Journal of Statistical Physics, Vol. 128, No. 3, August 2007 (© 2007)
DOI: 10.1007/s10955-007-9334-9

A New Method for the Solution of Models of Biological Evolution: Derivation of Exact Steady-State Distributions

David B. Saakian[1]

Received May 6, 2006; accepted April 30, 2007
Published Online: May 22, 2007

We investigate well-known models of biological evolution and address the open problem of how construct a correct continuous analog of mutations in discrete sequence space. We deal with models where the fitness is a function of a Hamming distance from the reference sequence. The mutation-selection master equation in the discrete sequence space is replaced by a Hamilton-Jacobi equation for the logarithm of relative frequencies of different sequences. The steady-state distribution, mean fitness and the variance of fitness are derived. All our results are asymptotic in the large genome limit. A variety of important biological and biochemical models can be solved by this new approach.

KEY WORDS: evolution, extract solution, Eigen model, Hamilton-Jacobi equation

PACS numbers: 87.10.+e, 87.15.Aa, 87.23.Kg, 02.50.-r

1. INTRODUCTION

The theory of biological evolution is intimately connected with mathematical and statistical physics. Two principal ideas of interest that have been thus far studied are: analogies between mutation processes and diffusion processes,[(1)] and, parallelism between describing combined selection-diffusion processes and established methodologies of statistical quantum mechanics.[(2)] For Mendelian populations a diffusion model of evolution in the case of a few alleles has been studied as early as 1945 by Wright.[(3)] Kimura was the first to consider evolution in neutral-fitness landscapes as a diffusion process.[(4)] In recent studies[(5−8)] discrete sequence-spaces of biological evolution models have been generalized to give

[1] Yerevan Physics Institute, Alikhanian Brothers St. 2, Yerevan 375036, Armenia and Institute of Physics, Academia Sinica, Nankang, Taipei 11529, Taiwan; e-mail: saakian@yerphi.am

a better description of an evolutionary diffusion as a continuous process. It is generally accepted that the above extensions provide a good qualitative picture of evolutionary phenomena, however, it is also known that the current formalism does not provide the required high-accuracy mapping between a discrete configuration space and diffusion in a continuum space.[9] Results of several models[10−13] indicate the importance of a proper handling of the thermodynamic limit.[14]

In this paper we present a new alternative way to describe evolutionary models in the limit of an infinite genome. We use previously studied models[15−17] as test beds for our theory. Within the new formalism introduced here we shall derive expressions for population landscapes, exact relations for variances of steady-state distributions in the parallel mutation-selection scheme (Crow-Kimura) and the connected mutation-selection scheme (Eigen) models. Our approach is complementary to the known four different exact approaches that include the maximum principle for quadratic forms,[14,18] the Suzuki-Trotter method,[16,17] the high-temperature expansion method,[19] and the functional approach of Ref. 20. Our results are compared against similar results obtained in these four approaches

In a simple Crow-Kimura model[2,13] any genotype is specified by N values of two-valued spins $s_k = \pm 1$, $1 \leq k \leq N$. In our notation, a genotype is assigned to the i-th configuration $S_i \equiv (s_i^1, \ldots, s_i^N)$. The state is specified by 2^N relative frequencies P_i, $1 \leq i \leq 2^N$:

$$\frac{dP_i}{dt} = P_i\left(r_i - \sum_{j=1}^{2^N} r_j P_j\right) + \sum_{j=1}^{2^N} m_{ij} P_j. \tag{1}$$

where r_i is the fitness, m_{ij} is the mutation rate from configuration S_j to configuration S_i[2] and for the probability balance $\sum_i m_{ij} = 0$. Configurations S_i and S_j are separated by the Hamming distance $d_{ij} = (N - \sum_k s_i^k s_j^k)/2$. We have that $m_{ii} = -\gamma_0 N$; for $d_{ij} = 1$ $m_{ij} = \gamma_0$ and $m_{ij} = 0$ otherwise. We describe a fitness landscape by defining the fitness function as $f(S_i) = r_i$. The mean fitness rate R_p for Eq. (1) is given by

$$R_p = \sum_{i=1}^{2^N} P_i r_i \tag{2}$$

As a single index i, taking the values $1 \leq i \leq 2^N$ is equivalent to the collection of the N spins s_k, taking values ± 1, we define a function $f(S_i) \equiv r_i$. This is quite a formal definition that describes the fitness landscape by a fitness function.

It is assumed in Eq. (1) that mutations and selections act independently from each other. This is unlike the Eigen's model,[10,11] where these two processes are interconnected, which gives:

$$\frac{dP_i}{dt} = \sum_{j=1}^{2^N}[Q_{ij}r_j - \delta_{ij}D_j]P_j - P_i\left[\sum_{j=1}^{2^N}(r_j - D_j)P_j\right]. \tag{3}$$

where D_i is the degradation rate, and the elements Q_{ij} of mutation matrix give probabilities that an offspring of state S_j belongs to the state S_i. In this model mutations are quantified by $Q_{ij} = q^{N-d(i,j)}(1-q)^{d(i,j)}$ and $\gamma = N(1-q)$, where $\exp[-\gamma] \equiv q^N$ is the exact copying probability. For Eq. (3) the mean fitness rate R_c is defined as

$$R_c = \sum_{i=1}^{2^N} P_i(r_i - D_i) \tag{4}$$

In Eqs. (1), (3) $r_i \equiv f(S_i)$ describes the *fitness landscape*. The theoretical derivation of the *population landscape*, specified by the relative populations P_i, from the fitness landscape and the mutation terms is still a difficult problem.[12] We will calculate it later. Other important characteristics to be defined are *mean fitness* R_p in the model described by Eq. (1) and the *mean excess production* R_c for the model given by Eq. (3).

Solutions to the system of nonlinear equations describing both models, i.e., Eq. (1) as well as Eq. (3), can be obtained by means of transformations[21,22]

$$P_i(t) = \frac{\hat{P}_i(t)}{\sum_{j=1}^{2^N}\hat{P}_j(t)} \tag{5}$$

where distributions $\hat{P}_j$ are obtained by solving linear parts of the models:

$$d\hat{P}_i/dt = \sum A_{ij}\hat{P}_j, \tag{6}$$

and A_{ij} is a quadratic form of the linear part. The linear part of the dynamics can be obtained from the long-time asymptotic[11,14]

$$\hat{P}_i(t) \sim \exp(Rt), \tag{7}$$

where for the Crow-Kimura model $R \equiv R_p$, and for the Eigen model $R \equiv R_c$. Alternatively, the same result can be obtained when R is defined as a maximum over all distributions $\hat{P}_i$:

$$R = \max\left[\frac{\sum_{i,j} A_{ij}\hat{P}_i\hat{P}_j}{\sum_i \hat{P}_i^2}\right], \tag{8}$$

where the maximum is defined over all distributions $\hat{P}_i$.

Fitness functions for real biological systems are highly irregular. Theoretically they can be described[11,23] by considering the Hamming distance between an i-th configuration and a reference "peak" configuration. The peak configuration can be selected, e.g., as $S_1 \equiv 1, \ldots, 1$. Having this selection, the fitness depends only on $d_{1i} = N(1 - m^i)/2$, where the parameter $m^i \equiv \sum_{l=1}^{N} s_i^l / N$ is defined in analogy with magnetization. In this definition the fitness is a function of m^i[11,23]:

$$f(S_i) = N f_0(m^i), \tag{9}$$

where f_0 is a simple function of one variable, m is the magnetization of the configuration i. Based on the value $m^i \equiv m_l, m_l = 0, 1, 2, \ldots, N$, configurations are grouped into cosets. The multiplication N_l and the magnetization m_l of the lth equivalency class are, respectively,

$$N_l = \frac{N!}{l!(N-l)!}, \quad m_l = 1 - \frac{2l}{N} \tag{10}$$

The 0th equivalency class contains only S_1, and the fitness values $r_i = N f_0(m_l)$ depend only on the corresponding ith configuration. Assuming that the mutation rates m_{ij} in Eq. (1) are the same for any i, one needs only consider symmetric solutions of Eq. (1), where within the l-th equivalency class $P_i = p_l$. Each equivalency class is characterized by one value of fitness $r_i \equiv J_l$, i.e. $J_l \equiv N f_0(m_l)$.

An important characteristics of the model is the surplus which describes the degree of how configurations are grouped around the peak one

$$s = \frac{\sum_i P_i m^i}{\sum_i P_i} = \frac{\sum_l p_l N_l m_l}{\sum_l N_l p_l} \tag{11}$$

Equation (1) is transformed to Ref. 14

$$\frac{dp_l}{dt} = p_l \left(N f_0(m_l) - \gamma_0 N \right) + \left(l p_{l-1} + (N-l) p_{l+1} \right) \gamma_0 \tag{12}$$

The solution of Eq. (12) can be mapped to Eq. (1) via the transformation Eq. (5). As all the N_l configurations $\hat{P}_i$ from the same class have the probability of p_l, Eq. (8) is simplified to

$$R = \max \left[\frac{\sum_{i,j} A_{ij} p_l p_{l'}}{\sum_i \hat{N}_l p_l^2} \right] = \max \left[\frac{\sum_{l,l'} \hat{A}_{ll'} \hat{y}_l \hat{y}_{l'}}{\sum_i \hat{y}_l^2} \right] \tag{13}$$

where

$$y_l = \sqrt{N_l} p_l, \tag{14}$$

and $\hat{A}_{ll'} = \sqrt{\frac{N_l}{N_{l'}}} \sum_j A_{ij}$, the sum over j is restricted to the class l', i belongs to the class l (see Ref. 10). In the first equation of (14) the sum is restricted via configurations i from the class l and j from the class l'. If y_l has a sharp maximum at some l_0. The same is with y_l^2. Therefore one can look for the maximum of $\sqrt{N_l} p_l$

to calculate both the numerator and the dominator of Eq. (13) via the saddle point and then derive R. We will follow this idea later while deriving mean fitness.

In this article we are going to solve Crow-Kimura's and Eigen's models. Our solution is asymptotic in the large genome limit while we take into account accurately the backward mutations. In case of Eigen's and Crow-Kimura's models we calculate the leading finite N corrections. These results are important as there are biologically interesting situations with a restricted actual genome length.[24–26] Our analysis is restricted to the models where the fitness is a function of the Hamming distances from the reference sequence and the genome has two-value alphabet. Our equations can also be generalized for the four-value alphabet[27] when the fitness is symmetric. Some results are possible to derive for the simple two-peak fitness landscapes (Crow-Kimura model with quadratic fitness, for example) but we avoid these problems in this paper.

This paper is organized as follows. In Sec. 2 we derive the Hamilton-Jacobi equation Eq. (19) to solve exactly the Crow-Kimura model, derive exact steady-state distribution in the limit of large genome length, and give finite-size corrections to mean fitness and fitness variance. The exact (in the large genome length limit) steady state distribution is derived, Eqs. (17), (22). The finite size corrections to mean fitness are calculated for the general symmetric fitness Eqs. (38), (44) and the fitness variance, Eq. (28). Numerical results are presented in Tables I and II. In Sec. 3 the same problems are solved for the Eigen model, where it is important to solve the model without ignoring the background mutations.[11] Here, Eq. (64) gives the leading order correction to the mean single-pick fitness model for finite N. According to our formula Eq. (64) one has to take into account the finite N corrections for short genome lengths ($N = 10 - 20$),[24] or even $N = 100$,[25] in case of neutrality and high mutation rates ($\sim$1 per genome per replication) for realistic fitnesses $(A - 1) \approx 0.2$ to get correct results.

2. SOLUTION OF PARALLEL MODEL

2.1. Correct Scaling for p_l

At the limit $N \to \infty$ we consider a smooth solution for p_l and define a function

$$p(m, t) \equiv p_{(N-mN)/2}(t) \tag{15}$$

In Ref. 16 for $f_0(m) = 0$ and for the initial condition $p_N = 1$ we derived the following exact solution for $p(m, t)$:

$$p(m, t) = \mathrm{e}^{N\left(\frac{1+m}{2} \ln\cosh(\gamma_0 t) + \frac{1-m}{2} \ln\sinh(\gamma_0 t) - \gamma_0 t\right)} \tag{16}$$

Let us generalize the form of Eq. (16) and consider a solution of Eq. (12) for the nonzero J_l in the form:

$$p(m,t) = \exp[Nu(m,t)] \tag{17}$$

The Eq. (16) is the key point of the work. For another ansatz instead of $p_l \sim e^{N\cdots}$, different problems arose at $N \to \infty$.[5,8] Let us choose $\gamma_0 = 1$ and put the expression of $p(m,t)$ from the Eq. (17) into Eq. (12). If we define $m \equiv m_l$, then $m_{l-1} = m + 2/N$, $m_{l+1} = m - 2/N$. Using a simple expansion we have

$$p_{l\pm 1} = \exp[Nu(m,t) \pm 2u'(m,t)](1 + o(1)), \tag{18}$$

and Eq. (12) transforms into the Hamilton-Jacobi equation for $u(m,t)$:

$$\frac{\partial u(m,t)}{\partial t} = f_0(m) - 1 + \left[\frac{(1+m)}{2}\exp\left(-2\frac{\partial u(m,t)}{\partial m}\right) + \frac{(1-m)}{2}\exp\left(2\frac{\partial u(m,t)}{\partial m}\right)\right] \tag{19}$$

Equation (16) gives an exact solution of Eq. (19) at $f_0(m) = 0$.

2.2. Investigation of the Hamilton-Jacobi Equation

We assume an asymptotic $u(m,t) = u_0(m) + kt$, where $k \equiv R_p$, see Eq. (2). We already calculated the mean fitness R_p using Suzuki-Trotter method.[17] Here we derive it by an alternative method exploring Eq. (19).

For the suggested ansatz we have the following equation

$$k = f_0(m) - 1 + \left[\frac{(1+m)}{2}\exp\left(-2\frac{du_0(m,t)}{dm}\right) + \frac{(1-m)}{2}\exp\left(2\frac{du_0(m,t)}{dm}\right)\right] \tag{20}$$

We obtain two solutions

$$u_0'(m) = \frac{1}{2}\ln\frac{k+1-f_0(m) \pm \sqrt{(k+1-f_0(m))^2 - 1 + m^2}}{1-m} \tag{21}$$

We are interested in a monotonic function $f_0(m)$ and a single peak for the function $f(x) + \sqrt{1-x^2}$. For the region $-1 \le x < m_0$, where m_0 is the solution of the equation $(k+1-f_0(m))^2 - 1 + m^2 = 0$, see Eq. (26), we should choose the $+$ solution.

$$u_0(m) = \int_{-1}^{m}\frac{dx}{2}\ln\frac{k+1-f_0(x) + \sqrt{(k+1-f_0(x))^2 - 1 + x^2}}{1-x} \tag{22}$$

This region $[-1, m_0]$ is the most important one, as the majority of population is located at the point s in side that region, $0 < s < m_0$. s is the surplus, will be derived later. Let us prove our choice.

For the total (class) probabilities $N_l p_l = \exp[N(u_0(m) + h(m))] \equiv \exp[Nv(x)]$, where

$$h(m) = -(1+m)/2 \ln(1+m)/2 - (1-m)/2 \ln(1-m)/2,$$

we have an equation

$$k = f_0(m) - 1 + \left[\frac{(1+m)}{2} \exp\left(2\frac{dv(m)}{dm}\right) + \frac{(1-m)}{2} \exp\left(-2\frac{dv(m)}{dm}\right)\right] \tag{23}$$

Equation (23) has also $\pm$ solutions, corresponding to the $\pm$ ones in Eq. (21).

$$v'(m) = \frac{1}{2} \ln \frac{k+1-f_0(m) \pm \sqrt{(k+1-f_0(m))^2 - 1 + m^2}}{1+m}$$

We assume that $v(x)$ has one maximum point $m = x_0$, where $v'(x_0) = 0$, $v''(x_0) < 0$. The first condition gives from Eq. (23)

$$k = f_0(x_0) \tag{24}$$

Comparing with Eq. (11), we identify x_0 with the surplus s. The reference sequence has been chosen specially to have $x_0 > 0$.

At this point two solutions for v' give $v'_+(x_0) = \frac{1}{2} \ln \frac{1+\sqrt{x_0^2}}{1+x_0} = 0$, and $v'_-(x_0) = \frac{1}{2} \ln \frac{1-\sqrt{x_0^2}}{1+x_0} = \ln \frac{1-x_0}{1+x_0} < 0$. Therefore, only $+$ solution has a maximum. Thus at $x = x_0$ we should choose the $+$ solution. Assuming a continuous solution for Eqs. (19), (20), we should choose the $+$ solution in the whole region $[-1, m_0]$, where at the point m_0 two solutions $\pm$ for u'_0 coincide. To choose the correct solution $\pm$ in the region $[m_0, 1]$ we assume that $u''(x)$ is continuous at the point m_0, then we should choose the solution $-$ in the region $[m_0, 1]$. The situation is similar to the question, what choice among two alternatives: $f_\pm(x) = 1 + \pm\sqrt{|x^2|}$ is correct for $x > 0$ and $x < 0$. A simple analysis gives a smooth solution $1 + \sqrt{|x^2|}$ for $x > 0$ and $1 - \sqrt{|x^2|}$ for $x < 0$, or just $1 + x$. Our numerics for the fitnesses $f_0(x) = kx^2/2 + x$ and $f_0(x) = kx^2/2 + x^3$ confirmed such $(-)$ choice of the sign in the region $[m_0, 1]$. For the case of non-monotonic function the choice of the proper solution of Eq. (21) is a nontrivial task.

Our solution for Eq. (22) is correct in the interval $-1 < m < 1$ when

$$k \geq k_0, k_0 \equiv \max_{-1<m<1} \left[f(m) - 1 + \sqrt{1-m^2}\right], \tag{25}$$

otherwise $u_0(m)$ has an imaginary part. Denote the point of maximum as m_0.

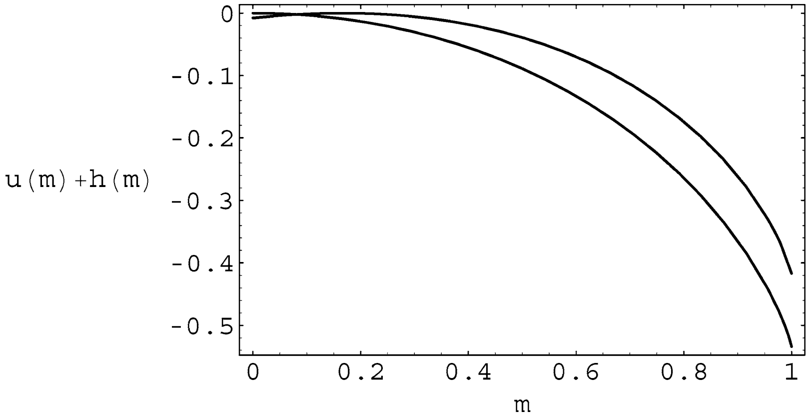

Fig. 1. The logarithm of a class probability ($\frac{1}{N}\ln(p_l\binom{N}{l}) \equiv u_0(m) + h(m)$) as a function of m for the fitness $f_0(m) = \mathrm{cm}^2/2$. At $c = 0.9$ the system is in the error catastrophe phase and the maximum is at $m = 0$ (down line). At $c = 1.2$ the maximum is at 1/6 (up line), where $s \equiv 1/6$ is the surplus, Eq. (11)

The low border of the inequality (25) just corresponds to the mean fitness value (equal to Nk_0, see Ref. 17). When the maximum in Eq. (25) is at $m = 0$, there is no selection. This is the error catastrophe phase.[10,11] In case of successful selection, the minimum in Eq. (25) is at some nonzero $m = m_0$:

$$k_0 = f_0(m_0) + \sqrt{1 - m_0^2} - 1, \quad f_0'(m_0) = \frac{m_0}{\sqrt{1 - m_0^2}} \tag{26}$$

Having the expression for k_0, we can define the surplus $s \equiv x_0$ from the Eq. (24) putting $k = k_0$. We can define the current value of m (m for the majority of population) as the maximum of the exponent $\exp[N(u(m,t) + h(m))]$.

We solved Eq. (19) numerically for the quadratic fitness $f_0(m) = \mathrm{cm}^2/2$ (see Fig. 1). The numerical solution of our equation is coherent with the theory. The critical value is $c = 1$.

2.3. The Fitness Variance

Let us derive an expression for the steady-state fitness variance. The class probability has a maximum at some $m = x_0$ and $k = f_0(x_0)$ according to Eq. (24). The variance is defined as

$$V = N^2 \sum_{l=0}^{N} N_l p_l (J_l - k_0)^2 = \frac{N^2 \int_{-1}^{1} dm (f_0(m) - k_0)^2 \exp N(h(m) + u_0(m))}{\int_1^1 dm \exp N(h(m) + u_0(m))} \tag{27}$$

To calculate the last expression via saddle point, we need in $h''(x_0) + u_0''(x_0)$. From the Eq. (21) we derive

$$h''(x_0) + u_0''(x_0) = -\frac{f_0'(x_0)}{2x_0}$$

Putting the last expression into Eq. (27) we obtain

$$V = \sqrt{\frac{Nf_0'(x_0)}{4x_0\pi}} \int_{-\infty}^{\infty} dx e^{-N\frac{f_0'(x_0)}{4x_0}x^2} N^2(f_0'(x_0)x)^2 = 2Nf_0'(x_0)x_0 \tag{28}$$

In Ref. 14 the following expression for the fitness variance has been derived for the homogenous fitness $f_0(lx) = f_0(x)l^n$

$$V = 2Nn\gamma_0 f(x_0) \tag{29}$$

which coincides with our Eq. (28) derived for the case of general fitness landscape (we have chosen $\gamma_0 = 1$).

Let us now consider the transformation Eq. (14) and define

$$\exp[NU(m,t)] \equiv \sqrt{N_l}\,p_l(t) \tag{30}$$

For this new variable the Eq. (19) transforms into

$$\frac{\partial U(m,t)}{\partial t} = f_0(m) - 1 + \sqrt{1-m^2}\cdot\cosh\left[2\frac{\partial U(m,t)}{\partial m}\right] \tag{31}$$

Considering the asymptotic solution $U(m,t) = U_0(m) + kt$ we derive

$$k = f_0(m) - 1 + \sqrt{1-m^2}\cdot\cosh\left[2\frac{dU_0(m)}{dm}\right] \tag{32}$$

Let us assume that $U_0(m)$ has a maximum at some $m = m_0$, where $U_0'(m_0) = 0$. We immediately derive $k = f_0(m_0) - 1 + \sqrt{1-m_0^2}$. Taking into account the inequality from Eq. (25), we conclude that $k = k_0$. The same result could be derived from the maximum principle of Eq. (13). Let us look at the second equation in (13). The configuration, giving the maximal value of R, should be either at the border of interval $[-1 \leq m \leq 1]$ or inside the interval. We assume the second situation. Calculating the value of R from Eq. (13) we have only $f_0(m_0) - 1 + \sqrt{1-m_0^2}$. We are looking for the maximal eigenvalue and derive the extremum value $k = k_0$ in Eq. (25).

2.4. High Order Corrections

We derived the principal term in the expansion of $y_l \equiv \sqrt{N_l}\,p_l$, see Eq. (30). Let us derive an expression for the next term in the $1/N$ expansion

$$\ln\sqrt{N_l}\,p_l = NU(m) + u_1(m) + o(1) \tag{33}$$

First consider the exact quadratic form instead of Eq. (32). We have the eigenvalue equation for $y_l \equiv \sqrt{N_l} p_l$

$$\begin{aligned}\lambda y_l &= y_l(Nf_0(m_l) - N) + \sqrt{l(N-l+1)} y_{l-1} + \sqrt{(N-l)(l+1)} y_{l+1}\\ &\approx y_l(Nf_0(m_l) - N) + \sqrt{l(N-l)}\left[1 + \frac{1}{2(N-l)}\right] y_{l-1}\\ &\quad + \sqrt{l(N-l)}\left[1 + \frac{1}{2l}\right] y_{l+1}\end{aligned} \tag{34}$$

Denote

$$\begin{aligned}L_{l,l'} &= (Nf_0(m_l) - N)\delta_{l,l'} + \sqrt{l(N-l)}[\delta_{l+1,l'} + \delta_{l-1,l'}]\\ L^b_{l,l'} &= \left[\frac{1}{2(N-l)}\delta_{l-1,l'} + \frac{1}{2l}\delta_{l+1,l'}\right]\sqrt{l(N-l)}\end{aligned} \tag{35}$$

With the second order accuracy ($o(1)$) Eq. (34) could be written as

$$\lambda y_l = \sum_{l'} [L_{ll'} + L^b_{ll'}] y_{l'} \tag{36}$$

Equation (32) gives the exact solution of the operator L_0, where $y_{l\pm 1}$ are replaced by $y_l \exp[-\pm 2U'_{0,m}]$. We are looking up for the maximal eigenvalue

$$(L_0 + L_a + L_b)(|\psi > + |\psi_1 >) = (Nk_0 + k_1)(|\psi > + |\psi_1 >), \tag{37}$$

where Nk_0 is the exact eigenvalue of the operator L^0, $\hat{L} = \hat{L}^0 + \hat{L}^a$. According to the formulas of quantum mechanics, the eigenvalue k', including the first order corrections (mean fitness expression with the finite size corrections), could be calculated as

$$k' = Nk_0 + k_1 \equiv Nk_0 + k_a + k_b = \frac{\langle\psi|L_0 + L_a + L_b|\psi\rangle}{\langle\psi|\psi\rangle}, \tag{38}$$

where $|\psi >$ is the eigenvector of the operator L_0 and $|\psi > + |\psi_1 >$ of the $(L_0 + L_a + L_b)$.

We obtain

$$k_b = \frac{\langle\psi|L_b|\psi\rangle}{\langle\psi \mid \psi\rangle} = \frac{1}{\sqrt{1 - m_0^2}} \tag{39}$$

Considering the higher order corrections for p_{l+1} in Eq. (34), we derive

$$\frac{\int_{-\infty}^{\infty} dm \exp[2NU_0(m)]\sqrt{1-m^2}\cosh(2U_0')\frac{2U_0''}{N}}{\int_{-\infty}^{\infty} dm \exp[2U_0(m)]} = \frac{2}{N}\sqrt{1-m_0^2}U_0''(m_0) \tag{40}$$

where

$$U_0''(m) = -\frac{f'(m) - \frac{m}{\sqrt{1-m^2}}}{2\sqrt{1-m^2}\,\sinh(2U_0')} \tag{41}$$

Near the maximum point

$$f(m) - 1 - k_0 + \sqrt{1-m^2} \approx -F(m-m_0)^2/2,$$
$$-F = f''(m_0) - \frac{1}{\left(1-m_0^2\right)^{3/2}} \tag{42}$$

Then we immediately derive

$$U_0''(m_0) = -\frac{1}{2}\left(1-m_0^2\right)^{-1/4}\sqrt{F} \tag{43}$$

Therefore, for the finite size corrections we have:

$$k_1 = \frac{1}{\sqrt{1-m_0^2}}\left[1 - \sqrt{1 - f''(m_0)\left(1-m_0^2\right)^{3/2}}\right] \tag{44}$$

In Table I we compare the results for the mean fitness $R_{\mathrm{num}} \equiv N k_{\mathrm{num}}$ derived from the numerical solution of Eq. (1) with the theoretical formula $k_0 N$, Eq. (26), and with the high accuracy formula $k_0 N + k_1$, see Eqs. (38), (44). One can check that our formula gives the exact value of mean fitness at the large genome limit

$$R_k - k_0 N - k_1 \sim O\left(\frac{1}{N}\right) \tag{45}$$

Table I. Comparison of First Order Accuracy Expression for the Mean Fitness k_{theor}, Eq. (26) and the Second Order Accuracy Expression $k_0 + k_1/N$, Eq. (44), with the Numerical Result for the Mean Fitness, Derived from Eq. (1) for the Fitness Function $f_0(m) = \frac{c}{2}m^2$, Where $\delta_1 \equiv k_{\mathrm{num}} - k_0$, $\delta_2 \equiv k_{\mathrm{num}} - k_0 - k_1/N$

N	100	100	100	150	150	150
c	1.2	1.5	2	1.2	1.5	2.
k_{num}	0.02226	0.08712	0.25716	0.02033	0.08591	0.25180
δ_1	0.00560	0.00389	0.00271	0.00367	0.00257	0.00180
δ_2	0.00022	0.00007	0.00003	0.00008	0.00003	0.00001

Let us consider the high order corrections to U_0, see Eq. (33). Using the formulas

$$y_{l\pm 1} \approx y_l \exp[2U_0''/N - \pm(2U_0' + 2u_1'/N)], \lambda = Nk_0 + k_1,$$

we derive from Eq. (34)

$$\begin{aligned} &2u_1' \sinh(2U_0')\sqrt{1-m^2} \\ &\quad = k_1 - 2U_0'' \cosh(2U_0')\sqrt{1-m^2} - \frac{\cosh(2U_0') - m\sinh(2U_0')}{\sqrt{1-m^2}} \end{aligned} \tag{46}$$

We have an equation for the u_1

$$\begin{aligned} u_1 = \int_{m_0}^{m} \frac{dM}{2(1-M^2)\sinh(2U_0')}\Bigg[\sqrt{1-M^2}k_1 \\ + \left|\frac{f' - M/\sqrt{1-M^2}}{\sinh(2U_0'(M))}\right| \cosh(2U_0'(M))\sqrt{1-M^2} \\ - \cosh(2U_0'(M)) + M\sinh(2U_0'(M))\Bigg] \end{aligned} \tag{47}$$

Let us define P_i/P_{i_0}, where P_i has a maximum at $i = i_0$

$$P_i = P_{i_0}\sqrt{\frac{N_{i_0}}{N_i}} \exp[NU_0(m) + u_1(m)],$$

$$U_0(m) = \int_{m_0}^{m} \frac{dm}{2} \ln \frac{(k_0 + 1 - f) \pm \sqrt{(k_0+1-f)^2 - 1 + m^2}}{\sqrt{1-m^2}} \tag{48}$$

We take the $+$ solution for $m < m_0$ and the $-$ one for $m > m_0$. Our Eq. (48) gives the values of P_i with a relative accuracy $O(1/N)$.

2.5. Finite N Corrections in the Single Peak Fitness Crow-Kimura Model

Let us consider the fitness $J_0 = cN$ and $J_l = 0, l \geq 1$. We have calculated the mean fitness and the p_l in Refs. 16, 17. For the mean fitness we obtain $k_0 = c - 1$, $p_0 = \frac{c-1}{c}$, $p_1 = p_0/(cN)$. We can calculate the first order corrections using the Eq. (38). We should calculate the accurate expression of the mean fitness $k = \frac{\sum_{i,j=1}^{2N} A_{ij}P_iP_j}{\sum_l N_l p_l^2}$. We have for the dominator $\sum_l N_l p_l^2 \approx p_0^2 + Np_1^2 = p_0^2(1 + \frac{1}{c^2N})$. Using the expressions $\sum_j A_{0,j}P_j \approx Np_0(c-1) + Np_1 = p_0N(c-1)[1 + \frac{1}{c(c-1)N}]$ and $\sum_j A_{1,j}P_j \approx -Np_1 + p_0 = p_0\frac{c-1}{cN}$. Collecting all the three terms we derive for the mean fitness per spin

$$(c-1)+\frac{1}{cN} \tag{49}$$

3. EIGEN MODEL

3.1. Solution for the Smooth Symmetric Landscapes

Instead of Eq. (11) we now have

$$\frac{dp_l}{dt}=\sum_{l'} p_{l'} J_{l'} \sum_n \hat{Q}^n_{ll'} - d_0(m_l)p_l, \tag{50}$$

where

$$\hat{Q}^n_{ll'}=\sum_{j,d_{ij}=n} Q_{ij},$$

configuration i belongs to the class l and j to the class l'. Now we denote $D_i = d_0(m_l)$ and $J_l = f_0(m)$. The calculations are similar to those in the parallel case. The only difference is that we should consider multiple spin flips. In the configuration of the class l there are l negative spins. To take into account n spin flips we should consider all mutation schemes $n = n_1 + n_2$, where class l' is derived from the class l after n_1 up and n_2 down flips. We have the following expression for the neighbors number with n_1 up and n_2 down flips

$$N(l,n_1,n_2)=\frac{l!}{n_1!(l-n_1)!}\frac{(N-l)!}{n_2!(N-l-n_2)!}. \tag{51}$$

For the principal terms in Eq. (50) we have $n \ll N$. Therefore we can simplify Eqs. (50), (51) using the expressions $\frac{l!}{n_1!(l-n_1)!} \to l^{n_1}/n_1!, \frac{(N-l)!}{n_2!(N-l-n_2)!} \to (N-l)^{n_2}/n_2!$ and $J_{l'} = f_0(m)$. We have that $l = N(1-m)/2$, $N-l = N(1+m)/2$ and for the mutation event with n_1 down and n_2 up spin flips $l' = l-(n_1-n_2)$. We again take the ansatz (17) and derive

$$p_{l'} = \exp(Nu(m))[\exp(2(n_1-n_2)u') + o(1)] \tag{52}$$

For n_1+n_2 spin flips we have

$$Q_{ij} = e^{-\gamma}\left(\frac{\gamma}{N}\right)^{n_1+n_2} \tag{53}$$

Multiplying the last three equations we obtain

$$\sum_j Q_{ij} J_{l'} p_{l'} = f_0(m) \sum_{l'} p_{l'} e^{-\gamma} \left(\frac{\gamma}{N}\right)^{n_1+n_2} N(l, n_1, n_2)$$
$$= \sum_{n_1,n_2} \frac{f_0(m) p_l}{n_1! n_2!} e^{-\gamma} \left[\gamma \frac{(1-m)e^{2u'}}{2}\right]^{n_1} \left[\gamma \frac{(1+m)e^{-2u'}}{2}\right]^{n_2} \quad (54)$$

where i belongs to the class l and j to the class l'. Taking the sum over $0 \le n_1 \le \infty$, $0 \le n_2 \le \infty$ we derive the main equation for this case

$$\frac{\partial u(m,t)}{\partial t} = f_0(m) e^{-\gamma} \exp\left\{\gamma \left[\cosh\left(2\frac{\partial u(m,t)}{\partial m}\right) - m \sinh\left(2\frac{\partial u(m,t)}{\partial m}\right)\right]\right\} - d_0(m) \quad (55)$$

Let us consider again an asymptotic $u(m,t) = kt/N + u_0(m)$. We have an equation similar to Eq. (21)

$$u_0'(m) = \frac{1}{2} \ln \frac{a(m) \pm \sqrt{a(m)^2 - 1 + m^2}}{1-m}$$
$$a(m) = 1 + \frac{1}{\gamma} \ln \frac{k + d(m)}{f_0(m)} \quad (56)$$

and a solution for u_0

$$u_0(m) = \frac{1}{2} \int_{-1}^{m} \ln \frac{a(x) \pm \sqrt{a(x)^2 - 1 + x^2}}{1-x} dx \quad (57)$$

Equation (57) has a real solution at

$$k \ge k_0$$
$$k_0 = \max_{-1 \le m \le 1} \left[f_0(m) e^{-\gamma(1-\sqrt{1-m^2})} - d_0(m)\right] \quad (58)$$

We obtained the value of mean fitness k_0 that was derived in Ref. 19 by an alternative method.

Let us consider the maximum of the class probabilities $\exp[Nu(m,t) + Nh(m)]$ to define the surplus. Here we choose the $\pm$ solutions in Eq. (57) as in case of parallel model. Calculating u_0' from Eq. (58), we derive an equation for the surplus $s \equiv x_0$ from the saddle point condition $h' + u_0' = 0$

$$f_0(x_0) - d_0(x_0) = k_0 \quad (59)$$

Taking the derivative of Eq. (56), we have

$$h''(x_0) + u''(x_0) = -\frac{f_0(x_0)' - d_0(x_0)'}{2f_0(x_0)x_0\gamma} \tag{60}$$

Similar to Eqs. (27)–(29), we derive for the mean fitness variance

$$V = \sqrt{N\frac{f_0(x_0)' - d_0(x_0)'}{4\pi f_0(x_0)x_0\gamma}\int_{-\infty}^{\infty} dx e^{-\frac{N(f_0(x_0)' - d_0(x_0)')}{4f_0(x_0)x_0\gamma}x^2}}$$

$$(f_0'(x_0) - d_0'(x_0))^2 x^2 = 2\frac{x_0\gamma}{N} f_0[f_0'(x_0) - d_0'(x_0)] \tag{61}$$

Equation (61) gives the expression of the fitness variance. It is an important characteristic of evolving systems and has chances to be defined from the experimental data.

For $\sqrt{N_l}p_l = \exp[NU(m,t)]$ Eq. (55) transforms to

$$\frac{\partial U(m,t)}{\partial t} = f_0(m)e^{-\gamma}\exp\left\{\gamma\left[\sqrt{1-m^2}\cosh\left(2\frac{\partial U(m,t)}{\partial m}\right)\right]\right\} - d_0(m) \tag{62}$$

Considering the equation for U_0 in the asymptotic regime $U(m,t) = U_0 + kt/N$ we derive

$$k = f_0(m)e^{n-\gamma}\exp\left\{\gamma\sqrt{1-m^2}\cosh\left(2\frac{dU_0(m)}{dm}\right)\right\} - d_0(m) \tag{63}$$

Assuming that $U_0(m)$ has a maximum inside the interval $[-1, 1]$ we get the equality condition in the Eq. (58).

3.2. Finite Size Corrections in the Single Peak Fitness Eigen Model

In Refs. 19, 28 the bulk distribution of the model has been derived. For the fitness landscape $J_0 = A$ and $J_l = 1, l > 0$ one has for the mean fitness QA, $p_0 = \frac{(QA-1)}{A-1}, p_1 = \frac{\gamma A}{(A-1)N}p_0$ for the distributions. Besides the three terms, considered in the Sec. 2.4, we have an additional correction term $A(1-\gamma/N)^N = AQ(1-\frac{\gamma^2}{2N})$ where $Q \equiv \exp(-\gamma)$. Repeating the derivation of the Sec. 2.5 we obtain $p_0^2 + Np_1^2 = p_0^2(1 + \frac{\gamma^2 A^2}{(A-1)^2 N})$.

Using the expressions $\sum_j A_{0,j}P_j \approx p_0QA + NAQ\gamma p_1/N = p_0QA[1 + \frac{\gamma^2 A}{(A-1)N}]$, $Np_1\sum_j A_{1,j}P_j \approx Qp_0^2\frac{\gamma^2 A}{N(A-1)}(\frac{A}{A-1}+1)$, we derive for the mean fitness

$$QA + \frac{QA\gamma^2}{N}\left[\frac{1}{(A-1)} - \frac{1}{2}\right] \tag{64}$$

4. DISCUSSION

We have addressed the known open problem,[9] namely, how to construt correct continuous analog for mutations in discrete sequence space and constructed a new exact method for investigation of models of biological evolution. We suggested an ansatz (17) and carefully took the large genome length limit for evolution equations. The infinite system of quasispecies Eqs. (1), (3) is mapped to the Hamilton-Jacobi equations (see Refs. 19, 55) and to similar Eqs. (31), (62). When initial distribution of P_i is symmetric (P_i for the same Hamming distance l from the peak (reference) configuration are equal to p_l) our equation describes the dynamic of $u \equiv \ln p_l/N$. These new equations are a special case of the Hamilton-Jacobi equations when the spatial derivatives are present in the equation only in the exponents like $\exp[\pm u'x]$. This is the key point of our success in the derivation of the asymptotic solutions and the exact mean fitness. To derive the asymptotic of these equations we looked up only the minimum of the equation regarding to the spatial derivatives u'_m. This simple qualitative argument gives exact asymptotic in case of the Crow-Kimura and Eigen models.

We have derived exact steady-state distributions for the Crow-Kimura model, Eqs. (17), (22) and finite size corrections for the mean fitness Eqs. (44), (48). For the Eigen model we have calculated the principal terms for the probabilities in the long genome length limit, Eqs. (56), (57). Equations (49), (64) give the leading order correction to the mean single-pick fitness model for finite N. We have given exact expressions for the fitness variance in the steady state Eqs. (29), (61). For the case of Crow-Kimura model our analytical results have been confirmed by the numerical solution of Eq. (19) (see Fig. 1). Our theoretical results have been well confirmed by numerical calculations of the mean fitness, see $O(1/N^2)$ accuracy for δ_2, Table I.

Let us compare different methods for the continuous time models of molecular evolution. As all the methods are exact, they give identical results while using different tools.

A. The maximum principle method of Refs. 14, 18 is especially useful in the case of four value spins (it is easier to apply that instead of the Suzuki-Trotter method) and Eigen model (Suzuki-Trotter method could not be applied in a simple way). It is difficult to apply for the case of multi-peak fitness landscapes or for the finite genome size corrections. I do not see any way to obtain exact results for the case.
B. The Suzuki-Trotter approach[16,17] is the simplest method in case of two value spins. It is the best to solve the case of multi-peak fitness.
C. The high temperature expansion method[19] works for the case of Eigen model and the first exact solution of Eigen model has been derived by means of this

method. The method could be applied for the case of multi-peak fitness to give the mean fitness.

D. The functional integral method[20] gives the solution of Eigen model for the multi-peak fitness case including the finite genome length corrections. It is useful for the case of disorder.
E. The exact dynamics method for the single-peak fitness. [15,16] It could be generalized for the case of hierarchic, Random Energy Model like fitness landscapes, as well as for the nonlinear (diploid like) evolution.
F. The Hamilton-Jacobi equation (HJE) method is especially efficient in case of two-value spins. It gives finite genome size corrections to the mean fitness, exact steady state, variance as well as the dynamics for both Crow-Kimura and Eigen models. The HJE method could be applied for the genome growth model,[29,30] gene regulation model,[26] nonlinear evolution models (like the diploid evolution). Another interesting application are evolution games,[31] where HJE method could give higher accuracy than the diffusion method,[32] applied in Ref. 31.

I hope that these new results could be useful for the virus research where the fitness variance has a direct biological meaning. The accurate exact expressions for mean fitness are important as the fitness differences for the virus mutants are sometimes very small.

ACKNOWLEDGMENTS

"I thank A. Allahverdyan, E. Baake, C. Biebricher, M. Deem, V. Priezhev, O. Rozanova for useful discussions, and Z. Kirakosyan for checking the formulas. The work has been supported by CRDF grant ARP2-2647-Ye-05 and U.S. Defense Advanced Research Projects Agency DARPA \#HR00110510057, by the National Science Council of the Republic of China (Taiwan) under Grant No. NSC 95-2112-M-001-008, National Center for Theoretical Sciences in Taiwan, and Academia Sinica in Taiwan under Grant No. AS-95-TP-A07."

REFERENCES

1. S. Wright, *Proceedings of the sixth International Congress on Genetics* **1**:356 (1932).
2. E. Baake, M. Baake and H. Wagner, *Phys. Rev. Lett.* **78**:559 (1997).
3. S. Wright, *Proc. Natl. Acad. Sci. USA* **31**:382 (1945).
4. M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983).
5. L.S. Tsimring, H. Levin and D. A. Kessler, *Phys. Rev. Lett.* **76**:4440 (1996).
6. F. Bagnoli and M. Bezzi, *Phys. Rev. Lett.* **79**:3302 (1997).
7. U. Gerland and T. Hwa, *J. Mol. Evol.* **55**:386 (2002).
8. W. Peng, U. Gerland, T. Hwa and H. Levin, *Phys. Rev. Lett.* **90**:088103 (2003).

9. N. M. Shnerb, Y. Louzoun, E. Bettelheim and S. Solomon, *Proc. Natl. Acad. Sci. USA* **97**:10322 (2000).
10. M. Eigen, *Naturwissenschaften* **58**:465 (1971).
11. M. Eigen, J. McCaskill and P. Schuster, *Adv. Chem. Phys.* **75**:149 (1989).
12. M. Eigen and C. Biebricher, *Virus Research* **107**:117 (2005).
13. J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory* (Harper Row, NY, 1970).
14. E. Baake and H. Wagner, *Genet. Res.* **78**:93 (2001).
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, H. Khachatryan and C.-K. Hu, *Phys. Rev. E* **70**:041908 (2004).
18. E. Baake, M. Baake and A. Bover, *J. Math. Biol.* **50**:83 (2005).
19. D. B. Saakian and C.-K. Hu, *Proc. Natl. Acad. Sci. USA* **109**:4935 (2006).
20. D. B. Saakian, E. Munoz, C.-K. Hu and M. W. Deem, *Phys. Rev. E* **73**:041913 (2006).
21. C. J. Thompson and J. L. McBride, *Mathematical Biosciences* **21**:127 (1974).
22. B. L. Jones, R. H. Enns, and R. S. Rangnekar, *Bull. Math. Biol.* **38**:15 (1975).
23. J. Swetina and P. Schuster,*Biophys. Chem.* **16**:329 (1982).
24. D. M. Weinreich, N. F. Delaney, M. A. DePristo and D. L. Hartl, *Science* **312**:111 (2006).
25. Kun, A., Santos, M., and Szathmáry, E., *Nature Genetics*, **37**:1008–1011 (2005)
26. J. Berg, S. Willmann and M. Lassig, *BMC Evolutionary Biology* **4**:42 (2004).
27. J. Hermisson, M. Baake and H. Wagner, *J. Stat. Phys.* **102**:315 (2001).
28. B. Drossel, *Advances in Physics* **50**:209 (2001).
29. W. Li, *Phys. Rev. A* **43**:5240 (1991).
30. P. W. Messer, P. F. Arndt and M. Lassig, *Phys. Rev. Lett.* **94**:138103 (2005).
31. M. Lassig, L. Peliti and F. Tria, *Europhys. Lett.* **62**:446 (2003)
32. N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (North Holland, Amstersdam, NY, 2001).