Biological evolution in a multidimensional fitness landscape

David B. Saakian,[1,2,3,*] Zara Kirakosyan,[2] and Chin-Kun Hu[1,†]
[1]*Institute of Physics, Academia Sinica, Nankang, Taipei 11529, Taiwan*
[2]*Yerevan Physics Institute, Alikhanian Brothers St. 2, Yerevan 375036, Armenia*
[3]*National Center for Theoretical Sciences (North): Physics Division, National Taiwan University, Taipei 10617, Taiwan*
(Received 15 March 2012; revised manuscript received 7 May 2012; published 20 September 2012)

We considered a multiblock molecular model of biological evolution, in which fitness is a function of the mean types of alleles located at different parts (blocks) of the genome. We formulated an infinite population model with selection and mutation, and calculated the mean fitness. For the case of recombination, we formulated a model with a multidimensional fitness landscape (the dimension of the space is equal to the number of blocks) and derived a theorem about the dynamics of initially narrow distribution. We also considered the case of lethal mutations. We also formulated the finite population version of the model in the case of lethal mutations. Our models, derived for the virus evolution, are interesting also for the statistical mechanics and the Hamilton-Jacobi equation as well.

DOI: 10.1103/PhysRevE.86.031920 PACS number(s): 87.23.Kg, 87.10.Mn

I. INTRODUCTION

The investigation of biological evolution models [1–5] is one of the most fruitful applications of statistical mechanics or theoretical physics to biological problems [6–23]. To solve the evolution models, one can apply the whole machinery of modern theoretical physics: spin-glass physics methods [11], quantum statistical mechanics [12,15–18], quantum field theory [15,18], and the Hamilton-Jacobi equation (optimal control) [19–21].

The genome, a collection of genes with different types, could be considered as a particular spin configuration of a statistical system, where the fitness (the rate to produce offspring of the given genome) is equivalent to the Hamiltonian of the spin system. In evolution theory, the notion of fitness is central in defining the general features of evolution or in modeling a concrete experiment. Fitness is a complicated function of gene content (types of genes) of the genome in sequence space; this function is assumed to have a mean-field like behavior. Most of the investigations have been devoted to the symmetric fitness case, when there is a master (reference) sequence, and fitness (energy) is a simple function of the Hamming distance from that sequence [2]. In [18], a generalization of symmetric fitness landscape was considered, when there are some K reference sequences, and the fitness is a function of K Hamming distances from these reference sequences. In [24–26], there were suggested evolution models where the genome consists of different blocks and the fitness is a function of the gene mean types at different blocks. In the current article, we follow the idea of [24], considering an infinitely long genome, a collection of a finite number of blocks, defining mean "magnetizations" at any such block and the fitness as a function of block magnetizations. We then use the Hamilton-Jacobi equation [19] to solve the equation. This approach is more powerful and technically easier than that used in [18]. Thus, in the present paper we can calculate the mean fitness of a recombination model in a multidimensional fitness landscape.

Recombination is one of the key factors in evolution. The mathematical aspects of recombination were analyzed in [27–29]. Recently, there was good progress in the statics of recombination [22,23] and there was some advance in the dynamics [30]. We formulate the recombination model in a multidimensional fitness landscape for a many-loci haploid model with two alleles (type of gene) at any locus (position of a gene in the genome).

The rest of the paper is organized as follows: In Sec. II, we formulate and solve (calculate the mean fitness for) the evolution model with selection and mutation in a multidimensional fitness landscape, including the case of lethal mutations [31,32]. We consider two-block models for the lethal mutations and an asymmetric initial distribution. In Sec. III, we formulate the recombination model in a multidimensional space. While we could not calculate the mean fitness, we derive a general result regarding the dynamics of population for the initial narrow distribution. In Sec. IV, we summarize our results and discuss problems for further research.

II. THE MULTIDIMENSIONAL MODEL

A. The model

We identify the alleles as spins and consider the genome as a collection of L spins taking the values ± 1. In the peak configuration, all spins take the value $+1$. Our model is a simple generalization of the Crow-Kimura model [4,12]. The genome is a collection of H pieces (blocks), with the length $L_n, 1 \leqslant n \leqslant H$, such that $\sum_{n=1}^{H} L_n = L$.

Any sequence is characterized by $l_1, \ldots, L_H$, the number of "$-$" (negative) spins in the blocks. We introduce the "magnetization" m_n, defined as

$$m_n = 1 - \frac{2l_n}{L_n}, \tag{1}$$

at the nth piece of genome for all of n with $1 \leqslant n \leqslant H$. Our fitness r is a function of $(l_1, \ldots, l_H)$. Thus, we define $r_{l_1,\ldots,l_H} \equiv Lf(m_1, \ldots, m_H)$. The discrete variables l_n are defined in the

*saakian@phys.sinica.edu.tw
†huck@phys.sinica.edu.tw

interval $[0,L_n]$, while the magnetizations m_n are becoming continuous at the limit $N \to \infty$ and $-1 \leqslant m_n \leqslant 1$. We define the function $f(m_1,\ldots,m_H)$ as a fitness function.

The description of the mutation process is the principal point in the definition of the model. In order to describe mutations we use the coefficients $x_\pm(l_n,L_n)$:

$$x_+(l_n,L_n) = \frac{L_n - l_n}{L_n}, \quad x_-(l_n,L_n) = \frac{l_n}{L_n}, \tag{2}$$

where L_n denotes the length of the nth piece, and $x_+(l_n,L_n)$ and $x_-(l_n,L_n)$ are the fractions of $+$ and $-$ spins in the nth piece, respectively.

If the initial distribution of the population is symmetric, i.e., all the sequences with the same l_n have the same probability, we describe the system through the $p(l_1,\ldots,l_H,t)$, the probability of all sequences having $l_1,\ldots,l_H$ minus spins in corresponding blocks. Then we write the following system of equations:

$$\begin{aligned}
&\frac{dp(l_1,\ldots,l_H,t)}{dt} \\
&\quad = (r_{l_1,\ldots,l_H} - LR)p(l_1,\ldots,l_H,t) - Lp(l_1,\ldots,l_H,t) \\
&\qquad + \sum_{\beta=\pm1,n} L_n x_\beta(l_n-\beta,L_n)p(l_1,\ldots,l_n-\beta,\ldots,l_H,t), \\
&R = \frac{1}{L}\sum_{0\leqslant l_n\leqslant L_n} p(l_1,\ldots,l_H,t)r_{l_1,\ldots,l_H}.
\end{aligned} \tag{3}$$

The sum over n extends from 1 to H and R and is the mean fitness. We considered mutations independently at all pieces of the genome; thus, in the middle line at the right hand side of Eq. (3) we changed an l_n at the position n, where n changes from 1 (first piece of genome) until H (the last piece of genome).

The current configuration $(l_1,\ldots,l_n,\ldots,l_H)$ could be obtained from either $(l_1,\ldots,l_n+1,\ldots,l_H)$, reversing one of the l_n+1 $-$ spins, or from the $(l_1,\ldots,l_n-1,\ldots,l_H)$ configuration, reversing one of (L_n-l_n+1) $+$ spins. There are (l_n-1) such possibilities for the first case, and (L_n-l_n+1) for the second case. Dividing by L_n, we derived the coefficients $x_-(l_n+1,L_n)$ and $x_+(l_n-1,L_n)$ in Eq. (3). For $H=1$, Eq. (3) coincides with the Crow-Kimura model [4,13,16].

Let us consider the linear part of the latter equation and write an equation for $P(m_1,\ldots,m_H,t) \equiv p(l_1,\ldots,l_H,t)$:

$$\begin{aligned}
&\frac{dP(m_1,\ldots,m_H,t)}{dt} \\
&\quad = L(f(m_1,\ldots,m_H)-1)P(m_1,\ldots,m_H,t) \\
&\qquad + \sum_{\beta=\pm1,1\leqslant n\leqslant H} L_n\left(\frac{1+\beta m_n}{2} + \frac{1}{L_n}\right) \\
&\qquad \times P\left(m_1,\ldots,m_n+\frac{2\beta}{L_n},\ldots,m_H,t\right).
\end{aligned} \tag{4}$$

Following [33,34], we define the mean fitness in the steady state of Eq. (3) as the largest eigenvalue of the quadratic matrix on the left hand side of Eq. (4).

Following [19], we assume an ansatz:

$$P(m_1,\ldots,m_H,t) = \exp[Lu(m_1,\ldots,m_H,t)]. \tag{5}$$

Then with $1/L$ accuracy we get the following Hamilton-Jacobi equation (HJE):

$$\begin{aligned}
&\frac{\partial u(m_1,\ldots,m_H)}{\partial t} \\
&\quad = -H\left(m_1,\ldots,m_H;\frac{\partial u}{\partial_{m_1}},\ldots,\frac{\partial u}{\partial_{m_H}}\right), \\
&\qquad -H(m_1,\ldots,m_H;\hat{P}_1,\ldots,\hat{P}_H) = f(m_1,\ldots,m_H) \\
&\qquad -1 + \sum_{1\leqslant n\leqslant H}\frac{L_n}{L}\left(\frac{1+m_n}{2}e^{2\hat{P}_n} + \frac{1-m_n}{2}e^{-2\hat{P}_n}\right),
\end{aligned} \tag{6}$$

where we missed $O(1/L)$ terms and introduced the momenta $\hat{P}_n = \partial u/\partial m_n$.

Considering the asymptotic solution

$$u(m_1,\ldots,m_H,t) = Rt + u_0(m_1,\ldots,m_H), \tag{7}$$

we get an equation

$$\begin{aligned}
R = {} & f(m_1,\ldots,m_H) - 1 \\
& + \sum_{1\leqslant n\leqslant H}\left[\frac{L_n}{L}\frac{1+m_n}{2}e^{2\frac{\partial u_0(m_1,\ldots,m_n,\ldots,m_H)}{\partial m_n}}\right. \\
& \left. + \frac{L_n}{L}\frac{1-m_n}{2}e^{-2\frac{\partial u_0(m_1,\ldots,m_n,\ldots,m_H)}{\partial m_n}}\right].
\end{aligned} \tag{8}$$

On the other hand, we have a condition that at any point m, our R should be higher than the minimum of the right hand side, considered as a function of momenta $\frac{\partial u_0(m_1,\ldots,m_n,\ldots,m_H)}{\partial m_n} = \partial u/\partial m_n = \hat{P}_n$. We define

$$\begin{aligned}
U(m_1,\ldots,m_H;\hat{P}_1,\ldots,\hat{P}_H) = \min\Bigg[& f(m_1,\ldots,m_H) - 1 \\
& + \sum_{1\leqslant n\leqslant H}\left[\frac{L_n}{L}\frac{1+m_n}{2}e^{2\hat{P}_n} + \frac{L_n}{L}\frac{1-m_n}{2}e^{-2\hat{P}_n}\right]\Bigg].
\end{aligned}$$

Examining the solution of the minimum problem with respect to $(\hat{P}_1,\ldots,\hat{P}_H)$ and looking at different points m, we find

$$\begin{aligned}
R &\geqslant \max[U(m_1,\ldots,m_n)]|_{m_1,\ldots,m_H}, \\
U(m_1,\ldots,m_H) &= f(m_1,\ldots,m_H) - 1 \\
&\quad + \sum_{1\leqslant n\leqslant H}\frac{L_n}{L}\sqrt{1-m_n^2}.
\end{aligned} \tag{9}$$

In Eq. (9) we take the maximum in the domain $-1 \leqslant m_n \leqslant 1$. The function $U(m_1,\ldots,m_H)$ is the equivalent of the potential in classic mechanics.

Following [19], we identify the mean fitness [the maximum eigenvalue of the matrix on the right hand side of Eq. (4)] with the lower bound of Eq. (9),

$$R = \max[U(m_1,\ldots,m_H)]|_{m_1,\ldots,m_H}. \tag{10}$$

One can calculate the mean fitness R by differentiating the function $U(m_1,\ldots,m_H)$.

Thus, we defined the mean fitness for the general multidimensional mean-field like fitness landscape for the evolution model with selection and mutation.

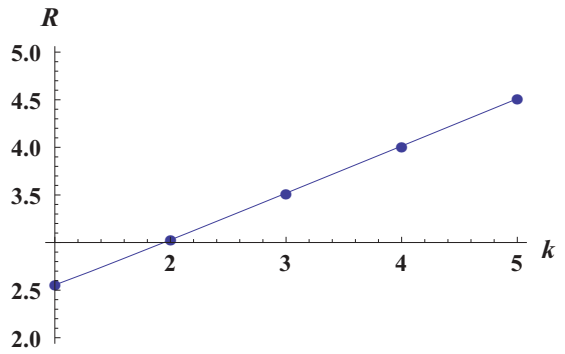

FIG. 1. (Color online) The comparison of analytical result (smooth line) with the numerics (dots) for the 3D model with $L_1 = L_2 = L_3 = 20$. The whole genome mutation rate is 1. The first part of the genome has a fitness $f_1(m_1) = km_1^2/2$. In the second part all the mutations are lethal. For the fitness contribution from this part, we have a zero for the sequences with zero mutations in this block and $-\infty$ for the nonzero mutations. Part three is described by a single peak fitness landscape with fitness $J = 3$ for the peak subsequence and zero for other subsequences. Thus, the fitness function is defined as $f(m_1,m_2,m_3) = km_1^2/2 - [1 - \delta(m_2 - 1)]\infty + J\delta(m3 - 1)$, where the discrete $\delta(x)$ function is equal to 1 at zero and is equal to 0 otherwise. The mean fitness is given as $k[1 - 1/(3k)]^2/2 + 3 - 2/3$.

Figure 1 gives the comparison of our analytical result for Eq. (9) with numerics of the three-dimensional (3D) model.

B. The multidimensional model with lethal mutations

Let us now consider a model where there exists some probabilities of lethal mutations: the fitness in parallel model is becoming $-\infty$ [30].

At any piece of the genome, we consider the master subsequence having nonlethal $L_n(1 - \lambda)$ neighbors with single mutations, where $0 \leqslant \lambda < 1$ is a parameter describing the fraction of lethal mutations. When the fitness is a function of the Hamming distance from the reference sequence, we simplify the evolution equations using this symmetry. We define some mutations from the reference sequence as lethal mutations and assume that any sequence having at least one lethal mutation (plus some nonlethal mutations) has a $-\infty$ fitness. Therefore, at the lth Hamming class we have

$$N_{l,\lambda_n} = \frac{L_n(1-\lambda_n)!}{[L_n(1-\lambda_n)-l]!l!}$$

viable l point mutants, and, as a maximal l, we take $L_n(1 - \lambda_n)$. For a small $l \ll L_n$, there is a dilution of the sequence space via a factor $(1 - \lambda_n)$, while the total number of viable sequences is

$$\sum_{l=0}^{L_n(1-\lambda_n)} N_{l,\lambda_n} = 2^{(1-\lambda_n)L_n}. \tag{11}$$

We define now the fitness function as

$$r_{l_1,\ldots,l_H} \equiv Lf(m_1,\ldots,m_H), \tag{12}$$

where, instead of Eq. (1), we now define

$$l_n = L_n\frac{1+m_n}{2}(1-\lambda_n). \tag{13}$$

Then the calculation, identical to those in [30], gives

$$R = \max_m \left[f(m_1,\ldots,m_H) - 1 + \sum_n \frac{L_n}{L}(1-\lambda_n)\sqrt{1-m_n^2}\right]. \tag{14}$$

C. The model in multipeak fitness landscape

We formulated the model by Eq. (3) for a rather general case. The multipeak model, considered in [18], could be derived as a particular case of our solution.

Let us choose $H = 2^{K-1}$ and consider K reference sequences with our s_i^n spins, $1 \leqslant i \leqslant L, 1 \leqslant n \leqslant H$. At any position i along the genome, we are looking at the alignment of spins in our K reference sequences. We have chosen the first configuration with all + spins and define the alignment of spin along the ith reference sequence at the nth piece of genome as $\alpha_{i,n}$. We group together the configurations $s_i^n = \alpha_{i,n}$ and $s_i^n = -\alpha_{i,n}$, where $\alpha_{i,n} = \pm 1$ and these two cases have a joint probability L_n/L. The magnetization of the ith sequence M_i is defined through our m_n as

$$M_i = \sum_{n=1}^{H} \frac{L_n}{L}\alpha_{ni}m_n. \tag{15}$$

We then take a fitness which is a function of our H reference sequences. Thus, we should find the maximum of

$$F(M_1,\ldots,M_K) - 1 + \sum_{1\leqslant n\leqslant H} \frac{L_n}{L}\sqrt{1-m_n^2} + \sum_i h_i\left[-M_i + \sum_{n=1}^{H}\frac{L_n}{L}\alpha_{ni}m_n\right], \tag{16}$$

where we introduced the auxiliary variables h_i. The maximum condition gives

$$h_i = \frac{\partial F(M_1,\ldots,M_K)}{\partial M_i}, \quad \sum_i h_i\alpha_{ni} = \frac{m_n}{\sqrt{1-m_n^2}}. \tag{17}$$

The last system of equations coincides with the one derived in [18] with the mapping

$$m_n = \frac{1}{1 + \left(\sum_{i=1}^{K}\alpha_{n,i}H_i\right)}, \tag{18}$$

where H_i are the fields, conjugate to the M_i in Eq. (10) of [18]. A single difference is that in [18] we defined L_n/L for 2^K situations [misprints in Eqs. (9) and (24) of [18], in which 2^K should be replaced by 2^{K-1}], instead of 2^{K-1} in the current article.

D. The two-dimensional case

1. The definition of the model

Let us consider the two-dimensional (2D) case. We have a system of equations:

$$\frac{dp(l_1,l_2,t)}{dt} = \left(r_{l_1,l_2} - L - LR\right)p(l_1,l_2,t) + \sum_{\beta=\pm1} L_1 x_\beta(l_1,L_1)p(l_1-\beta,l_2,t) + L_2 x_\beta(l_2,L_2)p(l_1,l_2-\beta,t),$$
$$R = \frac{1}{L}\sum_{0\leqslant l_n\leqslant L_n} p(l_1,l_2,t)r_{l_1,l_2}. \quad (19)$$

We have a HJE for this case:

$$\frac{\partial u(m_1,m_2)}{\partial t} = -H\left(m_1,m_2;\frac{\partial u}{\partial_{m_1}},\frac{\partial u}{\partial_{m_2}}\right),$$
$$-H(m_1,m_2;\hat{P}_1,\hat{P}_2) = f(m_1,m_2) - 1 + \sum_{1\leqslant n\leqslant 2}\frac{L_n}{L}\left(\frac{1+m_n}{2}e^{2\hat{P}_n} + \frac{1-m_n}{2}e^{-2\hat{P}_n}\right). \quad (20)$$

The mean fitness R is defined through the equations

$$R = f(m_1,m_2) - 1 + \sum_{1\leqslant n\leqslant 2}\frac{L_n}{L}\sqrt{1-m_n^2},$$
$$f_1'(m_1,m_2) = \frac{L_1}{L}\frac{m_1}{\sqrt{1-m_n^2}}, \quad f_2'(m_1,m_2) = \frac{L_2}{L}\frac{m_2}{\sqrt{1-m_n^2}}. \quad (21)$$

2. The two-block model with lethal mutations

In Fig. 2 we compare the analytical results with the numerics for the two-block model, where one part has the length $(L-n)$ with a lethal mutation (all the spin configurations of the block besides the one have $-\infty$ fitness), and the other block has the length n and a fitness $f(m_1) = km_1^2/2$. We obtain the

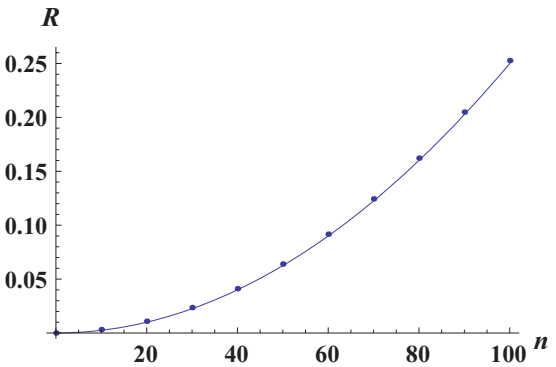

FIG. 2. (Color online) The mean fitness R versus the length of the first block in the 2D model with a fitness $f(m_1) = m_1^2$ for the first block with the length n and lethal mutations for the second block, with zero fitness for the peak configuration of the second block. The total length of the genome is 100. The analytical results are given by the smooth line.

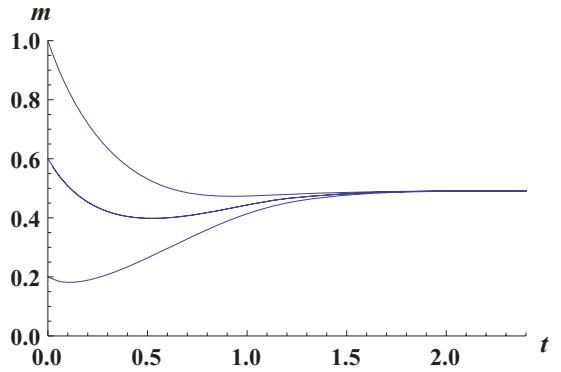

FIG. 3. (Color online) The dynamics of m,m_1,m_2 for the model by Eqs.(19),(24) with $m_1(0) = 1, m_2(0) = 0.2, L_1 = L/2, L_2 = L/2$. The middle line corresponds to the m by Eq. (24) or the Crow-Kimura model by Eq. (23) with $m(0) = 0.6$.

mean fitness of this model as

$$R = \frac{k}{2}\left(1 - \frac{n}{k(n+m)}\right)^2 - \frac{m}{n+m}. \quad (22)$$

3. The asymmetric original distribution

We consider the original distribution $m(0) = 0.6$ for the symmetric distribution, only considering the one-dimensional (1D) (Crow-Kimura) model:

$$f(m) = \frac{k}{2}m^2. \quad (23)$$

Later we take the simplest asymmetric distribution, where the part L_1 spins have l_1 minus spins and original narrow distribution with $m_1 = 1 - 2l_1/L_1$. Another part has l_2 minus spins and original narrow distribution around $m_2 = 1 - 2l_2/L_2$. We consider the model by Eq. (19) with the fitness

$$f(m_1,m_2) = \frac{k}{2}m^2, \quad m = \left(m_1\frac{L_1}{L} + m_2\frac{L_2}{L}\right). \quad (24)$$

Figure 3 gives the results of the dynamics for m,m_1,m_2.

4. The population distribution for the 2D case

Let us investigate the population distribution. We consider a fitness

$$f(m_1,m_2) = \frac{1}{2}\sum_{ij} A_{ij}m_i m_j,$$
$$A_{11} = k_1,\ A_{22} = k_2,\ A_{12} = k_3. \quad (25)$$

Assuming an ansatz

$$P(x) = \frac{\pi L\sqrt{\det(G)}}{2}\exp\left[-L\frac{\langle x|G|x\rangle}{2}\right]$$
$$u(x) = -\frac{\langle x|G|x\rangle}{2}, \quad \vec{x} = \vec{m} - \vec{s}, \quad (26)$$

we obtain for the correlation

$$K_{ij} \equiv \int dx\frac{p_i'(x)p_j'(x)}{p(x)} = \sum_{l,n} G_{li}G_{nj}\langle x_l x_n\rangle = G_{ij}. \quad (27)$$

TABLE I. Mean fitness for the 2D model by Eq. (25). $L_1 = L_2 = L/2, k_2 = k_1$.

L	100	100	100	100	100
k_1	4	8	4	4	3
K_3	3	2	5	6	7
R_{theor}	7.0312	9.025	8.0277	9.025	9.025
R_{num}	7.0315	9.0251	8.0280	9.025	9.0251
$\frac{g_1+g_3}{k_1+k_3}$ theor	1.	1	1	1	1
$\frac{g_1+g_3}{k_1+k_3}$ num	0.991	0.984	0.992	0.993	0.994

Differentiating the HJE (20) for the steady state by x_1, x_2, and putting $p_1 = 0, p_2 = 0$, we obtain

$$A_{ij}s_j - G_{ij}s_j. \tag{28}$$

For the symmetric fitness case

$$\begin{aligned} A_{11} = A_{22} = k_1, A_{12} = k_3, \\ G_{11} = G_{22} = g_1, G_{12} = g_3, \end{aligned} \tag{29}$$

$s_1 = s_2$ and Eq. (28) gives

$$k_1 + k_3 = g_1 + g_3. \tag{30}$$

We verified the validity of Eq. (30) by the numerics in Table I.

III. RECOMBINATION IN A MULTIDIMENSIONAL FITNESS LANDSCAPE

A. The model

In order to describe the recombination (horizontal gene transfer), we follow [22,23]. We consider the following system of equations:

$$\begin{aligned} &\frac{dp(l_1,\ldots,l_H,t)}{dt} \\ &= \left(r_{l_1,\ldots,l_H} - LR\right)p(l_1,\ldots,l_H) - Lp(l_1,\ldots,l_H,t) \\ &\quad + \sum_{\beta=\pm1,n} L_n x_\beta(l_n-\beta,L_n)p(l_1,\ldots,l_n-\beta,\ldots,l_H,t) \\ &\quad + c\left[\left(\sum_{\beta=\pm1,n} L_n x_\beta(l_n,L_n)\frac{1-\beta s_n}{2} - 1\right)p(l_1,\ldots,l_H,t)\right. \\ &\quad + \sum_{\beta=\pm1,n} L_n x_\beta(l_n-\beta,L_n)\frac{1+\beta s_n}{2} \\ &\quad \left.\times p(l_1,\ldots,l_n+\beta,\ldots,l_H,t)\right], \end{aligned} \tag{31}$$

where the sum over n extends from 1 to H, and

$$s_n = \sum_{l_1,\ldots,l_H} p(l_1,\ldots,l_H,t)\frac{L_n - 2l_n}{L_n} \tag{32}$$

is the equivalent of surplus or "surface" magnetization. The simple symmetric fitness landscape ($K = 1$) has one surplus parameter, but now there are H parameters.

The term $-Lp(l_l,\ldots,l_K,t)$ describes the mutations of the whole genome with a rate of 1 per allele; the following line describes the mutation. Using a coefficient c, we define the diagonal recombination terms: $-c$ is the total rate of changing the given sequence, and $x_\beta(l_n,L_n)\frac{1-\beta s_n}{2}$ describes the recombination event when we replace a spin from our current sequence with the same kind of spin from the pool of spins at the same position in the population. In the second term inside "$[\cdots]$," we define the recombination terms as the change in the current configuration: we replace a spin with an opposite spin from the spin pool.

Let us derive the Hamilton-Jacobi equation. We used the same ansatz, Eq. (5), as before; the simple derivations give

$$\begin{aligned} &\frac{\partial u}{\partial t} = H(m_1,\ldots,m_K;s_1,\ldots,s_H;u'_1,\ldots,u'_H), \\ &\quad -H = f(m_1,\ldots,m_H) - f(s_1,\ldots,s_H) - 1 - c \\ &\quad + \sum_{\beta=\pm1,1\leqslant n\leqslant H} \frac{l_n}{L}\left(\frac{1+m_n}{2}e^{2u'_n} + \frac{1-m_n}{2}e^{-2u'_n}\right) \\ &\quad + c\left(\sum_n \frac{l_n}{L}\left(\frac{(1+m_n)(1+s_n)}{4} + \frac{(1-m_n)(1-s_n)}{4}\right)\right. \\ &\quad - 1 + \sum_n \frac{l_n}{L}\left[\frac{(1+m_n)(1-s_n)}{4}e^{2u'_n}\right. \\ &\quad \left.\left. + \frac{(1-m_n)(1+s_n)}{4}e^{-2u'_n}\right], \end{aligned} \tag{33}$$

where we denote $u_n = \frac{\partial u(m_1,\ldots,m_H,t)}{\partial m_n}$. The function $u(m_1,\ldots,m_H,t)$ has the maximum at the point $(m_1,\ldots,m_H) = (s_1,\ldots,s_H)$.

We do not see a simple way to calculate the asymptotic solution of the last equation.

B. An approximate solution of recombination dynamics

Let us consider the dynamics of the initial normal distribution,

$$\begin{aligned} &P(m_1,\ldots,m_H,0) \\ &\quad = \exp\left\{-L\sum_{ln}\frac{y_{ln}}{2}[m_l - s_l(0)][m_n - s_n(0)]\right\}. \end{aligned} \tag{34}$$

Equation (34) describes a narrow distribution around some Hamming classes.

We assume that for some not too large periods of time, we have a similar solution,

$$\begin{aligned} &P(m_1,\ldots,m_H,t) \\ &\quad = \exp\left\{-L\sum_{ln}\frac{y_{ln}}{2}[m_l - s_l(t)][m_n - s_n(t)]\right\}, \end{aligned} \tag{35}$$

where y_{ln} describes the normal distribution.

We get the following system of equations for $ds_n(t)/dt$ using our Hamiltonian form Eq. (33):

$$\begin{aligned} -\sum_n y_{ln}\frac{ds_n}{dt} &= -\frac{dH(s_1,\ldots,s_H;s_1,\ldots,s_H;0,\ldots,0)}{dm_l} \\ &\quad + \sum_n \frac{dH(s_1,\ldots,s_H;s_1,\ldots,s_H;0,\ldots,0)}{dp_n}y_{ln}. \end{aligned} \tag{36}$$

Let us prove that the last two terms do not depend on c. From the first line we obtain

$$-\frac{dH(s_1,\ldots,s_H;s_1,\ldots,s_H;0,\ldots,0)}{dm_l}=\frac{\partial f(m_1,\ldots,m_H)}{\partial m_l}. \tag{37}$$

For the rest we derive

$$-2\sum_l \frac{L_l}{L}s_l y_{ln}. \tag{38}$$

Eventually, putting the results of Eqs. (37) and (38) into Eq. (36), we derive

$$\sum_n y_{ln}\frac{ds_n}{dt}=f_n'(s_1,\ldots,s_H)-2\sum_n \frac{L_l}{L}s_l y_{ln}. \tag{39}$$

Thus, for the initially narrow distribution of population by Eq. (34) and mean-field like fitness landscape, the recombination does not have any impact on the relaxation dynamics for some period of time T. If the number of mutations and recombination per genome per replication is on the order of 1, then we have the following condition for this time period: $1 \ll T \ll L$.

C. Asymmetric recombination

The theorem from the previous section is not valid for the asymmetric recombination, since we have different recombination rates for the allele changes to up and down. Consider the simple case of a one-dimensional fitness landscape:

$$\begin{aligned}\frac{dP_l}{dt}&=[(r_l-LR)]\,P_l+(l+1)P_{l+1}+[1-(l-1)]P_{l-1}\\&-L\left[c_1\left(1-\frac{\bar{l}}{L}\right)\frac{l}{L}P_l+c_2\frac{\bar{l}}{L}\left(1-\frac{l}{L}\right)P_l\right]\\&+L\left[c_1\left(1-\frac{\bar{l}}{L}\right)\frac{l+1}{L}P_{l+1}\right.\\&\left.+c_2\frac{\bar{l}}{L}\left(1-\frac{l-1}{L}\right)P_{l-1}\right],\end{aligned} \tag{40}$$

where c_1,c_2 describe the recombination rates to the up and down directions in Hamming classes and $\bar{l}=\sum_l P_l l$.

Using an ansatz $P_l=\exp[Lu(m,t)]$, we derive the following HJE:

$$\begin{aligned}\frac{du}{dt}&=f(m)-f(s)-c_1\frac{(1+m)(1-s)}{4}-c_2\frac{(1-m)(1+s)}{4}\\&+e^{2u'}\frac{1+m}{2}\left[1+\frac{1-s}{2}c_2\right]-1\\&+e^{-2u'}\frac{1-m}{2}\left[1+\frac{1+s}{2}c_1\right].\end{aligned} \tag{41}$$

Now we take $u(t)=-y[m-s(t)]^2/2$ and get an equation

$$y\frac{ds}{dt}=f'(s)-2ys(t)-\frac{(c_1-c_2)}{2}[1-s(t)^2]y. \tag{42}$$

We see that the recombination immediately starts to change the distribution; see Fig. 4 for the illustration.

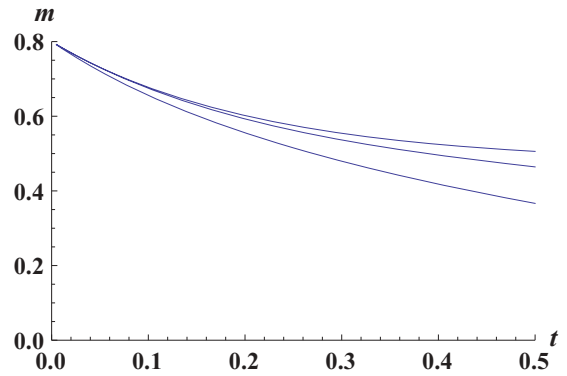

FIG. 4. (Color online) The dynamics of $m \equiv 1-2n/L$, where n is the mean number of mutations in the model by Eqs. (40) and (41) with $f(m)=m^2$, $L=1000$. The top line corresponds to the model without recombination, the middle line to the model with symmetric recombination with the rate $c=1$, and the bottom line to the asymmetric recombination with $c_1=1.5$, $c_2=0.5$. The time scale is chosen as in Eq. (41). For the zero selection case at time period 1 almost all the alleles in the genome are mutated.

D. The recombination model with lethal mutations

In order to describe the lethal mutations, we consider the genome which consists of two parts with the length $L_1=\lambda L$ and $L(1-\lambda)$. In the first piece, there is only one sequence with the fitness 0, and any mutation in this part gives a lethal sequence with the $-\infty$ fitness.

We can investigate the situation using our model by Eq. (40). Previously we used the mutation rate 1. Now we introduce the mutation rate μ_0 per nucleotide and c as a recombination rate per nucleotide.

We just write the equations for $p(0,l)\equiv p_l$, identifying also $r(0,l)\equiv r_l$:

$$\begin{aligned}\frac{dp_l}{dt}&=r_l p_l-p_l\mu_0 L\\&+\bar{L}\left[\mu_0\left(\frac{l-1}{\bar{L}}p_{l+1}+\frac{\bar{L}-l+1}{\bar{L}}p_{l-1}\right)\right.\\&+c\left(\frac{l-1}{\bar{L}}\frac{1+s_n}{2}+\frac{\bar{L}-l+1}{\bar{L}}\frac{1-s_n}{2}-1\right)p_l\\&\left.+c\left(\frac{l-1}{\bar{L}}\frac{1-s_n}{2}p_{l-1}+\frac{\bar{L}-l+1}{\bar{L}}\frac{1+s_n}{2}p_{l+1}\right)\right],\end{aligned} \tag{43}$$

where we denoted the length of the genome without lethal mutations as $\bar{L}=L(1-\lambda)$. While in the previous models we took $\mu_0=1$, now we write formulas for general μ_0.

Let us define

$$m=\frac{2l-\bar{L}}{\bar{L}},\quad r_l=f(m)\hat{L}, \tag{44}$$

then we can use the results of [23] to calculate the mean fitness. If we define the potential $U(m,s)$,

$$\begin{aligned}U(m,s)&=f(m)+\sqrt{(1-m^2)C}+\frac{cms}{2}-\frac{c}{2},\\C&=\left[\left(\mu_0+\frac{c}{2}\right)^2-\frac{c^2s^2}{4}\right],\end{aligned} \tag{45}$$

then the mean fitness of the genome is defined as

$$\max[\bar{L}(U(m,s) - L\mu_0)], \quad LR = Lf(s)(1-\lambda). \quad (46)$$

E. The finite population version of the model with lethal mutations

In the case of HIV, there are highly variable parts of the genome with about 100 nucleotides [35]. In [35] the use of an evolution model with shorter effective genome length to describe the virus evolution in such a case was suggested; later this idea was applied in [36]. We assume that the usage of an effective genome length is reasonable for the zero epistasis case, while in the case of lethal mutations as well, we cannot use an ordinary model with the short genome length.

Extending the ideas in [37], we suggest the following finite population versions of the model. The genome consists of two parts. The first part has a length $L\lambda$ where all the mutations are lethal, while the n mutations from the part with the length $L(1-\lambda)$ give a mutant with the fitness function r_n. The population is described via $L-\bar{n}$ viable sequences and the $\bar{n}$ lethal ones, and the total population size N is fixed. We describe the population via the number of viruses, n_l, in the lth Hamming class, $0 \leqslant l \leqslant L$ and $\bar{n}$. We have a conserved population size, $n + \sum_{l=0}^{L} n_l = N$.

During the time period δt, there are $\mu\delta t(1-\lambda)$ nonlethal mutations and $\mu\delta t\lambda$ lethal mutations.

We consider the following steps during the evolution:

(a) a birth of $\delta\bar{n}$ new lethal mutants which is a binomial random process with the probability parameter $\delta t\lambda$ and $(N-\bar{n})$ trials;

(b) a birth of δn_l new viruses in the lth class, which is a binomial random process with n_l trials and a probability parameter $r_l\delta t$;

(c) forward nonlethal mutations f_l, which are described via binomial random process with a probability parameter $\delta t(1-\lambda)\frac{l}{L}$ and n_l trials, and backward nonlethal mutations b_l, which are described via the binomial random process with probability parameter $\delta t(1-\lambda)\frac{L-l}{L}$ and n_l trials (thus, after these mutation processes, $n_l \to n_l - f_l - b_l$, $n_{l+1} = n_{l+1} + f_l$, $n_{l-1} = n_{l-1} + b_l$); and

(d) the dilution of the model, where we reduce the virus population via $\bar{n} + \sum_{l=0}^{L} \delta n_l$ numbers, uniformly distributed via $L+2$ classes.

IV. CONCLUSION

We formulated and solved the evolution model on the multidimensional fitness space, where we considered the genome as a collection of several pieces and the total fitness as the function of the allele type fractions of the pieces. Such a model is more general and more realistic than the multipoint fitness landscape considered in [18]. The numerics confirmed our analytical results well.

We calculated the mean fitness of this model, including the case of lethal mutations, and found a simple way of deriving the results of the multipeak fitness models.

We formulated the recombination model in the multidimensional fitness space. While we could not calculate the mean fitness, we derived the Hamilton-Jacobi equation for the dynamics of the population and deduced an important theorem about the dynamics. For the initially narrow initial distribution and mean-field fitness landscape, the recombination does not affect the dynamics of the population for a rather long period of time (see Fig. 2). This theorem is not valid in the case of asymmetric recombination.

We formulated the finite population version of the model with lethal mutations. Our results could be applied to model virus experiments, prescribing to different parts of the genome either lethal mutations or negative or positive selection. For example, we can apply our model in the case of the dengue virus, where 95% of the genome is epistasis free while there are strong correlations between the gene contributions of the remaining 5% [38].

The main open mathematical problem in the investigation of multidimensional evolution is the calculation of the surplus and the distribution around the peak of distribution. While we calculated the mean fitness, we failed to calculate the surplus. In classical mechanics, one can easily define the ground state energy and the position of the interacting particles, looking for the minimum of potential energy. Now, for our Hamiltonian by Eqs. (20), the situation is highly nontrivial. One should consider the asymptotic solution for the characteristics (the solutions of Hamilton equation), looking for the steady states. Another problem, which is important for applications, is to define the quadratic expansion of the solution $u(m_1,m_2)$ near the maximum of distribution. Again, the situation is highly nontrivial, and different statistical physics phases are possible like the phases in [39]. While we found some relations, Eqs. (28) and (30), we failed to find the complete solution of distribution. We hope that it is possible to succeed using the advanced methods of HJE to address this open problem.

ACKNOWLEDGMENTS

Several years ago, C. Biebricher criticized our former choice of fitness landscape [18]; he pointed out to us the current version of the fitness landscape as more adequate for virus evolution. We thank R. H. Griffey for the discussion of possible application of our model to analyze the evolution data of dengue virus. This work was supported by Academia Sinica, National Science Council in Taiwan, with Grant No. NSC 100-2112-M-001-003-MY2 and the National Center for Theoretical Sciences in Taiwan.

[1] M. Eigen, Naturwissenschaften **58**, 465 (1971).
[2] J. Swetina and P. Schuster, Biophys. Chem. **16**, 329 (1982).
[3] M. Eigen, J. S. McCaskill, and P. Schuster, Adv. Chem. Phys. **75**, 149 (1989).
[4] J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory* (Harper & Row, New York, 1970).
[5] L. Peliti, arXiv:cond-mat/9712027.
[6] E. Baake and W. Gabriel, Annu. Rev. Comput. Phys. **7**, 203 (2000).

[7] B. Drossel, Adv. Phys. **50**, 209 (2001).
[8] I. Leuthausser, J. Stat. Phys. **48**, 343 (1987).
[9] P. Tarazona, Phys. Rev. A **45**, 6038 (1992).
[10] H. Woodcock and P. G. Higgs, J. Theor. Biol. **179**, 61 (1996).
[11] S. Franz and L. Peliti, J. Phys. A **30**, 4481 (1997).
[12] E. Baake, M. Baake, and H. Wagner, Phys. Rev. Lett. **78**, 559 (1997).
[13] E. Baake and H. Wagner, Genet. Res. **78**, 93 (2001).
[14] D. Alves and J. F. Fontanari, Phys. Rev. E **57**, 7008 (1998).
[15] D. B. Saakian and C.-K. Hu, Phys. Rev. E **69**, 021913 (2004).
[16] D. B. Saakian and C.-K. Hu, Phys. Rev. E **69**, 046121 (2004).
[17] D. B. Saakian and C.-K. Hu, Proc. Natl. Acad. Sci. USA **103**, 4935 (2006).
[18] D. B. Saakian, E. Munoz, C.-K. Hu, and M. W. Deem, Phys. Rev. E **73**, 041913 (2006).
[19] D. B. Saakian, J. Stat. Phys. **128**, 781 (2007).
[20] K. Sato and K. Kaneko, Phys. Rev. E **75**, 061909 (2007).
[21] D. B. Saakian, Z. Kirakosyan, and C. K. Hu, Phys. Rev. E **77**, 061907 (2008).
[22] E. Cohen, D. A. Kessler, and H. Levine, Phys. Rev. Lett. **94**, 098102 (2005).
[23] J.-M. Park and M. W. Deem, Phys. Rev. Lett. **98**, 058101 (2007).
[24] A. S. Perelson and C. A. Macken, Proc. Natl. Acad. Sci. USA **92**, 9657 (1995).
[25] H. A. Orr, Evolution **60**, 1113 (2006).
[26] S. Seetharaman and K. Jain, Phys. Rev. E **82**, 031109 (2010).
[27] R. Burger, *The Mathematical Theory of Selection, Recombination, and Mutation* (Wiley, New York, 2000).
[28] W. J. Ewens, *Mathematical Population Genetics* (Springer-Verlag, New York, 2004).
[29] M. Baake and E. Baake, Can. J. Math. **55**, 3 (2003).
[30] Zh. Avetisyan and D. B. Saakian, Phys. Rev. E **81**, 051916 (2010).
[31] Z. Kirakosyan, D. B. Saakian, and C.-K. Hu, Phys. Rev. E **82**, 011904 (2010).
[32] D. B. Saakian, C.-K. Biebricher, and C.-K. Hu, PLoS One **6**, 21904 (2011).
[33] C. J. Thompson and J. L. McBride, Math. Biosci. **21**, 127 (1974).
[34] B. L. Jones, R. H. Enns, and R. S. Rangnekar, Bull. Math. Biol. **38**, 15 (1975).
[35] G. Bocharov, N. J. Ford, J. Edwards, T. Breinig, S. Wain-Hobson, and A. Meyerhans, J. Gen Virol. **86**, 3109 (2005).
[36] N. N. V. Vijay, V. R. Ajmani, A. S. Perelson, and N. M. Dixit, J. Gen. Virol. **89**, 1467 (2008).
[37] D. B. Saakian, M. W. Deem, and C.-K. Hu, EPL **98**, 18001 (2012).
[38] R. Griffey (unpublished).
[39] D. B. Saakian, C.-K. Biebricher, and C.-K. Hu, Phys. Rev. E **79**, 041905 (2009).