

$$\frac{dp_i}{dt} = \sum_{j=1}^{M}\left[Q_{ij}\sum_{l=1}^{M}A_{lj}p_lp_j - p_i\left(\sum_{l=1}^{M}A_{lj}p_jp_l\right)\right], \quad (36)$$

where $Q_{ij} = q^{N-d_{ij}}(1-q)^{d_{ij}}$ and is the mutation probability from S_j to S_i with d_{ij} being the Hamming distance between S_j and S_i, and $q^N \equiv e^{-\gamma}$ is the errorless copying probability.

A. Single-peak fitness function

For the single-peak fitness landscape, WBS [19] suggested that

$$A_{11} = (1+s)^2,$$

$$A_{1i} = A_{i1} = (1+s)^{2h}, \quad \text{for } i \neq 1,$$

$$A_{ij} = 1, \quad \text{for } i \neq 1, \quad j \neq 1. \tag{37}$$

WBS gave an error threshold formula: Eq. (19) [19]: $e^{-\gamma} \equiv Q \equiv q^N > 1/(1+s)^{2h}$. It corresponds to *OBC* line in Fig. 5 of the Appendix.

Denoting $p_1 = x$, we have, for the marginal fitness of the peak sequence S_1,

$$g_1 \equiv \sum_{j=1}^{M} A_{1j} p_j = (1+s)^2 x + (1+s)^{2h}(1-x), \tag{38}$$

for the other sequences,

$$g_2 \equiv \sum_{j=1}^{M} A_{2j} p_j = (1+s)^{2h} x + (1-x), \tag{39}$$

and for the mean fitness,

$$F \equiv \sum_{j,l=1}^{M} A_{jl} p_j p_l = x g_1 + (1-x) g_2$$
$$= [(1+s)^2 x^2 + 2(1+s)^{2h} x(1-x) + (1-x)^2]. \tag{40}$$

Neglecting the mutation term which contains Q_{1j} with $j>1$ in Eq. (36), we write the dynamic equation for the $x \equiv p_1$ as

$$\frac{dx}{dt} = -x\{x^2[(1+s)^2 - 2(1+s)^{2h} + 1]$$
$$+ x[-Q(1+s)^2 + Q(1+s)^{2h} + 2(1+s)^{2h} - 2]$$
$$+ 1 - Q(1+s)^{2h}\}. \tag{41}$$

In the Appendix, we use a stability analysis method to find a stable physical solution of Eq. (41). Using such a stable physical solution x' in F, we find the mean fitness F per genome length:

$$F = [(1+s)^2 (x')^2 + 2(1+s)^{2h} x'(1-x') + (1-x')^2]. \tag{42}$$

When

$$(1+s)^{2h} > \frac{1+\sqrt{1+8(1+s)^2}}{4}, \tag{43}$$

there is only one stable solution in Eq. (41). Otherwise, there is a bistability in region S_1 of Fig. 5, which is similar to the parallel diploid model [20].

B. General smooth fitness landscapes

Instead of the single-peak fitness, now we take $A_{ij} = f(m_1, m_2)$. For example, we can take

$$f(m_1,m_2) = [(1+s)^{2h} - 1](m_1^p + m_2^p)$$
$$+ [(1+s)^2 - 2(1+s)^{2h} + 1] m_1^p m_2^p + 1, \tag{44}$$

which becomes the single-peak fitness function of Eq. (37) as $p \to \infty$.

To solve the model with the general fitness landscape, we will again consider the HJE approach of [24], which gives the following equations for the Eigen asexual model and for the mean fitness R in the asymptotic expression $u(m,t) = Rt/N + u_0(m)$:

$$\frac{N \partial u(m,t)}{\partial t} = f_0(m) e^{-\gamma} \exp\left\{ \gamma \left[\cosh\left(2 \frac{\partial u(m,t)}{\partial m} \right) + m \sinh\left(2 \frac{\partial u(m,t)}{\partial m} \right) \right] \right\},$$

$$R = \max\{ f(m) \exp[-\gamma(1 - \sqrt{1-m^2})] \}. \tag{45}$$

In analogy with the previous section for the parallel diploid model, we now derive the following equation for the steady state $P_l = \exp[N u_0(m)]$:

$$f(M_0, M_0) = f(m, M_0) e^{-\gamma} \exp(\gamma \{ \cosh[2u_0'(m)] + m \sinh[2u_0'(m)] \}). \tag{46}$$

We define the mean fitness $R \equiv f(M_0, M_0)$ as

$$f(M_0, M_0) = \max_m [U(m, M_0)], \tag{47}$$

where the potential $U(m, M_0)$ is defined as

$$U(m, M_0) = f(m, M_0) \exp[\gamma(\sqrt{1-m^2} - 1)]. \tag{48}$$

Now we get the distribution P_l from the equation

$$P_l = \exp\left[N \frac{1}{2} \int_1^m dt \ln \frac{q + \sqrt{q^2 - 1 + t^2}}{1+t} \right], \tag{49}$$

where $q = 1 + \frac{1}{\gamma} \ln \frac{f(M_0, M_0)}{f(m, M_0)}$ and $m = 1 - 2l/N$.

For the fitness, given by Eq. (44), and for large p, we use the ansatz $1 - M_0 \sim 1/p^2$ and $1 - m \sim 1/p$; then, Eq. (47) gives

$$Q[(1+s)^{2h} - 1](1+z) + Q[(1+s)^2 - 2(1+s)^{2h} + 1]z + Q$$
$$= 2[(1+s)^{2h} - 1]z + [(1+s)^2 - 2(1+s)^{2h} + 1]z^2 + 1, \tag{50}$$

where $z = m^p, Q = \exp[-\gamma]$. We find that Eq. (47) gives the same expression for the mean fitness per genome length as Eqs. (41) and (42). For $N=100$, $\gamma=1$, and $p=101$, our formula gives for the mean fitness per genome length 2.084, which is very close to 2.087 obtained by numerical calculations.

We also solve the case of the quadratic fitness function

$$f(m_1, m_2) = \frac{1}{2} a (m_1^2 + m_2^2) + b m_1 m_2 + 1. \tag{51}$$

We have used Eqs. (47) and (48) to calculate M_0, then used Eq. (49) to calculate P_l. The results for $N=50$, $a=6$, $b=0.5$,

and $\gamma=1$ are shown in Fig. 3. Our analytic results are consistent with those obtained by numerical calculation.

IV. PARALLEL DIPLOID MODEL WITH RECOMBINATION

For the real diploid evolution, the recombination rate is often much higher than the mutation rate. The recombination in the multiple-loci case is very complicated [17,18]. Here we consider a simple parallel diploid model with recombination, corresponding to the case of intragenic gene conversion [28]. Sometimes it is the primary mechanism of recombination [29]. In the case of the diploid evolution, biologists usually investigate the stability of the states. We investigate only the mean fitness and the variance of the distribution. The latter is one of main targets of a quantitative analysis of the data.

Consider the model with many loci and two alleles at every locus. Now Eq. (1) is modified with a new quadratic term for describing the recombination:

$$\frac{dp_i}{dt} = p_i\left[\sum_j A_{ij}p_j - \sum_{k=1}^{M}\sum_{j=1}^{M} A_{jk}p_j p_k\right] + \sum_j m_{ij}p_j + \sum_{j,l} T^i_{jl}p_j p_l, \tag{52}$$

where p_i are haplotype frequencies, the probability of the ith collection of genes in sexual cells, and T^i_{jl} describes the recombination. Usually one separates the fitness contribution from one-locus and many-locus interactions. We consider the simplest case where there is a symmetry between different loci. Cohen, Kessler, and Levine [25] have considered the recombination matrix T^i_{jl} for haploid models, in which haploid types S_j and S_l exchange a single spin to create the new type S_i. Let us take the matrix of [25,26] to describe the recombination. T^i_{jl} is nonzero only for configurations S_i, S_j, S_l, such that S_i differs from S_j or S_l by one spin. For the simple case of symmetric fitness landscapes we have

$$\begin{aligned}\frac{dP_l}{dt} &= P_l F_l - P_l\left(1 + \sum_k F_k P_k\right) \\ &+ \frac{1}{N}[(N-l+1)P_{l-1} + (l+1)P_{l+1}] \\ &- c\left[\left(1-\frac{\bar{l}}{N}\right)\frac{l}{N}P_l + \frac{\bar{l}}{N}\left(1-\frac{l}{N}\right)P_l\right. \\ &\left. - \left(1-\frac{\bar{l}}{N}\right)\frac{l+1}{N}P_{l+1} - \frac{\bar{l}}{N}\left(1-\frac{l-1}{N}\right)P_{l-1}\right],\end{aligned} \tag{53}$$

where c is the recombination rate, denoted as f_s in [25] [its connection with T^i_{jl} of Eq. (52) can be found in [25,26]], and $\bar{l}=\Sigma_l P_l l$. The last two lines are just the recombination terms of [26], first introduced in [25]. The meaning of those is explicit: any letter of a given sequence could be replaced via the corresponding letter from the genes pool (collection of letters at the same locus in the population). For the sequence from the lth Hamming class the probability of the cut letter to be +1 is $(1-l/N)$ and equal to l/N to be −1. This dropped letter is replaced via the randomly chosen letter from the whole population.

Repeating the derivation of Sec. II, we get, instead of Eq. (18),

$$\begin{aligned}\frac{N du(m,t)}{dt} &= f(m,m_0) - f(m_0,m_0) + \frac{1+m}{2}e^{2u'(m,t)} \\ &+ \frac{1-m}{2}e^{-2u'(m,t)} - 1 - c\frac{(1-mm_0)}{2} \\ &+ c\left[\frac{(1+m)(1-m_0)}{4}e^{2u'}\right. \\ &\left. + \frac{(1-m)(1+m_0)}{4}e^{-2u'}\right].\end{aligned} \tag{54}$$

Repeating the derivations of previous sections, we get an equation for the mean fitness:

$$R = f(M_0,M_0) = \max_m[U(m,M_0)], \tag{55}$$

where

$$\begin{aligned}U(m,M_0) &= f(m,M_0) - \frac{1}{2}c(1-mM_0) \\ &+ \sqrt{(1-m^2)\left[\left(1+\frac{1}{2}c\right)^2 - \frac{1}{4}c^2M_0^2\right]} - 1.\end{aligned} \tag{56}$$

For practical applications it is important to calculate the variance of distributions [30]. Consider the steady-state solution of Eq. (54) near the maximum of distribution at the point $m=M_0$, differentiate Eq. (54), and put $u'(m)=u''(M_0)(m-M_0)$. We get, for the variance,

$$\langle (m-M_0)^2\rangle = \frac{2M_0}{Nf'(M_0,M_0)} \tag{57}$$

and

$$\begin{aligned}N\left\{\left\langle \sum_i p_i\left(\sum_j A_{ij}p_j\right)^2\right\rangle - \left[\left\langle \sum_i p_i\left(\sum_j A_{ij}p_j\right)\right\rangle\right]^2\right\} \\ = 2M_0 f'(M_0,M_0) \equiv V.\end{aligned} \tag{58}$$

We derived the above expressions for the mutation rate $\gamma=1$. For $\gamma\neq 1$ the expression should be multiplied by γ.

We have performed numerical calculations for the quadratic fitness of Eq. (34) and compared the results with analytic calculations. Our theoretical and numerical results are quite consistent as shown in Fig. 4.

V. DISCUSSION

In population genetics it is important to find the steady-state distribution. This distribution has been well investigated mainly for few alleles, although multiple-allele models (a well-known example is the infinite allele model [3]) are more realistic. In recent years, the study of human population via the distribution of a large set of genes has become an impor-

tant topic in genome research [31]. One of the central issues is the existence of a selection pressure for several genes. We have solved the model with a selection for the rather general case (one locus and many alleles or many loci with two alleles at any one). Moreover, it is possible to solve directly our Eq. (18) and get the dynamics for the realistic situation. We derived exact analytical expressions for the steady-state distribution for the multiple-allele model with general fitness landscape for the parallel diploid model, Eq. (31), and coupled diploid model, Eq. (49). We defined the mean fitness $f(M_0,M_0)$ and surplus M_0 in Eq. (29) for the parallel model and in Eq. (47) for the coupled model. Our analytical results for the mean fitness, error threshold, and steady-state distribution are well confirmed by numerical solutions.

The principal difference between the asexual evolution model and diploid evolution model could be clarified by comparing the potentials $U(m)$ of Eq. (21) and $U(m,M_0)$ of Eq. (26). In the case of asexual evolution, we have a single solution for the mean fitness R, while in the diploid case several solutions are possible, as the potential $U(m,M_0)$ depends on the value of the parameter M_0. The existence of several solutions for the extremum problems of a potential $U(m,M_0)$ indicates the possibility of multistability. The existence of different locally stable solutions for the diploid model has been discussed in [20,32]. Park and Deem [26] have investigated this phenomenon in a simple recombination haploid model. We solved a simple case of recombination in a diploid model, assuming a constant recombination rate between loci (regardless of the distance between loci). It is a crude approximation (valid at least for the case of intragenic conversion only [28,29]), but it can catch the qualitative features of the recombination phenomenon.

We have derived the variance of the distribution for the recombination model with general fitness, Eq. (57). The calculation of this variance is one of the main subjects of quantitative population genetics [30]. In our approach we could calculate the population profile for the recombination model, which was impossible to do in the approach of [26]. Our Eq. (57) shows that there is no serious difference in the population distribution for the cases with and without recombination. For the homogenous fitness function $f(m_1,m_2)$, the population variance does not depend on the recombination. Thus, while the recombination suppresses the mean fitness (together with mutation), it plays a quite different role in shaping the population profile and does not change it in the case of homogenous fitness landscapes.

The evolution advantage of this simple form of recombination, if any, could be only in dynamics. There have been claims about the possible advantage of dynamics [25]. To check this hypothesis of [25], one needs a careful investigation of the dynamical solution of Eq. (54) (very similar to the corresponding dynamics of the case without recombination), as well as its finite population version. There is another possibility as well. All models considered in this work, as well as in [25,26], are completely mean-field-like: the fitness, the mutations, and the recombination. The mean-field-like behavior ceases to work for recombination in several situations [33]. There is a chance that the evolution behavior for those cases is different from the one by mean-field-like recombination. The evolution model with differential equation has a more restricted meaning for diploids than for haploids [18,34]. Our results could be generalized for the case of discrete-time models with smooth fitness landscapes. In this paper we considered only symmetric fitness and permutation-invariant distributions. It is possible to write and investigate the HJE equation for the multiple-peak case [15], breaking the permutation symmetry.

ACKNOWLEDGMENTS

This work was supported by CRDF Grants No. ARP2-2647-YE-05, No. NSC 95-2112-M-001-008, and No. AS-95-TP-A07 and the National Center for Theoretical Sciences in Taiwan.

APPENDIX: STABILITY ANALYSIS OF THE SINGLE-PEAK FITNESS COUPLED DIPLOID MODEL

Denoting $R_0=(1+s)^2, r=(1+s)^{2h}$ and neglecting the mutation term which contains Q_{1j} with $j>1$ in Eq. (36), we write the dynamic equation for $x\equiv p_1$:

$$\frac{dx}{dt} = Qg_1x - x^2g_1 - x(1-x)g_2$$
$$= -[R_0 - 2r + 1]x(x-x_-)(x-x_+). \tag{A1}$$

Here

$$x_\pm = \frac{-b \pm \sqrt{b^2-4ac}}{2a}$$

are solutions of the equation

$$ax^2 + bx + c = 0, \tag{A2}$$

where $a=[R_0-2r+1]$, $b=[-QR_0+Qr+2r-2]$, and $c=[1-Qr]$. Solutions x_+ and x_- are real numbers when

$$b^2 - 4ac \geq 0. \tag{A3}$$

We can use signs of the coefficients a, b, and c of Eq. (A2) to analyze the stability of solutions $x=0$, $x=x_-$, and $x=x_+$ of Eq. (A1). Figure 5 shows the phase diagram in the $1/Q$-r plane. In Fig. 5, $a=0$ is the vertical line through $r=(R_0+1)/2$ with $a>0$ and $a<0$ representing regions on the left-hand and right-hand sides of the line, respectively; $c=0$ is the diagonal line OBC with $c<0$ and $c>0$ representing regions below and above the line, respectively; $b=0$ is the hyperbolic line BE, $1/Q=(R_0-r)/2(r-1)$, with $r=1$ and $1/Q=-1/2$ as asymptotic axes, and $b<0$ and $b>0$ representing regions below and above the line, respectively. The hyperbolic line goes through point B in Fig. 5 and passes through the horizontal line $1/Q=1$ at point E with $r=(R_0+2)/3<(R_0+1)/2$.

The inequality (A3) is represented by line AB in Fig. 5; the points in phase S_1 satisfy the inequality. For $r=1/Q$, inequality (A3) gives the coordinate of the point B: $([1+\sqrt{1+8R_0}]/4,\ [1+\sqrt{1+8R_0}]/4)$. At $r=1$, inequality (A3) is simplified as

$$Q \geq \frac{2}{1+\sqrt{R_0}}, \tag{A4}$$

which gives the coordinate of the point A as $(1,[1+\sqrt{R_0}]/2)$.

The lines $a=0$, $b=0$, $c=0$ and the line $b^2-4ac=0$ in the region $a>0$, $b<0$, $c>0$ divide the plane $r\geq 1$ and $1/Q \geq 1$ into seven regions. We can analyze the stability of solutions $x=0$, $x=x_-$, and $x=x_+$ of Eq. (A1) in these regions as follows.

(i) $a<0$, $b>0$, and $c<0$. From the inequalities $r>(R_0+1)/2$ and $Qr>1$, we have the inequality

$$b^2-4ac=d^2+(a-c)^2-2d(a+c)>0,$$

where $d=R_0(1-Q)>0$. Therefore we have $x_->x_+>0$. Near x_+, to the leading order we have

$$\frac{dx'}{dt}=-Ax',$$

where $A=ax_+(x_+-x_-)>0$ and $x'=x-x_+$. It is easy to derive that $x'\sim\exp(-At)\to 0$ for $t\to\infty$ and x_+ is a stable solution. We use a similar analysis to find that the solutions $x=x_-$ and $x=0$ are unstable. Thus, in the steady state, the system will reach the stable solution $x=x_+$.

Using a similar analysis, we can reach following results.

(ii) $a>0$, $b>0$, and $c<0$. $x_+>0$, $x_-<0$, and $|x_-|>x_+$. We find that x_+ and x_- are stable solutions, but $x=0$ is not a stable solution. As initially $x(0)>0$, only the $x=x_+$ stable solution is relevant. Thus, in the steady state, the system will choose the physical solution $x=p_1=x_+$.

(iii) $a>0$, $b<0$, and $c<0$. We have $x_+>|x_-|>0$ and $x_-<0$. There is only one physical stable solution x_+.

(iv) $a<0$, $b>0$, and $c>0$. $x_+<0<x_-$. Solutions $x=x_-$ and $x=x_+$ are unstable; the solution $x=0$ is stable. The system will reach the physical solution $x=0$ in the steady state.

(v) $a>0$, $b>0$, and $c>0$. When $b^2-4ac>0$, we have $x_-<x_+<0$. Solutions $x=x_-$ and $x=0$ are stable; the solution $x=x_+$ is not stable. The system will reach the physical solution $x=0$ in the steady state.

(vi) When $b^2-4ac<0$, $x=0$ is the only stable physical solution.

(vii) $c>0$ and $a>0$. When $b^2-4ac>0$, we have $0<x_-<x_+$. Solutions $x=0$ and $x=x_+$ are stable and the solution $x=x_-$ is unstable. When the initial value of $x(t)$, $x(0)$, is smaller than x_-, the system will be driven to the fixed point $x=0$; when $x(0)$ is larger than x_-, the system will be driven to the fixed point $x=x_+$. Thus the final steady state depends on the initial condition $x(0)$. The bistable phase is represented by S_1 (OBA) in Fig. 5.

When $c<0$, regions 1, 2, and 3 form the selective phase S_2 of Fig. 5. When $c>0$, regions 4, 5, and 6 form the nonselective phase NS and region 7 has a bistable phase S_1. The phase S_2 has been found in [19], while the phase S_1 has been missed. The phase S_1 has been found later for the parallel model in [20].

Let us derive the steady-state distribution. For the distribution p_i we have

$$p_2=\frac{\gamma x}{N}\frac{g_1}{g_1-g_2},$$

$$p_{n+1}=\frac{p_1 g_1\left(\frac{\gamma}{N}\right)^n+\left[\cdots\frac{n!p_{l+1}\left(\frac{\gamma}{N}\right)^{n-l}}{(n-l)!l!}+\cdots+np_n\frac{\gamma}{N}\right]g_2}{g_1-g_2}, \tag{A5}$$

where $p_1\equiv x$ and f_1,f_2 are defined by Eqs. (38) and (39).

[1] M. Eigen, Naturwissenschaften **58**, 465 (1971).
[2] M. Eigen, J. McCaskill, and P. Schuster, Adv. Chem. Phys. **75**, 149 (1989).
[3] J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).
[4] C. L. Thompson and J. L. McBride, Math. Biosci. **21**, 127 (1974); B. L. Jones, R. H. Enns, and S. S. Rangnekar, Bull. Math. Biol. **38**, 15 (1976).
[5] I. Leuthausser, J. Stat. Phys. **48**, 343 (1987).
[6] P. Tarazona, Phys. Rev. A **45**, 6038 (1992).
[7] S. Franz and L. Peliti, J. Phys. A **30**, 4481 (1997).
[8] L. Peliti, e-print arXiv:cond-mat/9712027.
[9] E. Baake, M. Baake, and H. Wagner, Phys. Rev. Lett. **78**, 559 (1997).
[10] E. Baake and H. Wagner, Genet. Res. **78**, 93 (2001).
[11] J. Hermisson, O. Redner, H. Wagner, and E. Baake, Theor. Popul. Biol. **62**, 9 (2002).
[12] E. Baake, M. Baake, and A. Bover, J. Math. Biol. **50**, 83 (2005).
[13] D. B. Saakian and C.-K. Hu, Phys. Rev. E **69**, 021913 (2004); **69**, 046121 (2004).
[14] D. B. Saakian, C.-K. Hu, and H. Khachatryan, Phys. Rev. E **70**, 041908 (2004).
[15] D. B. Saakian, E. Munoz, C.-K. Hu, and M. W. Deem, Phys. Rev. E **73**, 041913 (2006).
[16] D. B. Saakian and C.-K. Hu, Proc. Natl. Acad. Sci. U.S.A. **103**, 4935 (2006); A recent review about quantum spin models of biological evolution can be found at C.-K. Hu, in *Noise and Fluctuations*, AIP Conf. Proc. No. 922, edited by M. Tacano,Y. Yamamoto, M. Nakao (AIP, New York, 2007), pp. 589–594.
[17] W. J. Ewens, *Mathematical Population Genetics* (Springer-Verlag, New York, 2004).
[18] R. Burger, *The Mathematical Theory of Selection, Recombination, and Mutation* (Wiley, New York, 2000).
[19] T. Wiehe, E. Baake, and P. Schuster, J. Theor. Biol. **177**, 1 (1995).
[20] E. Baake and T. Wiehe, J. Math. Biol. **35**, 321 (1997).
[21] N. H. Barton, Evol. Int. J. Org. Evol. **46**, 551 (1992).
[22] M. Shpak and A. S. Kondrashov, Evol. Int. J. Org. Evol. **53**, 600 (1999).
[23] N. H. Barton and M. Shpak, Theor. Popul. Biol. **57**, 249 (2000).

[24] D. B. Saakian, J. Stat. Phys. **128**, 781 (2007).
[25] E. Cohen, D. A. Kessler, and H. Levine, Phys. Rev. Lett. **94**, 098102 (2005).
[26] J.-M. Park and M. W. Deem, Phys. Rev. Lett. **98**, 058101 (2007).
[27] D. B. Saakian and E. Munoz (unpublished).
[28] T. Wiehe, J. Mountain, P. Parham, and M. Slatkin, Genet. Res. **75**, 61 (2000).
[29] P. Parham, E. J. Adams, and K. L. Arnett, Immunol. Rev. **143**, 141 (1995).
[30] D. S. Falconer and T. F. C. Mackay, *Introduction to Quantitative Genetics* (Longman, Essex, 1996).
[31] E. Pennisi, Science **316**, 1690 (2007).
[32] R. J. Burger, J. Theor. Biol. **101**, 585 (1985).
[33] M. Shpak and K. Atteson, Bull. Math. Biol. **64**, 703 (2002).
[34] I. Hildebrandt and E. Baake, e-print arXiv:0612024.