Evolution models with base substitutions, insertions, deletions, and selection

D. B. Saakian*
Institute of Physics, Academia Sinica, Nankang, Taipei 11529, Taiwan;
Yerevan Physics Institute, Alikhanyan Brothers Street 2, Yerevan 36, Armenia;
and National Center for Theoretical Sciences: Physics Division, National Taiwan University, Taipei 10617, Taiwan
(Received 10 May 2008; revised manuscript received 11 September 2008; published 22 December 2008)

The evolution model with parallel mutation-selection scheme is solved for the case when selection is accompanied by base substitutions, insertions, and deletions. The fitness is assumed to be either a single-peak function (i.e., having one finite discontinuity) or a smooth function of the Hamming distance from the reference sequence. The mean fitness is calculated exactly in large-genome limit. In the case of insertions and deletions the evolution characteristics depend on the choice of reference sequence.

DOI: 10.1103/PhysRevE.78.061920 PACS number(s): 87.23.Kg, 64.60.De

I. INTRODUCTION

The existence of insertions and deletions (indels) is well established experimentally [1]. There has been considerable interest recently in molecular evolution eigenmodels, i.e., in the connected mutation-selection scheme [2,3], and in the parallel, i.e., "decoupled" mutation schemes [4–8]. The studies included mean fitness for different fitness landscapes [5–8] and population distributions under mutation-selection balance constraint [9]. In Refs. [10,11] has been investigated sexual evolution models with unequal recombination as a mechanism for generating indels. There have been several studies of molecular evolution models that incorporate base substitutions, insertions, and deletions [11–14]. In this article we integrate a concept of indels with parallel mutation-selection processes to solve our asexual evolution model with general fitness landscape and derive an exact formula for the mean fitness.

In biology research the term indel stands for either insertion alone or deletion alone or both these processes present simultaneously. Indels play an important role in phylogenic analysis in practical population genetics [15], where incorrect handling of indels may give unrealistic outcomes.

In the parallel mutation-selection model any genotype configuration i is specified as a sequence of N two-valued letters (alleles) $s_n=\pm 1$, $1\leqslant n\leqslant N$. We denote such configuration i by $S_i\equiv(s_1^i,\ldots,s_N^i)$. The probability p_i that configuration S_i occurs in genome, $1\leqslant i\leqslant 2^N$, satisfies

$$\frac{dp_i}{dt}=p_i\left(r_i-\sum_{j=1}^{2^N}r_jp_j\right)+\sum_{j=1}^{2^N}\mu_{ij}p_j, \tag{1}$$

where r_i is the fitness, and μ_{ij} is the mutation rate from S_i to S_j per unit time. For the Crow-Kimura model [4]: $\mu_{ij}=-aN$ if the Hamming distance d_{ij} is zero, $\mu_{ij}=a$ if $d_{ij}=1$, and $\mu_{ij}=0$ if $d_{ij}>1$, where $d_{ij}=(N-\Sigma_n s_n^i s_n^j)/2$.

In the models studied here we consider the following three independent parallel processes in the genome: base substitutions, deletions, and insertions. Assuming constant genome-variation rates per site, we denote a/N_0, b/N_0, and c/N_0 the rates of mutation, insertion, and deletion, respectively, where $N_0\gg 1$ is the scale length of the genome. Unlike in the well-studied cases of the parallel mutation-selection scheme and the eigenmodel, now the genome length can be varied. In this paper we focus only on the symmetric fitness landscape, i.e., when the fitness of the genome is a function of Hamming distance from a reference genome sequence. The fitness is assumed to be either a single-peak function (i.e., having one finite discontinuity) or a smooth function of the Hamming distance.

In the first model, analyzed in Sec. II, we are investigating indels acting in a toy problem when the reference sequence is ordered, i.e., when it contains only one letter (either +1 or −1) at all positions. Obtaining the solution to this toy problem is by no means trivial because neither the maximum principle [6] nor the Hamilton-Jacobi method [9] can be applied directly. A more realistic case is analyzed in Sec. III for a random reference sequence when the letters +1 and −1 are randomly distributed along the genome length. For symmetric fitness in parallel mutation-selection models without indels the choice of the reference sequence does not affect the solution, which is a consequence of the existing symmetry of the governing equations. The introduction of indels to the model breaks this symmetry and the effect of indels acting on sequence space is to change the solution. This change depends on the choice of the reference sequence. If we choose as the reference sequence the one with all + alleles, the result of deletion is the same for all the N possible position of deleted allele. In case a random reference sequence we have different results (sequences) after different positions of deleted allele. In our model, an individual indel event means either insertion or deletion of a single letter in the genome sequence, one at a time, but there may be many indel events during evolution. In this article we focus on investigating a "successful selection" phase, i.e., the phase (range of parameters) with the majority of populations being localized around the reference sequence. Our results are discussed in Sec. IV.

II. ORDERED REFERENCE SEQUENCE

We choose the reference sequence that has all the alleles +1 and the initial distribution of sequences that is symmetric

*saakian@yerphi.am

under permutations. An individual configuration is denoted by (N,L), where N is genome length and L is the number of $(+1)$ alleles in the configuration. The fitness is $N_0 r(N,L)$. Considering only one-letter deletion or insertion at a time there may be three processes that start at (N',L') and end at (N,L).

Simple base substitutions at the rate of a/N_0. At the beginning there are L configurations $(N,L-1)$ and $N-L$ configurations with $(N,L+1)$.

Deletions at the rate of b/N_0. At the beginning there are $N+1$ configurations $(N+1,L+1)$ and $N+1$ configurations with $(N+1,L)$.

Insertion at the rate of c/N_0. At the beginning there are L configurations $(N-1,L-1)$ and $(N-L)$ configurations $(N-1,L)$.

During base substitutions the letters (alleles) change their signs. During deletion one of the letters disappears. During insertion a new letter (either +1 or −1) is added randomly at any of the $(N+1)$ positions along the chain.

There are two interests in solving this model, usually treated separately. One interest concerns genome growth [10–12]. The other interest is the study of successive selection phase, which we present here for the case when selections and base substitutions are accompanied by deletions and insertions.

The occurrence probability $p(N,L)$ denotes a fractional number of configurations (N,L) in the population. For symmetric fitness landscape and permutation-symmetric initial distribution, probabilities $p(N,L)$ satisfy the equations [16]

$$\begin{aligned}\frac{dp(N,L)}{dt} &= p(N,L)\left(N_0 r(N,L) - \frac{N}{N_0}(a+b) - c\frac{N+1}{N_0}\right) \\ &+ a\left(\frac{L}{N_0}p(N,L-1) + p(N,L+1)\frac{N-L}{N_0}\right) \\ &+ b[p(N+1,L+1) + p(N+1,L)]\frac{N+1}{N_0} \\ &+ \frac{c}{2}\left(p(N-1,L-1)\frac{L}{N_0} + p(N-1,L)\frac{N-L}{N_0}\right) \\ &- p(N,L)\sum_{N',L'} r(N',L')p(N',L')\binom{N'}{L'}. \end{aligned} \tag{2}$$

In the case of symmetric fitness landscape, for finding steady-state mean fitness it is sufficient to consider only symmetric evolution. We solved Eq. (2) numerically, varying genome length between N_1 and N_2 subject to $N_2 - N_1 \gg 1$. The numerical results for two values of genome length are presented in Fig. 1.

Equation (2) is slightly modified near the border values of N, as at N_2 there are only deletions and at N_1 only insertions. The weighted sum over all equations is zero, where the weights $\binom{N}{L}$ are numbers of configurations with the same N and L.

For single-peak fitness landscape we set all fitness values to zero but one:

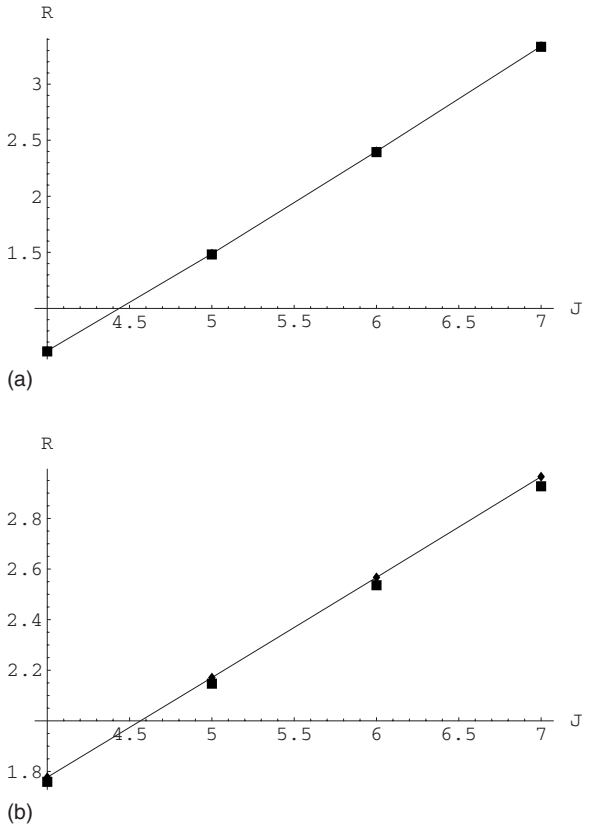

FIG. 1. Mean fitness R vs fitness parameter J for $a=b=1$ and $c=2$. Theoretical results are plotted as continuous lines. Numerical results are represented by symbols. Error bars give percent difference between numerical and theoretical results. (a) Single-peak mean fitness for $N=1000$. Error bars are about 0.05%, smaller than symbol size. (b) Quadratic fitness, $r=Jm^2$, for $N=200$. Error bars are about 0.5%.

$$r(N_0,N_0) = J, \quad r = 0 \text{ otherwise} \tag{3}$$

and $N_0 = \frac{N_1+N_2}{2}$. In the continuous-time model, considered in this work, fitness landscape (3) can be rescaled by an additive constant, which is a standard procedure in statistical physics. In discrete-time models all fitness values would have to be positive. For base substitutions acting alone without indels, with the choice given by Eq. (3) the system of equations (2) decouples, which leads to a single equation for only one master-type (reference sequence) probability [6] from which other probabilities are obtained recursively. When base substitutions and indels are simultaneously acting, for the single-peak fitness defined by Eq. (3) the system (2) does not decouple but, nonetheless, can be reduced to a tractable problem that can be treated analytically. The reduction procedure is outlined in the next paragraphs.

We consider the following scaling of Eq. (2):

$$p(N,N) \sim 1,$$

$$p(N,L) \sim 1/N_0^{N-L}. \tag{4}$$

Since the scaling (4) suppresses contributions from all terms (N,L) for $L<N$ as $1/N$, after the scaling we obtain a com-

plete set of equations for the class $p(N,N)$, $N_1 \leqslant N \leqslant N_2$ of frequencies. For example, for $p(N,N-1)$ we derive from Eq. (2)

$$p(N,N-1) = \frac{a}{N_0} p(N,N) + \frac{c}{2N_0} p(N-1,N-1), \tag{5}$$

which is easily verified to be consistent with the scaling ansatz (4).

Denoting by $\vec{P}$ a collection of all $p(N,N)$, where $P_l = p(N_1-1+l, N_1-1+l)$, we write Eq. (2) for $p(N,N)$ as

$$\frac{d\vec{P}}{dt} = \hat{A}\vec{P} - R\vec{P},$$

where R is the mean fitness, $R = Jp(N_0,N_0)$, and the elements of matrix $\hat{A}$ are

$$A_{ll} = -(a+b+c) + J\delta_{l,l_0},$$

$$A_{l,l+1} = b,$$

$$A_{l,l-1} = c/2, \tag{6}$$

where $l_0 = N_0 - N_1 + 1$ and δ_{l,l_0} is Kronecker's symbol. In a successful selection phase the majority of population is distributed around the configuration (N_0,N_0). The details of matrix $\hat{A}$ near the borders at $l=1$, $l=M+1$, and $M \equiv N_2 - N_1$ are irrelevant for the computation of the mean fitness in the successful-selection phase in the sense that the mean fitness is insensitive to variations in these border values. The mean fitness R is obtained in the standard way as the largest eigenvalue of $\hat{A}$ by solving the secular equation

$$\det(\hat{A} - \lambda\hat{I}) = 0, \tag{7}$$

where $\hat{I}$ is the identity matrix, and $R = \max(\lambda)$.

To calculate R within the $1/N_0$-accuracy we utilize the properties of determinant and in Eq. (7), modify the matrix A, taking $A_{1,M+1} = b$ and $A_{M+1,1} = c/2$. Then, we define an auxiliary function $g(J)$ by $g(J) \equiv \det(\hat{A} - R\hat{I})$. Since $g(J)$ is linear we write

$$g(J) = g(0) + Jg'(0),$$

where $g(0) = \det(\hat{B} - R\hat{I})$, the matrix $\hat{B}$ is the value of the matrix $\hat{A}$ computed at $J=0$, and $g'(0)$ is the first derivative of $g(J)$ computed at $J=0$. Because matrix $\hat{B}$ is symmetric and cyclic it is relatively straightforward to write $g(0)$ and $g'(0)$ explicitly:

$$g(0) = \prod_{l=0}^{M} \left(be^{i2\pi l/M} + \frac{c}{2} e^{-i2\pi l/M} - (a+b+c) - R \right),$$

$$\frac{g'(0)}{g(0)} = \frac{1}{M} \sum_{l=0}^{M} \frac{1}{be^{i2\pi l/M} + \frac{c}{2} e^{-i2\pi l/M} - (a+b+c) - R}. \tag{8}$$

In the thermodynamic limit of large N_0, N_1, and N_2, the infinite summation on the right-hand side of Eq. (8) becomes a contour integral in complex plane. The left-hand side of Eq. (8) is $g'(0)/g(0) = [g(J)/g(0) - 1]/J$ and $g(J) = 0$ because of Eq. (7). Thus, making the substitution $z = \exp(i2\pi l/M)$, Eq. (8) gives the relation between the mean fitness R and the fitness J of the peak configuration

$$1 = -\frac{J}{2\pi i} \oint \frac{dz}{z} \frac{1}{bz + \frac{c}{2z} - (a+b+c) - R}$$

$$= \frac{J}{\sqrt{(R+a+b+c)^2 - 2bc}}. \tag{9}$$

Inverting Eq. (9) gives the mean fitness R and fractional population P_m of the peak configuration

$$R = \sqrt{J^2 + 2bc} - (a+b+c),$$

$$P_m \equiv p(N_0,N_0) = \frac{\sqrt{J^2 + 2bc} - (a+b+c)}{J} \tag{10}$$

and the error-threshold condition

$$J \geqslant \sqrt{(a+b+c)^2 - 2bc}. \tag{11}$$

The results of Eqs. (3) and (10) are illustrated in Fig. 1(a).

A. Nonzero fitness at one N value and many L values

The sharp-peak fitness defined by Eq. (3) is an oversimplification as it is believed that realistic fitness landscapes are highly complicated and irregular. As a step towards generalization we now consider fitness that is nonzero at only one N value, set to $N=N_0$, and at many values of L. Here, L is the number of the (+1)-alleles in the genome and Hamming distance to the reference configuration is $N-L$. For this more general fitness we take

$$r(N,L) = \delta_{N,N_0} f(2L/N_0 - 1), \tag{12}$$

where $f(\cdots)$ is a smooth function. Following a method introduced by Baake and Wagner [6] we transform Eq. (1) to a more convenient form with the use of the substitution

$$y(N,L) = p(N,L) \sqrt{\frac{N!}{L!(N-L)!}}. \tag{13}$$

Equations for the weighted fractional populations $y(N,L)$ simplify in the large-genome limit. For the computation of the mean fitness they are easier to handle than the original Eq. (1). We have checked rigorously by calculating the distribution $y(N,L)$ that it is a smooth function of L/N_0 for the given $N=N_0$, although it is not smooth for all N. Assuming that $y(N,L)$ is a smooth function of $(2L/N-1)$ near $N=N_0$ and near the location $L_0(N_0)$ of its maximum, we replace $y(N,L)$ with $y[N,L_0(N_0)]$ in the coupled system of equations for $y(N,L)$ that was obtained from Eq. (1) after applying

transformation (13). As described for single-peak fitness, this gives a partial decoupling. For the decoupled part we have the eigenvalue problem

$$\hat{A}\vec{y} = \lambda\vec{y},$$

where $y_l = y[N, L_0(N_0)]$, $l = N - N_1 + 1$, and $R = \max(\lambda)$. Again, the matrix $\hat{A}$ is tridiagonal. In the limit of large N the eigenproblem for $\hat{A}$ gives

$$\lambda y_l = y_l[\delta_{l,l_0} f(m) - (a+b+c) + a\sqrt{1-m^2}] + \left(y_{l+1} b + y_{l-1}\frac{c}{2}\right)\frac{\sqrt{1+m}+\sqrt{1-m}}{\sqrt{2}}, \quad (14)$$

where $m = (N - 2L_0)/N_0$, $l_0 = N_0 - N_1 + 1$. There is full analogy between this problem and the problem already solved for the single-peak function. The quadratic form for the single-peak problem can be obtained from the quadratic form for the current problem by performing the mapping $b \to b\frac{\sqrt{1-m}+\sqrt{1+m}}{\sqrt{2}}$, $c \to c\frac{\sqrt{1+m}+\sqrt{1-m}}{\sqrt{2}}$, $R \to R - a\sqrt{1-m^2}$. By repeating the steps that lead to Eq. (9) we derive

$$R = \max_m\{-(a+b+c) + a\sqrt{1-m^2} + \sqrt{f^2(m) + bc(\sqrt{1+m}+\sqrt{1-m})^2}\}. \quad (15)$$

Theoretical results of Eqs. (12) and (15) for quadratic fitness are presented in Fig. 1(b).

B. General fitness landscape

For the ordered reference sequence we assume now nonzero fitness at many N and L values, i.e., fitness is a function of both genome length and the number of (+1) alleles:

$$r(N,L) = f\left(\frac{N}{N_0}, \frac{2L-N}{N_0}\right). \quad (16)$$

We assume that fitness $f(n,m)$ is a smooth function of both its arguments, i.e., it is also smooth in genome length. This is in contrast to the model of Sec. II A where the fitness may change drastically even within one unit of genome length. As in the previous two examples, to find the mean fitness R it is requested in our approach that distribution $y(n,m)$ must have a maximum localized at N and $L(N)$. Then, if the maximum exists the system of equations Eq. (1) for fractional populations can be partially decoupled at configuration $[N, L(N)]$. This leads to the algebra problem of finding the largest eigenvalue of a matrix $R = \max(\lambda)$. In analogy with Eq. (14), the intermediary result is

$$\lambda y_l = y_l[\delta_{l,l_0} f(n,m) - n(a+b+c) + a\sqrt{n^2-m^2}] + (y_{l+1} + y_{l-1})(\sqrt{n+m}+\sqrt{n-m})\frac{\sqrt{bc}}{2}. \quad (17)$$

Finally, the mean fitness R is the largest eigenvalue λ that is obtained by solving Eq. (17):

$$R = \max_{n \geqslant m}\{f(n,m) - n(a+b+c) + a\sqrt{n^2-m^2} + (\sqrt{n+m}+\sqrt{n-m})\sqrt{bc}\}. \quad (18)$$

Note, the final result (18) requires finding the maximum in two-dimensional space of arguments n and m.

III. RANDOM REFERENCE SEQUENCE

In this model a reference sequence contains both +1 and −1 alleles that are randomly distributed along the entire genome length. Now the evolution model is described via P_{Nj}, where N is the genome length and index j specifies the concrete sequence of length N, $1 \leqslant j \leqslant 2^N$. In the steady state we have an eigenvalue equation for the mean fitness

$$M_{Nj,N'j'} P_{N'j'} - R P_{Nj}, \quad (19)$$

where $N_1 \leqslant N \leqslant N_2$, $1 \leqslant j \leqslant 2^N$ and

$$M_{Nj,N'j'} = r_{Nj}\delta_{N,N'}\delta_{j,j'} + A_{Nj,Nj'}\delta_{N,N'} + B_{Nj,N'j} + C_{Nj,N'j'}. \quad (20)$$

In the last equation matrix $\hat{A}$ describes the base substitutions, $\hat{B}$-describes the insertions, and $\hat{C}$-describes the deletions. We have for the diagonal part $\hat{D}$ of the matrix $\hat{M}$

$$D_{Nj,N'j'} = \left[r_{Nj} - (a+b+c)\frac{N}{N_0}\right]\delta_{Nj,N'j'}. \quad (21)$$

In the case of a single peak fitness landscape we have a fitness $r_{N_0 j_0} = J$ for the reference sequence $N_0 j_0$ and 0 fitness for other sequences. We calculate the mean fitness R from the condition $\det\{\hat{M} - R\hat{I}\} = 0$, $\hat{I}$ is the identity matrix, or

$$\mathrm{Tr}\ln(\hat{M} - R\hat{I}) \to -\infty. \quad (22)$$

We rewrite the last equation as

$$\mathrm{Tr}\ln(\hat{D} - z\hat{I})\left[\hat{I} - \frac{\hat{I}}{\hat{D} - z\hat{I}}(\hat{A}' + \hat{B}' + \hat{C}')\right] \to -\infty, \quad (23)$$

where $\hat{A}, \hat{B}, \hat{C}$ are nondiagonal parts of corresponding operators. We can calculate Eq. (23) expanding via $\frac{\hat{I}}{\hat{D}-z\hat{I}}$. Thus we have terms $\sim\langle Nj_0|\hat{B}'\hat{C}|Nj_0\rangle$. In the case of an ordered reference sequence,

$$\langle N_0 j_0|\hat{B}'\hat{C}'|N_0 j_0\rangle \sim 1. \quad (24)$$

For the random reference sequence,

$$\langle N_0 j_0|\hat{B}'\hat{C}'|N_0 j_0\rangle \sim \frac{1}{N_0}. \quad (25)$$

For the random reference sequence we can neglect the nondiagonal parts of operators $\hat{B}, \hat{C}$ and Eq. (22) is equivalent to a similar equation in the single peak fitness case without indels and with substitution $R \to R - (b+c)$. Thus the mean fitness of the random reference sequence is exactly the same as though there were only base substitutions with the rate $a+b+c$:

$$R = J - (a + b + c), \tag{26}$$

and the error threshold condition is

$$J > (a + b + c). \tag{27}$$

For the continuous-time eigenmodel, where $r_1 = A$ and $r_i = 1$, $i > 1$, the presence of indels modifies error threshold to

$$QA \geqslant 1, \tag{28}$$

where Q is the probability of errorless reproduction of the entire genome.

Consider the general symmetric fitness landscape. In the previous sections we defined the fitness via the Hamming distance from the reference sequence. Now for any genome length N we specify some reference sequence and the fitness of sequence with the genome length N is defined via Hamming distance d from the corresponding reference sequence with the length N:

$$r(N,d) = f\left(\frac{N}{N_0}, \frac{N-2d}{N_0}\right), \tag{29}$$

where $f(\cdots,\cdots)$ is assumed to be a smooth function of two parameters. By the method outlined in Sec. II we obtain the following general result for the mean fitness:

$$R = \max_{n>m}\{f(n,m) - n(a+b+c) + a\sqrt{n^2 - m^2}\}. \tag{30}$$

IV. DISCUSSION

The toy model with ordered reference sequence and single-peak fitness, studied in Sec. II, is an interesting case from a methodological point of view. It describes the evolution when the fitness is defined by genome length N and the number L of (+1)-alleles. Genotype frequencies for this class are smooth functions of L in the neighborhood of the peak distribution. Error-threshold condition (11) depends on all the rates in a nonlinear fashion because of the existence of reversal processes introduced by indels, which processes may drive the evolution towards the reference sequence. This picture is unlike to what we learn from Crow-Kimura (parallel) model without indels, where error threshold depends linearly on the rates. In the generalized model with single-peak fitness, studied in Sec. III, where the reference sequence is a random sequence of +1 and −1 alleles, the error-threshold formula (27), simplifies again to that for the Crow-Kimura model with an efficient base substitution rate.

A general symmetric-fitness model with a random reference sequence, analyzed in Sec. III, presents the most realistic situation where both genome length and reference sequence are allowed to vary. In this work we investigated only steady-state characteristics of this model.

In our computational approach we used standard methods of linear algebra to partially decouple a system of evolutionary equations around the peak distribution and find the mean fitness as the largest eigenvalue of the decoupled subsystem. As the excellent agreement between our analytical and numerical solutions demonstrates (see Fig. 1) our methodology has the promise of becoming a routine approach in solving general evolutionary problems with varying genome length, alongside the methods of quantum mechanics [5] for the fixed genome length, and the methods of quantum field theory [17,18] for the changing genome length.

Haploid models with indels that were studied in this work can be extended to diploid evolution models with parallel insertions and deletions. Similar complex models were already considered to study the evolution of gene families via conversion processes [11] and gene crossover processes [10]. The latter mentioned mechanisms could be, in principle, handled analytically by modern methods [19] that proved to be successful in treating diploid evolution [20,21].

In evolution research the models often ignore either selection processes [11–13] or indels [2–7], however, it is generally accepted that the concurrent selection and indels play an important role in biology. The models that are capable to describe these two processes as acting simultaneously could give a connection with the phylogeny analysis and with the investigation of gene families. Our introductory study of this work shows that it is possible to analytically derive some class of results when selection is accompanied by indels.

In summary, in this work we introduced a method that allows one to investigate a broad class of evolution models. We solved a parallel mutation-selection model with general symmetric-fitness landscapes in the case of simultaneously acting base substitutions, insertions, and deletions. Our findings indicate that in the steady state of this evolution model the mean fitness depends strongly on the choice of the reference sequence.

ACKNOWLEDGMENTS

I thank J. Felenstein, A. Kolakowska, and H. C. Lee for useful discussions. This work was supported by the Volkswagenstiftung grant "Quantum Thermodynamics," by the National Science Council of Taiwan under Grant No. NSC 96-2911-M 001-003-MY3, by National Center for Theoretical Sciences in Taiwan, and by Academia Sinica (Taiwan) under Grant No. AS-95-TP-A07.

[1] W. H. Li and D. Graur, *Fundamentals of Molecular Evolution* (Sinnaier Associates, Sunderland, 1991).
[2] M. Eigen, Naturwiss. **58**, 465 (1971).
[3] M. Eigen, J. 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] E. Baake, M. Baake, and H. Wagner, Phys. Rev. Lett. **78**, 559 (1997).
[6] E. Baake and H. Wagner, Genet. Res. **78**, 93 (2001).

[7] D. B. Saakian, C.-K. Hu, and H. Khachatryan, Phys. Rev. E **70**, 041908 (2004).
[8] D. B. Saakian and C.-K. Hu, Proc. Natl. Acad. Sci. U.S.A. **103**, 4935 (2006).
[9] D. B. Saakian, J. Stat. Phys. **128**, 781 (2007).
[10] T. Nagylaki, Proc. Natl. Acad. Sci. U.S.A. **81**, 3796 (1984).
[11] T. Ohta, *Lecture Notes in Biomathematics* (Springer Verlag, Berlin 1980), Vol. 37.
[12] W. Li, Phys. Rev. A **43**, 5240 (1991).
[13] P. W. Messer, P. F. Arndt, and M. Lassig, Phys. Rev. Lett. **94**, 138103 (2005).
[14] L. C. Hsieh, L. Luo, F. Ji, and H. C. Lee, Phys. Rev. Lett. **90**, 018101 (2003).
[15] J. Cohen, Science **316**, 1836 (2007).
[16] J. Swetina and P. Schuster, Biophys. Chem. **16**, 329 (1982).
[17] M. Sasai and P. G. Wolynes, Proc. Natl. Acad. Sci. U.S.A. **100**, 2374 (2003).
[18] A. F. Ramos and J. E. M. Hornos, Phys. Rev. Lett. **99**, 108103 (2007).
[19] D. B. Saakian, Z. Kirakosyan, and C.-K. Hu, Phys. Rev. E, **77**, 061907 (2008).
[20] N. H. Barton, Evolution **46**, 551 (1992).
[21] M. Shpak and A. S. Kondrashov, Evolution **53**, 600 (1999).