IOPscience iopscience.iop.org

Home Search Collections Journals About Contact us My IOPscience

Finite-size scaling of the error threshold transition in finite populations

This article has been downloaded from IOPscience. Please scroll down to see the full text article.

1999 J. Phys. A: Math. Gen. 32 L1

(http://iopscience.iop.org/0305-4470/32/1/001)

View the table of contents for this issue, or go to the journal homepage for more

Download details:
IP Address: 171.66.16.104
The article was downloaded on 02/06/2010 at 07:23

Please note that terms and conditions apply.

J. Phys. A: Math. Gen. **32** (1999) L1–L7. Printed in the UK PII: S0305-4470(99)97627-8

LETTER TO THE EDITOR

Finite-size scaling of the error threshold transition in finite populations

P R A Campos and J F Fontanari
Instituto de Física de São Carlos, Universidade de São Paulo, Caixa Postal 369, 13560-970 São Carlos SP, Brazil

Received 15 September 1998, in final form 10 November 1998

Abstract. The error threshold transition in a stochastic (i.e. finite population) version of the quasispecies model of molecular evolution is studied using finite-size scaling. For the single-sharp-peak replication landscape, the deterministic model exhibits a first-order transition at $Q = Q_c = 1/a$, where Q is the probability of exact replication of a molecule of length $L \to \infty$, and a is the selective advantage of the master string. For sufficiently large population size, N, we show that in the critical region the characteristic time for the vanishing of the master strings from the population is described very well by the scaling assumption $\tau = N^{1/2} f_a[(Q - Q_c)N^{1/2}]$, where f_a is an a-dependent scaling function.

An elusive issue in the extension of Eigen's quasispecies model [1] of molecular evolution to finite populations is the characterization of the so-called error threshold phenomenon which limits the length of the molecules and, consequently, the amount of information they can store [2]. This phenomenon poses an interesting challenge for the theories of the origin of life, since it prevents the emergence of huge molecules which could carry the necessary information for building a complex metabolism. Moreover, since modern theories of integration of information in pre-biotic systems involve the compartmentation of a small number of molecules, the understanding of the effects of the error propagation in finite populations has become a major issue for the theories of the origin of life [3].

The quasispecies model was originally formulated within a deterministic chemical kinetic framework based on a set of ordinary differential equations for the concentrations of the different types of molecules that compose the population. Such formulation, however, is valid only in the limit where the total number of molecules, denoted by N, goes to infinity. In the binary version of the quasispecies model, a molecule is represented by a string of L digits $(s_1, s_2, \ldots, s_L)$, with the variables s_α allowed to take on only two different values, say $s_\alpha = 0, 1$, each of which representing a different type of monomer used to build the molecule. The concentrations x_i of molecules of type $i = 1, 2, \ldots, 2^L$ evolve in time according to the equations [1, 2]

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \sum_j W_{ij} x_j - \Phi(t) x_i \tag{1}$$

where $\Phi(t)$ is a dilution flux that keeps the total concentration constant. This flux introduces a nonlinearity in equation (1), and is determined by the condition $\sum_i \mathrm{d}x_i/\mathrm{d}t = 0$. In particular, assuming $\sum_i x_i = 1$ yields

$$\Phi = \sum_{i,j} W_{ij} x_j . \tag{2}$$

The elements of the replication matrix W_{ij} depend on the replication rate or fitness A_i of the strings of type i, as well as on the Hamming distance $d(i, j)$ between strings i and j. They are given by [1, 2]

$$W_{ii} = A_i q^{\nu} \tag{3}$$

and

$$W_{ij} = A_j q^{L-d(i,j)}(1-q)^{d(i,j)} \qquad i \neq j \tag{4}$$

where $0 \leqslant q \leqslant 1$ is the single-digit replication accuracy, which is assumed to be the same for all digits.

The quasispecies concept and the error threshold phenomenon are illustrated more neatly for the single-sharp-peak replication landscape, in which we ascribe the replication rate $a > 1$ to the so-called master string, say $(1, 1, \ldots, 1)$, and the replication rate 1 to the remaining strings. In this context, the parameter a is termed the selective advantage of the master string. As the replication accuracy q decreases, two distinct regimes are observed in the population composition in the deterministic case: the *quasispecies* regime characterized by the presence of the master string together with its close neighbours, and the *uniform* regime where the 2^L strings appear in the same proportion. The transition between these regimes takes place at the error threshold q_c, whose value depends on the parameters L and a [1, 2]. However, even in the deterministic case, $N \to \infty$, a genuine thermodynamic order–disorder phase transition will occur only in the limit $L \to \infty$ [4–6]. To study this transition for large L, it is convenient to introduce the probability of exact replication of an entire string, namely,

$$Q = q^L \tag{5}$$

so that the *discontinuous* transition occurs at

$$Q_c = \frac{1}{a} \tag{6}$$

for $L \to \infty$ [1, 6]. A recent finite-size scaling study of the sharpness of the threshold transition indicates that the characteristics of the transition persist across a range of Q of order L^{-1} about Q_c [7].

Although several theoretical frameworks have been proposed to generalize the deterministic kinetic formulation of the quasispecies model so as to take into account the effect of finite N [8–13], the somewhat uncontrolled approximations used in those analyses have hindered a precise characterization of the error threshold for finite populations. In particular, Nowak and Schuster [10] employed a simple birth and death model, whose deterministic limit, however, did not yield the stationary distribution predicted by equation (1), as well as numerical simulations based on Gillespie's algorithm [14] to show that an appropriately defined $Q_c(N)$ tends to the deterministic value $1/a$ with $N^{-1/2}$ for sufficiently large populations. A similar result was obtained by neglecting the possibility of multiple errors occurring during the replication of a molecule [12]. A more drastic approximation that neglects linkage disequilibrium at the population level yields that $Q_c(N)$ increases linearly with $1/N$ [13]. Of course, since there is no generally accepted definition of the error threshold for finite N (and for finite L as well), denoted above by $Q_c(N)$, there are some arbitrariness in those analyses.

In this paper we follow a more direct approach to characterize the error threshold transition for finite N, which dispenses with a definition for $Q_c(N)$. As mentioned before, since a genuine phase transition occurs in the limits $N \to \infty$ and $L \to \infty$ only, we study a stochastic (i.e. finite N) version of the quasispecies model with $L \to \infty$ and $q \to 1$ so that $Q = q^L$ is finite. In this limit the problem simplifies enormously as the probability of any string becoming a master string due to replication errors is of order $1/L$ and so can be safely neglected. Besides, since for the single-sharp-peak replication landscape the strings can be classified in two types only: the master strings and the error tail, which comprises all other strings, the population at any given generation can be described by the single integer $n = 0, 1, \ldots, N$, which gives the number of master strings in the population. The goal then is to calculate the probability distribution that at generation t there are exactly n master strings in the population. This quantity, denoted by $\mathcal{P}_t(n)$, obeys the recursion equation

$$\mathcal{P}_{t+1}(n) = \sum_{m=0}^{N} T(n, m)\mathcal{P}_t(m) \tag{7}$$

with the elements of the transition matrix $\mathbf{T}$ given by

$$T(n, m) = \sum_{k=n}^{m} \binom{N}{k}\binom{k}{n} w_m^k (1 - w_m)^{N-k} Q^n (1 - Q)^{k-n} \tag{8}$$

where

$$w_m = \frac{ma}{N - m + ma} \tag{9}$$

is the relative fitness of the master strings. In writing equation (8) we have followed the prescription used in the implementation of the standard genetic algorithm [15]: first the natural selection process acting via differential reproduction is considered and then the mutation process. A similar formulation was used to study the Muller's ratchet phenomenon in a smooth, multiplicative replication landscape in which only unfavourable mutations are allowed, leading to a continuous decrease in the average fitness of the population [16]. We note that $\sum_n T(n, m) = 1\ \forall m$ and $T(0, 0) = 1$. Moreover, the largest eigenvalue of $\mathbf{T}$ is $\lambda_0 = 1$ and its corresponding eigenvector is $\mathbf{l}_0^\dagger = (1, 0, \ldots, 0)$. This stochastic model is easily recognized as the celebrated Kimura–Crow infinite alleles model [17,18] which has been extensively studied within the diffusion approximation for large N. However, for arbitrary values of Q and a the solutions of the partial differential equations are too complicated to be of any use for our purposes [18].

As for finite N the fluctuations, either in the reproduction or mutation processes, will ultimately lead to the irreversible loss of all copies of the master string from the population, the asymptotic solution of equation (7) is simply $\mathcal{P}_\infty(n) = \delta_{n0}$. Our goal is to determine how the characteristic time, τ, that governs the vanishing of the master strings from the population depends on N, Q and a.

Before proceeding to the analysis of the stochastic problem, it is instructive to briefly discuss the deterministic limit $N \to \infty$. In this case the average number of master strings obeys the recursion equation

$$\begin{aligned}\langle n\rangle_{t+1} &= \sum_{n=0}^{N}\sum_{m=0}^{N} nT(n, m)\mathcal{P}_t(m) \\ &= Qa\langle n\rangle_t \end{aligned} \tag{10}$$

whose solution is $\langle n\rangle_t = (Qa)^t\langle n\rangle_0$. Hence, in the deterministic regime we find

$$\tau = -\frac{1}{\ln(Qa)} \tag{11}$$

which diverges at $Q = Q_c = 1/a$, thus signalling the existence of a phase transition in the limit $N \to \infty$. Clearly, for $Q > Q_c$ the master strings are always present in the population so that τ is infinite in this entire region.

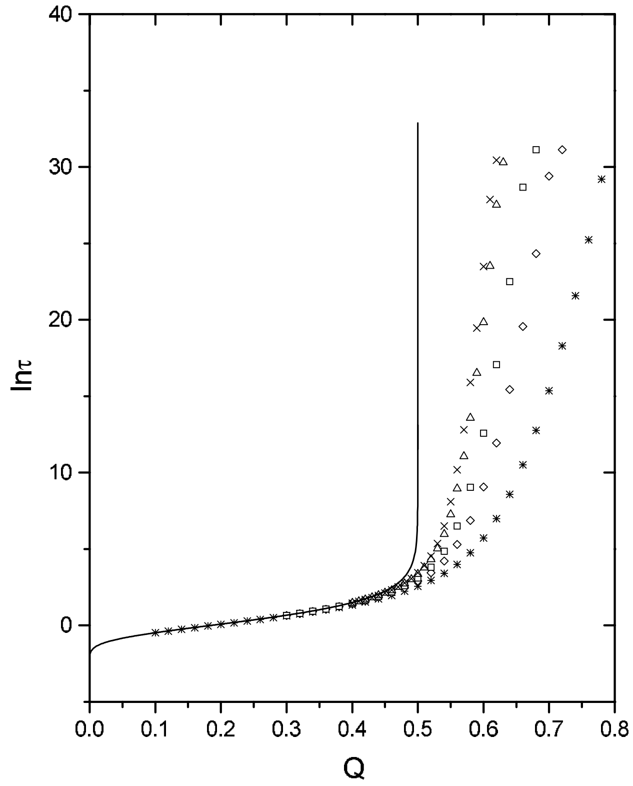

Figure 1. Logarithm of the characteristic time for the vanishing of the master strings from the population, $\ln\tau$, as a function of the probability of exact replication, Q, for $a = 2$, and $N = 100$ ($*$), 200 ($\Diamond$), 300 ($\Box$), 500 ($\triangle$), and 600 ($\times$). The solid line is the theoretical prediction for $N \to \infty$.

We now consider the finite N regime. In this case the recursion equations for the moments of n do not yield useful information since, as usual, the moment of order p depends on the moment of order $p + 1$ evaluated at the previous generation [16]. We then resort to a direct calculation of the probability distribution $\mathcal{P}_t(m)$. More specifically, we will focus on the calculation of $\mathcal{P}_t(0)$, since this is the quantity that measures the rate of vanishing of the master strings from the population. Although $\mathcal{P}_t(0)$ could be evaluated through a series of matrix multiplications, a simple linear algebra calculation yields [18]

$$\begin{aligned}\mathcal{P}_t(0) &= \sum_{n=0}^{N} c_n l_{n0} \lambda_n^t \\ &= 1 + c_1 l_{10} \lambda_1^t + \cdots + c_N l_{N0} \lambda_N^t \qquad (12)\end{aligned}$$

where λ_n are the eigenvalues of $\mathbf{T}$, l_{n0} are the zeroth components of the eigenvectors $\mathbf{l}_n$, and c_n are parameters that depend on the initial state $\mathcal{P}_0(n)$. Also we have used $\lambda_0 = l_{00} = c_0 = 1$. Assuming, without loss of generality, that $1 \geqslant \lambda_1 \geqslant \cdots \geqslant \lambda_N \geqslant 0$, in the limit of large t we find

$$1 - \mathcal{P}_t(0) \approx C\mathrm{e}^{-t/\tau} \qquad (13)$$

where

$$\tau = -\frac{1}{\ln\lambda_1} \qquad (14)$$

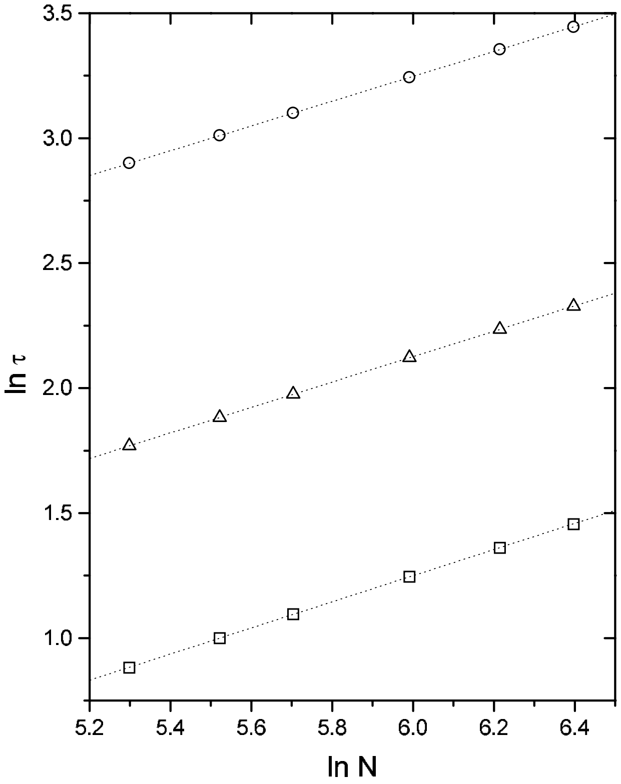

Figure 2. Logarithm of the characteristic time for the vanishing of the master strings from the population, ln τ, calculated at $Q_c = 1/a$ as a function of the logarithm of the population size, ln N, for $a = 2$ (○), 10 (△), and 50 (□).

and C is a constant that depends on the initial state. Thus the problem becomes the one of finding the second largest eigenvalue of the nonsymmetric matrix **T**. Since the largest eigenvalue and its corresponding eigenvector are already known, this numerical problem yields easily to the vector iteration method [19]. Alternatively, we could find τ by following the time evolution of $\mathcal{P}_t(0)$, obtained directly through the recursion equations (7), for a few generations and then plotting $\ln[1 - \mathcal{P}_t(0)]$ against the generation number t. We have verified that both methods yield the same results for τ.

In figure 1 we present the dependence of ln τ on the probability of exact replication of an entire string, Q, for $a = 2$ and several values of N. The finite N effects are negligible for values of Q smaller than, though not too close, Q_c, as indicated by the very good agreement between the finite N data and the theoretical prediction for $N \to \infty$ given in equation (11). Since we expect τ to increase exponentially with increasing N for $Q > Q_c$, and to tend towards its limiting value, equation (11), also exponentially with N for $Q < Q_c$, the issue is then to determine the dependence of τ on N at the critical point $Q = Q_c$. In figure 2 we present ln τ calculated at $Q_c = 1/a$ against ln N for different values of a. These results clearly indicate that at the critical point τ increases like $N^{1/2}$, irrespective of the value of a. Once we have identified the rescaling of ln τ that leads to the collapsing of the data for different N at $Q = Q_c$, the next step is to determine the sharpness of the transition, namely, the range of Q about Q_c where the transition characteristics persists. This is achieved by assuming that

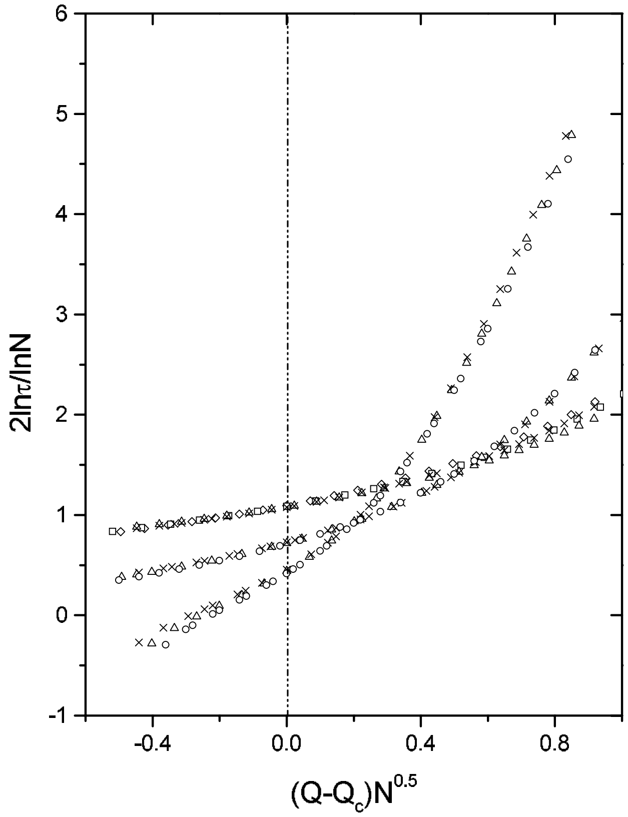

Figure 3. Properly rescaled logarithm of the characteristic time for the vanishing of the master strings from the population, $\ln \tau / \ln N^{1/2}$, as a function of $(Q - Q_c)N^{1/2}$ for (from top to bottom at $Q = Q_c$) $a = 2$, 10 and 50. The convention is $N = 200$ (◇), 300 (□), 400 (○), 500 (△), and 600 (×). For $a = 10$ and 50 only the data for $N \geqslant 400$ are presented.

the size of this region shrinks to zero like $N^{-1/\nu}$ as $N \to \infty$, where the exponent $\nu \geqslant 0$ is estimated using finite-size scaling or, more precisely, the data collapsing method [20]. In figure 3 we show the collapse of the data for different N obtained with $\nu = 2$ for $a = 2$, 10 and 50. Although for $a = 2$ we can achieve a good-quality data collapse using relatively small population sizes ($N \geqslant 200$), for larger values of a, however, a similar quality of collapsing can only be obtained using larger values of N (i.e. $N \geqslant 400$). In summary, the results of the data collapsing method indicate that the dependence of τ on N in the critical region is very well described by the scaling assumption

$$\tau = N^{1/2} f_a[(Q - Q_c)N^{1/2}] \tag{15}$$

where f_a is a scaling function, whose specific form depends on the parameter a.

To appreciate the effect of the selective advantage parameter a on the quality of the data collapsing results presented in figure 3, we next consider in some detail the case $a \to \infty$ and N finite. Using $w_m \to 1$ for $m > 0$ yields

$$T(n, m) = T(n) = \binom{N}{n} Q^n (1 - Q)^{N-n} \qquad m > 0. \tag{16}$$

As before, $T(0, 0) = 1$ and $T(n, 0) = 0$ for $n > 0$. In this case the eigenvalues of $\mathbf{T}$ can easily be analytically calculated yielding $\lambda_0 = 1$, $\lambda_1 = \sum_{n=1}^{N} T(n)$, $\lambda_2 = \cdots = \lambda_N = 0$. Hence,

$$\tau = -\frac{1}{\ln[1-(1-Q)^N]}. \tag{17}$$

Finally, taking the limits $Q \to Q_c = 0$ and $N \to \infty$, we can easily verify that $\nu = 1$ in this limit. This interesting result suggests that uncontrolled approximations and simplifications of the original model which enhance the selective advantage of the master string or the finite population sampling effects are expected to give unreliable estimates of the exponent ν. Moreover, care must be taken in restricting the finite-size scaling analysis to the regime $N \gg a$ to avoid underestimating the value of ν. We note, of course, that the situation of interest is $N \to \infty$ while a remains finite.

To conclude, the collapse of the data for different N into a-dependent scaling functions presented in figure 3 and summarized in the scaling assumption (15) provide a full characterization of the error threshold transition, signalled in our model by the divergence of τ at $Q_c = 1/a$, for large N. We emphasize that the main advantage of our approach is that it does not rely upon any arbitrary definition of error threshold for finite populations.

The work of JFF was supported in part by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). PRAC is supported by FAPESP.

References

[1] Eigen M 1971 *Naturwissenchaften* **58** 465
[2] Eigen M, McCaskill J S and Schuster P 1989 *Adv. Chem. Phys.* **75** 149
[3] Maynard Smith J and Szathmáry E 1995 *The Major Transitions in Evolution* (Oxford: Freeman)
[4] Leuthäusser I 1987 *J. Stat. Phys.* **48** 343
[5] Tarazona P 1992 *Phys. Rev.* A **45** 6038
[6] Galluccio S 1997 *Phys. Rev.* E **56** 4526
[7] Campos P R A and Fontanari J F 1998 *Phys. Rev.* E **58** 2664
[8] Ebeling W and Feistel R 1977 *Ann. Phys., Lpz.* **34** 81
[9] McCaskill J S 1984 *Biol. Cybern.* **50** 63
[10] Nowak M A and Schuster P 1989 *J. Theor. Biol.* **137** 375
[11] Zhang Y-C 1997 *Phys. Rev.* E **55** R3817
[12] Bonnaz D and Koch A J 1998 *J. Phys. A: Math. Gen.* **31** 417
[13] Alves D and Fontanari J F 1998 *Phys. Rev.* E **57** 7008
[14] Gillespie D 1976 *J. Comput. Phys.* **22** 403
[15] Goldberg D E 1989 *Genetic Algorithms in Search, Optimization and Machine Learning* (Reading, MA: Addison-Wesley)
[16] Higgs P G and Woodcock G 1995 *J. Math. Biol.* **33** 677
[17] Kimura M and Crow J F 1964 *Genetics* **49** 725
[18] Crow J F and Kimura M 1970 *An Introduction to Population Genetics Theory* (New York: Harper and Row)
[19] Engeln-Müllges G and Uhlig F 1996 *Numerical Algorithms with C* (Berlin: Springer)
[20] Binder K 1985 *J. Comput. Phys.* **59** 1