1. Introduction

In recent decades, ideas and methods of physics have been widely used to study biological problems at the molecular level, e.g., the structure and folding of proteins,1–4) structure and thermal properties of DNA and RNA.5,6,7 molecular models of biological evolution and the origin of life,8–21) etc. In this paper, we will address an interesting problem related to the origin of life.

In 1971, Eigen proposed an autocatalytic reaction model for the origin of life.10,11) The model describes the self-replication of the molecules having some $L$ monomers (letters). There is a mutation probability $1-q$ for each letter, which defines the probability of the letter being changed into another letter in the set of alphabets. If the replication rate of a master or main sequence is $A$ times higher than the replication rate (fitness) of other sequences, then the correct self-replication is possible when the mutation probability is less than a threshold value. Otherwise, the sequence with the high replication rate will be lost.

Given the typical experimental mutation probability for RNA molecules,12,14) the maximal allowed length is about 100–1000, while the length of the genome to posses a minimal set of genes15) was estimated to be about 8000–20000.16) This is the famous error catastrophe paradox in the origin of life. There have been many attempts to solve the paradox (see refs. 16 and 21, and references there in). More recent works include Rajamani et al.18) who proposed to solve the paradox through slowing replication processes due to mutations and our work tackling the problem using a Hamilton–Jacobi equation method to solve the model and calculate the mean fitness. We find that the error threshold depends on the correlation, and the suggested mechanism may give a simple solution to the error catastrophe paradox in the origin of life.

The error threshold phenomenon in the case of high mutation rates, considering the general probability distribution instead of the Poisson distribution for multiple mutations. In the latter case, multiple mutations are independent, while in the former case, multiple mutations are correlated. We applied the Hamilton–Jacobi equation method23–25) to solve the model and calculated the mean fitness. The error threshold could be changed due to considered correlations. The suggested mechanism can give a simple solution to the error catastrophe paradox in the origin of life without introducing the lethal mutation and a truncated fitness landscape.20,21)

In the rest of this section, we briefly review the Eigen model10,11) with random multiple mutations and its analytic solution. We also define related notations used in this paper.

In the Eigen model,10,11) there are some types of monomers. For the sake of simplicity, we consider two types of monomers, $\pm 1$ as in references.10,11) There are $M = 2^L$ different types (sequences) of molecules, $0 \leq i \leq M - 1$, and we choose the 0-th (peak) sequence with all +1 monomers. We denote $S_i = s_1^i, \ldots, s_L^i$ as the $i$-th genome, where $s_j^i$ describes the type of the monomer at the $j$-th position. We denote the probability for the appearance of $S_i$ at time $t$ by $p_{S_i} \equiv p_i(t)$. The $i$-th sequence has a fitness $r_i$.

The mutation matrix element $Q_{ij}$ is the probability that an offspring produced by state $j$ changes to state $i$, and the evolution is given by the set of equations for $M$ probabilities $p_i$

$$\frac{dp_i}{dt} = \sum_{j=0}^{M-1} (Q_{ij}) r_j p_j - p_i \left( \sum_{j=0}^{M-1} r_j p_j \right).$$

Here $p_i$ satisfies the constraint $\sum_{i=0}^{M-1} p_i = 1$, and

$$Q_{ij} = q^{d(i,j)} (1 - q)^{d(i,j)},$$

with $q$ being the mean nucleotide incorporation fidelity, and $d(i,j) = (L - \sum_{j=1}^{L} s_j^i s_j^j)/2$ being the Hamming distance between $S_i$ and $S_j$.10,11) The parameter $\gamma = L (1-q)$ describes the mutation efficiency. The second term in eq. (1) ensures the balance condition $\sum_{i} p_i = 1$. It describes the dilution of the population in the chemical reactor.

We define the Hamming distance $d(i,0)$ of the sequence $S_i$ from the peak sequence $S_0$ through the parameter $m$.

E-mail: huck@phys.sinica.edu.tw
\[ d(i, 0) = L \frac{1 - m}{2}. \] (3)

We assume that \( r_i \) is a function of the Hamming distance \( d(0, i) \) or \( m \)
\[ r_i \equiv f(m), \] (4)
where \( m \) is given by eq. (3). We have chosen the indices \( i \) of the first \( L + 1 \) sequences to have \( d(i, 0) = i, \) \( 0 \leq i \leq L. \)

On the basis of the connection between Eigen model and a quantum spin model,\(^{26}\) the mean growth rate \( R = \sum_r r \rho_t \) for the whole population has been calculated in ref. 27:
\[ R = \max \{e^{-\gamma(1-\sqrt{1-k^2})}f(k)\}_{-1 \leq k \leq 1}. \] (5)

The maximum is at some \(-1 \leq k_0 \leq 1.\)

The population is focused on the Hamming distance \( L(1 - s)/2 \) from the peak sequence, where \( s \) is defined by the equation:\(^{28}\)
\[ R = f(s). \] (6)

We have the non-selective phase with \( k_0 = 0, \) and the following expression for the mean growth rate
\[ R = f(0) \equiv 1, \] (7)
where we have chosen \( f(0) = 1.\)

The Eigen error threshold is defined as a point when the expression of \( R \) jumps from eq. (5) with some \( k_0 \neq 0 \) to eq. (7). For the single peak fitness model with \( f(1) = A \) and \( f(m) = 1 \) for \( m < 1, \) the error threshold is at\(^{10,11}\)
\[ Ae^{-\gamma} = 1. \] (8)

This paper is organized as follows. In §2, we define the Eigen model with correlated multiple mutations and derive its HJE.\(^{23-25}\) In §3, we follow\(^{23-25}\) to solve the HJE and compare the analytical results with numerical solutions to test the reliability of the analytic equations. In §4, we discuss our results.

2. Eigen Model with Correlated Multiple Mutations

2.1 Definition of the model

Consider the model given by eq. (1), which has correlated multiple mutations, we modified the mutation matrix as follows:
\[ Q_{ij} \equiv Q_n = \hat{Q}q^{-n}(1 - q)^sg_n, \] (9)
where \( n = d(i, j), \) \( g_n \) are some coefficients which decrease with the growth of \( n, \) and \( \hat{Q} \) defines the normalization of the probabilities \( Q_{ij}.\) In the original Eigen model,\(^{10,11}\) \( g_n = 1.\)

We consider the symmetric distribution of \( p_i = P_i, \) where \( l = d(i, 0) \) is the Hamming distance (HD) of the \( i \)-th sequence from the master sequence, so that all sequences with the same HD from the master sequence have the same \( p_i = P_i.\) From such a sequence, one can generate a sequence \( S_i \) with HD \( n \) from \( S_i \) through \( n_1 \) up and \( n_2 \) down mutations and \( n = n_1 - n_2.\) The total number of such mutations is
\[ \frac{l!}{n_1!(l - n_1)!n_2!(L - l - n_2)!}. \] (10)

Then we derive the following equation for \( P_l: \)
\[ \frac{dP_l}{dt} = \sum_{n_1=0}^{L-1-i} \sum_{n_2=0}^{L-1-l-i} \frac{l!}{n_1!(l-n_1)!n_2!(L-l-n_2)!} (L - l)! \times Q_{n_1+n_2} P_{l-n_1-n_2}. \] (11)

2.2 Derivation of the HJE equation

We assume the following ansatz for the \( P_l \)
\[ P_l = \exp[Lu(m, t)], \] (12)
where \( m = 1 - 2/2L. \) With 1/L accuracy we replace in eq. (11):
\[ P_{l-n} ightarrow P_l \exp[2nu(m)], \] (13)
where \( u(m) = \partial u(m)/\partial m. \) Then using the formulas
\[ \frac{l!}{(l-n_1)!} \approx P^{n_1} = \left( \frac{L}{2} \right)^{n_1}, \]
\[ \frac{(L - l)!}{(L - l - n_2)!} \approx (L - l)^{n_2} = \left( \frac{L}{2} \right)^{n_2}, \] (14)
we obtain
\[ \frac{dP_l}{dt} = \sum_{n_1=0}^{L-1-i} \sum_{n_2=0}^{L-1-l-i} \frac{l!}{n_1!(l-n_1)!n_2!(L-l-n_2)!} \times Q_{n_1+n_2} \frac{l!}{L^{n_1+n_2}} e^{2nu(n-2)u} P_{l-n_1-n_2}. \] (15)

We consider the case \( l \gg 1, N - l \gg 1. \) Then using an equality
\[ \sum_{n_1=0}^{L} \sum_{n_2=0}^{L} \frac{d^{n_1}n_1}{n_1!} g_{n_1+n_2} = \frac{\sum_{n=0}^{\infty} (a + b)^{n}}{n!} g_n, \] (16)
and re-scale the time \( t \rightarrow t/L, \) from eqs. (12) and (15) we obtain the HJE
\[ \frac{\partial u(m, t)}{\partial t} + H(m, u') = 0, \]
\[ -H = \sum_{n=0}^{L} \left( \frac{1 + m}{2} e^{-2u} + \frac{1 - m}{2} e^{2u} \right) \gamma \frac{g_n}{n!} f(m) \hat{Q}. \] (17)

where \( f(m) = r_1.\)

3. Solution of the Model

Let us calculate the mean fitness following the ideas of Saakian et al.\(^{23-25}\) We assume an asymptotic solution
\[ u(m, t) = kt + u_0(m). \] (18)

Then we have an ordinary differential equation for the \( u_0(m): \)
\[ k = -H(m, u_0'). \] (19)
First we define the potential \( U(m): \)
\[ U(m) = \min[-H(m, p)]_p = \hat{Q} \sum_{n} (1 - m^2)^{n/2} \frac{g_n}{n!} f(m) \gamma^n. \] (20)
To have a real value solution \( u_0(m) \) for \( u_0(m) \) in eq. (20), we put a constraint. Therefore
\[ k \geq U(m), \] (21)
We choose \( R \) as a minimal \( k \) under the constraint eq. (21),\(^{23}\) thus we have an equation for the mean fitness:
\[ R = \max[U(m)]_m. \]  
(22)

Now consider a simple choice of \( g_n \):
\[ g_1 = 1, \quad n \leq 2; \]
\[ g_n = \frac{1}{(n+1)(n+2)(n+3)}, \quad n > 2. \]  
(23)

Equations (9) and (23) imply that for a larger number of \( n \geq 3 \), the probability for \( n \) mutations is smaller than the case of independent multiple mutations (i.e., \( g_n = 1 \)). Thus eq. (23) has the effect of reducing fitness function for a larger \( n \), similar to that of the truncated fitness landscape, in which the fitness function is 0 for \( n \) larger than a critical value \( n_c \). The reduction of fitness function for a larger \( n \) can be understood as follows. For a larger number of \( n \), the probability for silent mutations is reduced, and the probability for missense or nonsense mutations is increased causing the reduction of fitness function. We choose the specific form of eq. (23) because we can obtain the exact solution of the model with this specific form.

From the expansion of \( e^\gamma \), one can easily show that
\[
1 + \gamma + \gamma^2/2 + \sum_{n \geq 3} \frac{\gamma^n}{(n+3)!} = \sum_{n \geq 0} \frac{\gamma^n}{(n+3)!} + \left(1 - \frac{1}{3!}\right) + \left(1 - \frac{1}{4!}\right) \gamma + \left(1 - \frac{2}{4!}\right) \gamma^2 = \frac{e^\gamma - 1 - \gamma}{\gamma^2} + \left(1 - \frac{1}{3!}\right) + \left(1 - \frac{1}{4!}\right) \gamma + \left(1 - \frac{2}{4!}\right) \gamma^2. \]  
(24)

From eqs. (9) and (23) and the normalization of \( Q_n \), one can obtain an expression for \( \dot{Q} \). Using such \( \dot{Q} \), and eqs. (17) and (24), one can obtain the Hamiltonian:
\[
-H(m,u) = \frac{f(m)\gamma^3}{e^\gamma - 1 - \gamma - \frac{1}{2} \gamma^2 + \frac{5}{6} \gamma^3 + \frac{23}{24} \gamma^4 + \frac{59}{120} \gamma^5} \times \left[ \frac{e^{G - 1 - G - G^2/2}}{G^3} \right. \\
+ \left. \frac{5}{6} + \frac{23}{24} G + \frac{59}{120} G^2 \right],
\]
\[ G = \gamma \left(\frac{1 + m}{2} e^{-2u} + \frac{1 - m}{2} e^{2u}\right). \]  
(25)

For the mean fitness we have an expression:
\[
R = \max \left[ \frac{f(m)\gamma^3}{e^\gamma - 1 - \gamma - \frac{1}{2} \gamma^2 + \frac{5}{6} \gamma^3 + \frac{23}{24} \gamma^4 + \frac{59}{120} \gamma^5} \times \left( \frac{w - 1 - w^2}{w^3} + \frac{5}{6} + \frac{23}{24} w + \frac{59}{120} w^2 \right) \right]_m \]
w = \gamma \sqrt{1 - m^2}. \]  
(26)

For the single peak fitness we derive for the mean fitness in the selective phase:
\[ R = \frac{1}{A} \gamma^3 \]
(27)

For the non-selective phase we again have eq. (7). Thus we derive the error threshold condition putting \( R = 1 \) in eq. (27):
\[ 1 = A \frac{1}{A} \gamma^3 \]
(28)

Figure 1 gives the comparison of the critical value of \( A \) for different values of \( \gamma \) for the Eigen model by eq. (8) and our model by eq. (28). Figure 1 shows that for the given values of \( A \) and \( (1 - q) \), eq. (28) gives a larger number of critical chain length than that given by eq. (8).

In Table I we compare the direct numerical results for eqs. (1), (9), and (23) with our analytical solution by eq. (27). Table I shows that our analytical results are very reliable.

The method used in this section can be extended to calculate the mean fitness and the error threshold condition for other functional form of \( g_n \). For example, instead of using eq. (23) for \( g_n \), one can choose \( g_n = 1/n^3 \) with \( n \geq 3 \). However, in this case one should use the numerical method to calculate the mean fitness and the phase diagram.

4. Discussion

In conclusion, we solved the Eigen model for the case of correlated mutations, with reducing the probability of three or more mutations, while holding ordinary expression for the
probability of double mutations. We relaxed the condition on the independence of mutations at different positions of the genome. The numerics confirmed well with our analytical results. There are some experimental results supporting the hypothesis that the mutations in bacteria are not random. In ref. 32 the limited randomness of mutations has been assumed as one of the key features of life. We first speculated that the pre-biotic matter also possess such a property, then constructed the model with the simplest correlation between mutations, and estimated the change of error threshold. In ref. 16 the minimal genome length has been estimated $L = 8000$, which for $1 - q = 0.001$ gives $\gamma = 8$. For the error threshold in our model we have $\lambda = 46$ (if we assume 40% lethal mutations) we need even less $\lambda = 20$, which is sometimes possible in RNA replication reactions. Thus assuming a simple mechanism of attenuation of multiple mutations, we can solve the error threshold paradox. The error threshold in Eigen model has a deep information-theoretical meaning: it coincides with the Shannon constraint for optimal codes to submit the encoded information through the channels with the noise. It will be very interesting to investigate the correlated noise case under the information-theoretical point of view.

Another possible application of our results are RNA viruses, evolving at high mutation rates near the error threshold. The Poisson distribution for the number of multiple mutations is only one of the possible choices.

This work inspires some interesting problems for further study. It is well known that in the Crow–Kimon model of biological evolution with parallel mutation-selection scheme, only one (DNA or RNA) base is changed from one generation to the next generation. Such a model has been extended to the biological evolution model with the general fitness function and multiple mutations. One can further extend this model to a model with correlated multiple mutations and study the influence of the correlation on the error threshold. Recently, analytic equations for finite-size corrections and finite population problem of the molecular evolution models had been derived. One can try to extend such study to Eigen model with correlated multiple mutations considered in this paper.

The extension of the Eigen model with multiple random mutations [in eq. (9), $g_n = 1$] to the Eigen model with multiple correlated mutations [in eq. (9), $g_n \ll 1$] is similar to the extension of random percolation models to percolation models corresponding to the Ising model, the Potts model, and a lattice model for hydrogen-bonding of water molecules. In the correlated percolation models the deviation from the random percolation models can be derived from subgraph expansions of the partition functions of the original lattice models. It is of interest to know whether one can obtain $g_n$ in eq. (9) from some theory in molecular biology.

Acknowledgments

We thank Gwoing Yu for a critical reading of the paper. This work was supported by Academia Sinica, National Science Council in Taiwan with Grant Number NSC 100-2112-M-001-003-MY2 and National Center for Theoretical Sciences (Taipei Branch).