Error threshold transition in the random-energy model

Paulo R. A. Campos*
Instituto de Física de São Carlos, Universidade de São Paulo, Caixa Postal 369, 13560-970 São Carlos, São Paulo, Brazil
(Received 23 August 2002; published 18 December 2002)

We perform a statistical analysis of the error threshold transition in quasispecies evolution on a random-energy fitness landscape. We obtain a precise description of the genealogical properties of the population through extensive numerical simulations. We find a clear phase transition and can distinguish two regimes of evolution: The first, for low mutation rates, is characterized by strong selection, and the second, for high mutation rates, is characterized by quasineutral evolution.

DOI: 10.1103/PhysRevE.66.062904 PACS number(s): 87.10.+e, 87.23.Kg

I. INTRODUCTION

Since Eigen's prominent paper [1] describing the evolution of self-replicating units subjected to the action of natural selection and mutations, the investigation of evolutionary models has been a constant issue in the scientific community. The quasispecies model has been used to study a large variety of different settings, including time dependent landscapes [2,3], coevolution [4], spatially extended systems [5], evolution of mutational robustness [6,7], maternal effects [8], and so on. One of the most intriguing results of the quasispecies theory is the existence of an error threshold transition, beyond which the adaptive information in the population is lost. The existence of an error threshold transition has been proved rigorously for a simple fitness landscape, the single-peak landscape [9–12]. The error threshold phenomenon has also been detected in more complex landscapes [12,13].

The similarity between the error threshold transition and the phase transitions observed in certain physical systems has motivated the study of the phenomenon within the context of statistical mechanics. Leuthäusser showed the equivalence between the quasispecies model and a two-dimensional Ising system with nearest-neighbor interactions [14]. Work by Galluccio [11] uses ideas and tools borrowed from polymer theory and statistical mechanics to solve exactly the quasispecies model for the case of a single-peaked landscape.

The resemblance among the landscapes derived from spin glass systems and real biological systems makes the former an appropriate model for replication landscapes [17,18]. In this sense, the NK landscapes proposed by Kauffman and Levin [19] arise as a good description for adaptation landscapes. In investigations of an evolutionary version of the random-energy model (REM) [20], Franz and co-workers showed that for a REM landscape the deterministic quasispecies model also exhibits a phase transition, similar to that found for the single-peaked landscape [21,22]. However, Franz and co-workers found, for small mutation rates, the existence of a *frozen* regime, in which the population is trapped in an adaptation optimum and all individuals in the population are identical. For the single-peaked landscape, the equivalent regime is characterized by a stable coexistence of the master sequence and a cloud of closely related mutants, and does not ensure the existence of only a single molecular species. Although the approach used by Franz and co-workers for solving the quasispecies model for the REM landscape permits the identification of a phase transition, a complete description of the phenomenon for finite systems is still missing.

In this paper, we investigate the random energy landscape for the infinite-sites model [15,16] and consider finite populations. Our formulation is based on a statistical analysis of the genealogical properties of the population. Since the advent of Kingman's theory of coalescence [23], our understanding of the structure of genealogical trees has vastly improved. However, although the coalescence theory is a powerful method for the description of neutral evolution, the inclusion of selection is still a challenge, and many open questions remain [24–26]. Here, we examine the genealogical process by doing rigorous statistics on two summary quantities, the pairwise hamming distance between two individuals in the population, and the time since their last common ancestor. Since these quantities are well known in certain limiting cases, they are well suited for the identification of the different regimes of evolution.

II. THE MODEL

We consider a finite population of constant size N. Each self-replicating unit in the population is represented by an infinite sequence of sites $\mathbf{S}=(s_1,s_2,\ldots)$, where each site s_α can assume two different states, i.e., $s_\alpha \in \{0,1\}$. In the replication process, each sequence acquires k new mutations, with probability P_k given by a Poisson distribution,

$$P_k = e^{-U}\frac{U^k}{k!}, \tag{1}$$

where U is the mean number of mutations per individual per generation. Since the sequences are infinitely long, we assume that new mutations appear only at sites that have never been hit by a mutation, i.e., the probability of producing reverse mutations is zero. Thus, we suppose that the state $s_\alpha=0$ denotes that the site α has maintained its original state throughout the evolution process, whereas $s_\alpha=1$ means that the site α has been hit by a mutation. This model corresponds to the celebrated infinite-sites model [15,16], which

*Electronic address: prac@if.sc.usp.br

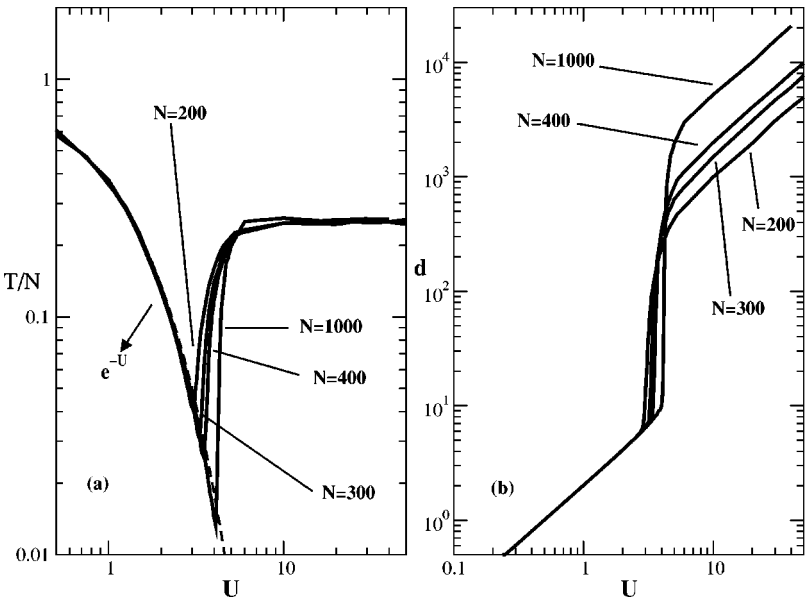

FIG. 1. (a) The average time since the most recent common ancestor (measured in generations) as a function of U (mean number of mutations per individual per generation). (b) The average hamming distance from the consensus sequence (measured in number of digits) as a function of U. In this plot $\beta=5.0$.

has been widely used by population geneticists to describe the DNA variability observed in samples of genes in the case of neutral mutations.

In order to introduce natural selection, we must define the fitness dependency on the genotype configuration **S**. In the random-energy model of Derrida [20], the energy levels are independent random variables, which are drawn from the probability distribution

$$P(E)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp^{-E^2/2\sigma^2} \tag{2}$$

with mean zero and variance σ^2. We assume that each individual contributes offspring to the next generation with probability $p_{off}\sim e^{-\beta E}$, where β is the selective pressure on the population. In the reprodution procedure, we employ the standard genetic algorithm. According to the dynamics, those individuals with higher p_{off} values have a better chance to contribute to the composition of the next generation with one or more offspring. We consider nonoverlapping generations, i.e., after the reproduction step the old sequences are removed from the population.

Since we assume infinitely long sequences, which results in an infinitely large genotype space, the mean number of distinct sequences that are produced with time is given by the relation $N_s=Nt(1-e^{-U})$, where t is the number of generations and the term $(1-e^{-U})$ is the probability of erroneous replication. So N_s also corresponds to the number of states that are visited in an adaptive trajectory whose duration is t. Henceforth we will consider that the variance σ^2 in Eq. (2) is a function of N_s. Since in the original random-energy model the variance is a function of the logarithm of the number of states, we assume that the variance is given by

$$\sigma^2=1/\ln(N\times t_{simul}), \tag{3}$$

where t_{simul} is the maximum time of simulation. We neglect the term dependent on the mutation rate U, since it does not affect the final result. We verified that the use of other forms for the variance does not reproduce the critical character of the model to be shown in the following.

III. THE GENEALOGICAL PROPERTIES AND RESULTS

The relevant features of genealogical trees can be obtained from the matrix $\mathbf{T}=\{T_{\alpha\gamma}\}$, whose elements $T_{\alpha\gamma}$ describe the time in generations since the latest common ancestor of individuals α and γ, and from the matrix $\mathbf{d}=\{d_{\alpha\gamma}\}$, where the elements $d_{\alpha\gamma}$ are the hamming distances between individuals α and γ, i.e., the number of digits that differ in the two sequences [27]. During each step of evolution, we update both matrices, and thus we estimate the mean time since the last common ancestor between two individuals, denoted by T, and given by

$$T=\frac{1}{N(N-1)}\sum_{\alpha,\gamma} T_{\alpha\gamma}, \tag{4}$$

and the average pairwise hamming distance

$$d=\frac{1}{N(N-1)}\sum_{\alpha,\gamma} d_{\alpha\gamma}. \tag{5}$$

In Fig. 1 we present T/N and d as a function of U for different values of N. The averages were taken after the system had reached the steady state regime. We let the system evolve for 98 000 generations, and then carried out a temporal averaging in the following 2000 generations. The final results were obtained after averaging over 100 independent samples for $N=200$, 300, and 400, and over 50 independent samples for $N=1000$. At $U=0$ the system is in the neutral regime and so $T/N=1$ (not shown on the log scale). As we increase U we observe that the data collapse on a single curve that shows an exponential decay. Specifically, we find that in this regime the quantity T/N is well described by the relation $T/N=e^{-U}$, as corroborated in the figure. The quantity e^{-U} is just the probability of exact replication as intro-

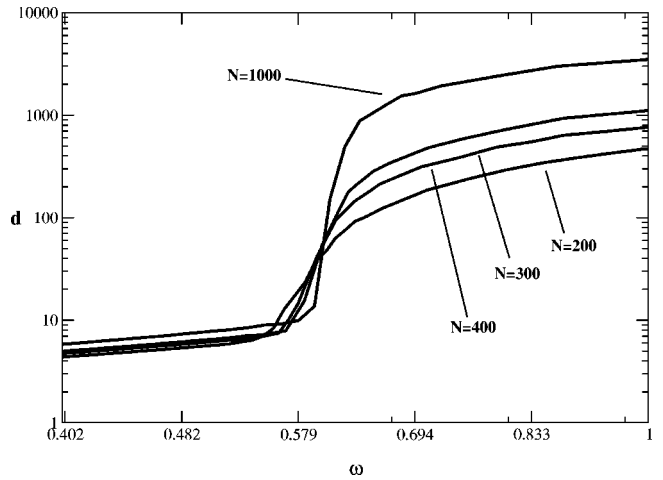

FIG. 2. The mean hamming distance d (measured in number of digits) as a function of the rescaled variable ω ($U/\ln N$) on a log-log scale. In this plot $\beta=5.0$.

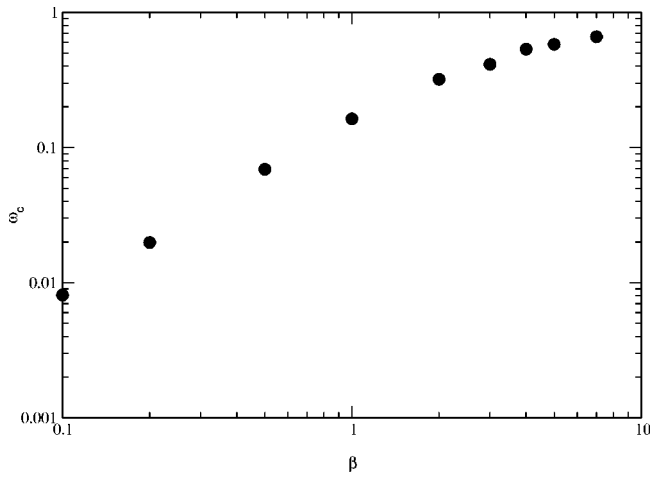

FIG. 3. Phase diagram: the critical point ω_c as a function of β.

duced in the model. This regime corresponds to a strong selection limit where only the best adapted individuals contribute to the offspring generation. This result is confirmed in part (b) of the same figure. In this region of U we find that $d\approx 2U$ as expected in the strong selection limit.

From Fig. 1(a) we also notice that the point of minimum of T/N shifts to the right when we increase the population size. Beyond the minimum value of U, the mean time T rapidly reaches a constant value that seems to be indenpendent of N. In Fig. 1 (b), this behavior is also clearly seen. We observe that, when we increase U, a transition occurs, and the system then changes from an ordered regime to a disordered regime. The ordered phase corresponds to a strong selection regime, at which only those individuals with the highest fitness value can replicate. In this phase the average hamming distance between any pair of individuals is twice the mean number of mutations in a single lineage, i.e., $d=2U$, as discussed above. The disordered phase can be visualized as a quasineutral regime. In this phase, the lines for differents values of N are parallel to the neutral solution $d_{neut}=2UN$, i.e., $d=2CUN$, where C is a constant and thus the product CN can be interpreted as a measurement of an effective population in the population. Actually, the constant C is approximately equal to the variance of the offspring distribution as was pointed out in Ref. [28]. Interestingly, the point of minimum in T/N corresponds to the transition point observed for d. We can understand the quantity d/N as an order parameter of the model, and U is the control parameter.

As usual in critical phenomena, the transition point $U=U_c$ depends on the system size N. This suggests that we need to find a rescaling for the control parameter, in order to obtain a variable that exhibits a unique value for the transition point. In order to measure the dependence of the critical point $U=U_c$ on N, we will focus on the limiting case $\beta\rightarrow\infty$. In this limit, we know that only those individuals that have the highest fitness in the population are able to reproduce. However, in order to maintain their selective advantage for future generations, at least one of those individuals must be able to replicate exactly, i.e., it must not acquire any mutation on its genome. Since the probability of exact replication is $Q=e^{-U}$, the critical point $U=U_c$ at which the selective information is no longer sustained, is determined by

$$N\times Q\sim 1. \tag{6}$$

Therefore, we deduce that U_c is approximately given by $U_c\approx\ln N$. Thus the correctly rescaled variable seems to be $\omega=U/\ln N$. In Fig. 2 we plot d as a function of the rescaled variable ω for different values of population size. As we can see, there is a single point $\omega=\omega_c$ at which all the lines for system sizes intersect with each other.

In Fig. 3 we present the phase diagram ω_c versus β. We can observe that ω_c is a monotonically increasing function of the selective strength β, although the increasing of β results in a smaller increase of ω_c.

To conclude, the determination of the phase diagram for the model permits a complete characterization of the error threshold transition, and also its generalization to finite population sizes. In the previous investigations by Franz and coworkers [21,22], in addition to the lack the characterization of the error threshold transition for finite systems, a complete understanding of the phase diagram was difficult to obtain, because of the definition of the parameter ω_c as a function of the probability of mutation per site of the sequence, which is meaningless for infinite sequences.

ACKNOWLEDGMENTS

P.R.A.C. is grateful to J. F. Fontanari and C. O. Wilke for helpful suggestions and discussions about the problem. This research was supported by Fundação de Amparo à Pesquisa do Estado de São Paulo, Project No. 99/09644-9.

[1] M. Eigen, Naturwissenschaften **58**, 465 (1971).
[2] M. Nilsson and N. Snoad, Phys. Rev. Lett. **84**, 191 (2000).
[3] C.O. Wilke, C. Ronnewinkel, and T. Martinetz, Phys. Rep. **349**, 395 (2001).
[4] C. Kamp and S. Bornholdt, Phys. Rev. Lett. **88**, 068104 (2002).
[5] S. Altmeyer and J.S. McCaskill, Phys. Rev. Lett. **86**, 5819 (2001).
[6] C.O. Wilke, J.L. Wang, C. Ofria, R.E. Lenski, and C. Adami, Nature (London) **412**, 331 (2001).
[7] E.V. Nimwegen, J.P. Crutchfield, and M. Huynen, Proc. Natl. Acad. Sci. U.S.A. **96**, 9716 (1999).
[8] C.O. Wilke, Phys. Rev. Lett. **88**, 078101 (2002).
[9] M. Eigen, J. McCaskill, and P. Schuster, J. Phys. Chem. **92**, 6881 (1988).
[10] P.R.A. Campos and J.F. Fontanari, Phys. Rev. E **58**, 2664 (1998).
[11] S. Galluccio, Phys. Rev. E **56**, 4526 (1997).
[12] P. Tarazona, Phys. Rev. A **45**, 6038 (1992).
[13] T. Wiehe, Genet. Res. Comb. **69**, 127 (1997).
[14] I. Leuthäusser, J. Chem. Phys. **84**, 1884 (1986).
[15] J.F. Crow and M. Kimura, *An Introduction to Population Genetics Theory* (Harper and Row, New York, 1970).
[16] G.A. Watterson, Theor Popul. Biol. **10**, 256 (1975).
[17] P.F. Stadler and R. Happel, J. Phys. A **25**, 3103 (1992).
[18] E.D. Weinberger and P.F. Stadler, J. Theor. Biol. **163**, 255 (1993).
[19] S.A. Kauffman and S. Levin, J. Theor. Biol. **128**, 11 (1987).
[20] B. Derrida, Phys. Rev. B **24**, 2613 (1981).
[21] S. Franz, L. Peliti, and M. Sellitto, J. Phys. A **26**, L1195 (1993).
[22] S. Franz and L. Peliti, J. Phys. A **30**, 4481 (1997).
[23] J.F.C. Kingman, J. Appl. Probab. **19**, 27 (1982).
[24] S.M. Krone and C. Neuhauser, Theor Popul. Biol. **51**, 210 (1997).
[25] C. Neuhauser and S.M. Krone, Genetics **145**, 519 (1997).
[26] C. Wiuf and J. Hein, Genetics **151**, 1217 (1999).
[27] P.G. Higgs and B. Derrida, J. Mol. Evol. **35**, 454 (1992).
[28] C.O. Wilke, P.R.A. Campos, and J.F. Fontanari, J. Exp. Zool., Mol. Dev. Evol. (to be published).