Error Threshold for Spatially Resolved Evolution in the Quasispecies Model

S. Altmeyer and J. S. McCaskill

Biomolecular Information Processing, BioMIP, GMD-German National Research Center for Information Technology, Schloss Birlinghoven, 53754 St. Augustin, Germany

(Received 6 July 2000)

The error threshold for quasispecies in 1, 2, 3, and ∞ dimensions is investigated by stochastic simulation and analytically. The results show a monotonic decrease in the maximal sustainable error probability with decreasing diffusion coefficient, independently of the spatial dimension. It is thereby established that physical interactions between sequences are necessary in order for spatial effects to enhance the stabilization of biological information. The analytically tractable behavior in an ∞-dimensional (simplex) space provides a good guide to the spatial dependence of the error threshold in lower dimensional Euclidean space.

DOI: 10.1103/PhysRevLett.86.5819 PACS numbers: 87.10.+e, 87.23.Cc, 87.23.Kg

The quasispecies theory provides a physically grounded kinetic model for the evolution of biological information [1]. For large, well-mixed, and noninteracting populations, it provides the first answer to the question as to how much information can be maintained stably in a population of erroneously self-replicating molecular sequences and thereby also to the question how far simple physicochemical interactions can take matter towards complex biological information. The sustainable length of specific sequence information is limited by an error threshold which depends primarily on the probability of errors during replication, but also on differences in replication rates of the sequences involved [2–4]. The error threshold can be computed as a function of the statistical distribution of sequence dependent replication rates [5], somewhat analogously to the electron localization threshold phenomenon of disordered solids [6]. Although there are some recent papers that deal with the physical aspects of the quasispecies model [7,8], it is curious that there appears to have been no detailed investigation of the interplay between physical space and evolution for that simple model. Most attention has naturally focused instead on the key role of space in fostering the stable coevolution of interacting sequences [9,10]. However, one would like to be able to compare the information dynamics, in such more complex spatially resolved systems with nonlinear kinetics, to the noninteracting quasispecies model basal case.

In the deterministic homogeneous quasispecies model, different heteropolymers composed of κ types of monomers with sequences i of length ν and with concentrations x_i can replicate and mutate as governed by the differential equations

$$\frac{dx_i}{dt} = \sum_j \mathcal{W}_{ij} x_j - \Phi x_i \,, \tag{1}$$

where the total concentration $\sum_i x_i$ is held constant by a compensating dilution flux with rate coefficient Φ. Making the standard assumption of a uniform fidelity q of copying each monomer in a sequence (leading to the whole sequence fidelity $Q = q^\nu$), the net production rate of sequence i from sequence j, $\mathcal{W}_{ij}$, depends on their hamming distance $d(i, j)$ (i.e., number of monomer substitutions necessary to convert sequence i to sequence j);

$$\mathcal{W}_{ij} = \left(\frac{q^{-1} - 1}{\kappa - 1}\right)^{d(i,j)} \mathcal{A}_j Q - \mathcal{D}_j \delta_{ij} \,, \tag{2}$$

with $\mathcal{A}_j$ and $\mathcal{D}_j$ denoting the rates of replication and degradation, respectively. For $i = j$, $\mathcal{W}_{ij}$ corresponds to the net rate coefficient for accurate replication, $\mathcal{A}_i Q - \mathcal{D}_i$, and for $i \neq j$, it corresponds to mutation. In the homogeneous quasispecies model, there is a minimum fidelity Q_c for which the steady state concentration of the fastest replicating sequence is significantly greater than the uniform background level $x_i = 1/\kappa^\nu$.

The properties of the model depend on the mapping of sequences to the rate coefficients $\mathcal{A}_i$ and $\mathcal{D}_i$ appearing in Eq. (2), called the fitness landscape. To expose the main effects of space most clearly, attention is restricted to conventionally the simplest nontrivial landscape (with $\mathcal{D}_i = 0$), consisting of a single peak at $i = 0$ (a particular sequence called the wild type) with magnitude $\mathcal{A}_0 = \sigma$ (called the "superiority parameter") above a lower plane of identical values $\mathcal{A}_{i\neq 0} = 1$ (which can be chosen arbitrarily as it rescales the time only) [11]. In this Letter, the approximation of neglecting the back mutation from the pool of sequences to the master sequence is made as in [2] and becomes valid for long sequences since the mean probability for such a process is proportional to $\nu^{-1}\kappa^{-\nu}$. The approximation allows the complex reactions which follow from Eq. (1) to be simplified by lumping the sequence space into two parts with concentrations $x = x_0$ and $y = \sum_{i\neq 0} x_i$. With this definition, the deterministic kinetics become

$$\frac{dx}{dt} = \sigma Q x - \Phi x \,, \tag{3}$$

$$\frac{dy}{dt} = \sigma(1 - Q)x + y - \Phi y \,, \tag{4}$$

which provides a good approximation to the critical error threshold as shown in [2]. Moreover, this simplified

 © 2001 The American Physical Society

case will prove to be accessible to analytical treatment in ∞-dimensional space.

In spatially resolved models, local fluctuations cannot be neglected as is possible in the kinetic equations of the quasispecies model. A discrete stochastic formulation is adopted below as in [3], where the rate coefficients determine transition probabilities (proportional to dt). The deterministic quasispecies model involves a constant total concentration. Various stochastic models of constant population size N have been employed [12] and lead to similar results. Here, a Moran model was adopted, in which individuals are lost by replacement with replication products, so that only fluctuations in composition are considered, not in population size. This standard general framework of stochastic reaction kinetics by the birth and death master equations [13] is employed locally.

The overall population is divided into discrete local subpopulations (sites with n individuals) on a lattice. The spatial dimension d is then reflected in the connection topology of neighboring sites. In addition to local reactions, individuals may exchange position between adjacent sites stochastically. The rate of such exchanges is described by the diffusion coefficient D. The simplex topology, corresponding to ∞-dimensional space, where every site is a neighbor of every other, will turn out to admit an exact analytical solution for small n. This topology is equivalent to the island model of Wright [14], in which migration takes place with equal probability between any pair of islands.

Let k be the number of wild-type individuals on a site, and $\bar{k}$ their mean number in the system with $\bar{k} = \sum_{k=0}^{n} k P_k(t)$. $P_k(t)$ gives the probability of there being k wild types on a chosen site at time t. Stochastically, the kinetics of the system are described by the (birth and death) master equations for $0 \le k \le n$ (with $P_{-1} = P_{n+1} = 0$):

$$\frac{dP_k}{dt} = w_{k+1\to k} P_{k+1} + w_{k-1\to k} P_{k-1} - (w_{k\to k-1} + w_{k\to k+1}) P_k \,, \tag{5}$$

setting $P_k \equiv P_k(t)$. The transition probabilities for increase and decrease of k, according to Eqs. (3) and (4) are

$$w_{k\to k+1} = \sigma Q k \frac{n-k}{n-1} + D\zeta \frac{n-k}{n} \,, \tag{6}$$

$$w_{k\to k-1} = \sigma(1-Q) k \frac{k-1}{n-1} + (n-k)\frac{k}{n-1} + D(n-\zeta)\frac{k}{n} \,. \tag{7}$$

The quantity ζ is the mean number of wild-type individuals on all other sites, and for a sufficiently large number of sites this is uncorrelated with k, i.e., $\zeta = \bar{k}$. For a finite number of coupled sites, as in lower dimensional space, this would require a mean field approximation. The origin of the other nondiffusive terms can be seen by comparing with Eqs. (3) and (4). In the stationary state, $dP_k/dt = 0$, one obtains the recurrence relation

$$P_k = P_0 \prod_{i=0}^{k-1} \frac{w_{i\to i+1}}{w_{i+1\to i}} \,, \tag{8}$$

where the probability of having zero wild-type sequences in a site P_0 is given by the normalization condition $\sum_{k=0}^{n} P_k = 1$ and $P_0 = (1 + \sum_{k=1}^{n} \prod_{i=0}^{k-1} \frac{w_{i\to i+1}}{w_{i+1\to i}})^{-1}$. By setting $a_k = \frac{w_{k+1\to k}}{w_{k\to k+1}} \prod_{i=0}^{k} \frac{w_{i\to i+1}}{w_{i+1\to i}}$, the self-consistency equation $\zeta = \bar{k} = \sum_{k=0}^{n} k P_k$ can be written as

$$\zeta \sum_{k=0}^{n} a_k = \sum_{k=0}^{n} k a_k \,, \tag{9}$$

which defines $Q(\zeta)$ implicitly for arbitrary n since the a_k depend on Q. For $n = 2$, one obtains after some manipulation $Q(\zeta) = \frac{\sigma + D[1+\zeta(\sigma-1)/2]}{\sigma(1+D)}$, with exponentially more complicated polynomial equations needing to be solved for higher n. Near the error threshold, ζ is small, allowing the critical error probability $R_c = 1 - Q(\zeta \to 0)$ to be calculated explicitly as a function of the diffusion rate D:

$$R_c = \frac{D(\sigma - 1)}{\sigma(1 + D)} \,, \tag{10}$$

which is the main result of the analytical calculation.

For the numerical results, a grid architecture with two molecules per site (as for the $n = 2$ case) was used both for simulations in the simplex case ($d = \infty$) and for the finite spatial dimensions $d = 1, 2, 3$ (which have not proven tractable by analytical means). The simulations involve a Monte Carlo simulation of the stochastic kinetics, with the states of the system described by the local population number $k \in \{0, 1, 2\}$ at each site. The transition probabilities are defined as in Eqs. (6) and (7) at each time step; the sum of the next reaction or molecule interchange between sites is chosen with a weighting factor proportional to these transition probabilities using a variant of the Gillespie algorithm [15]. In contrast with the above analytical model, in lower spatial dimensions the simulations potentially allow the buildup of local correlations between neighboring sites beyond those implied by the mean population sizes.

Two values of the superiority parameter σ (2 and 10) were chosen to illustrate the main cases of relatively mild and strong selection pressures. The critical error threshold R_c in the simulation is simply defined by the highest error probability $R = 1 - Q$, where the population of the wild-type sequence survives. An overall population size of $N = 120\,000$ was used which was checked to be sufficiently large for the results to be independent of N. Moreover, the simulations were done on a time scale of some 10 000 complete updates (generations) which assures that the system is in the steady state. The results of the simulations are shown in Fig. 1. In each dimensionality, a monotonic dependence of the critical error probability R_c on the diffusion rate D was obtained. In the case of low

diffusion rate, the number of sites visited by an individual on the time scale of replication (accessible sites) becomes small, which implies a low effective population size. It is known that the homogeneous error threshold value tends to zero for small populations [3,4,16], and this is also true for the spatial stochastic analysis and simulations here in the limit of low diffusion. The result of the simulation for the simplex case and of the analytical description are compared in Fig. 2. One obtains excellent agreement even for the situation of extremely low diffusion.

The basic properties of the model should be independent of the chosen site size, n, once the space scale induced by diffusion is larger than this size. For the infinite-dimensional case, the scaling behavior for different n can be deduced for moderate and high diffusion rates, making use of the simple observation that a diffusing molecule will never return to a former site, since its number is infinite. Hence, a molecule visiting m sites once, meets $m(n-1)$ others. Two different site sizes (n, n') can be related by rescaling diffusion coefficients (D to D') so that the number of molecules encountered in time τ, $n_\tau = (1 + D\tau)(n-1)$, is the same. One obtains

$$D' = \frac{n-1}{n'-1} D + \frac{n-n'}{n'-1} 1/\tau \approx \frac{n-1}{n'-1} D , \quad (11)$$

for sufficiently large D. Comparing site sizes n with the reference case $n' = 2$, combining Eqs. (10) and (11),

$$R_c(n) \approx \frac{D(n-1)(\sigma-1)}{\sigma[1 + D(n-1)]} . \quad (12)$$

The scaling behavior of $R_c(n)$ for large and moderate D (see Fig. 3) is well captured by Eq. (12) and reaches the correct large population limit for all n as $D \to \infty$. As seen in Fig. 1, the behavior is similar in lower dimensions.

The quasispecies model describes the limitations due to errors in large well-mixed (haploid) populations [2] and approximate expressions have been derived for the case of finite populations [3,5]. Locally, finite population effects play a vital role, so a continuous population approximation in terms of partial differential equations has been avoided. The infinite-dimensional limit strongly resembles the island model first proposed by Wright [14]. This theory was recently extended by some authors (e.g., [17–19]). In another context, Kimura [20] studied approximations in simplex topologies for the limit of large island populations.

In this Letter, an exact analytical description of the stochastic behavior for the smallest site size n has been achieved. An exact analytical treatment is still possible for $n = 3$ and $n = 4$, and for higher n, the characteristic polynomial of Eq. (9) can be solved numerically. However, it was also demonstrated that by scaling the diffusion coefficient appropriately, the $n = 2$ case describes the reduction in maximal sustainable error probability, R_c, down to diffusion rates which isolates individuals on single sites for time scales similar to that of replication. The reduction of the critical error threshold with decreasing diffusion coefficient is similar to that induced by finite populations. This connection between effective population size and diffusion coefficient was also used to

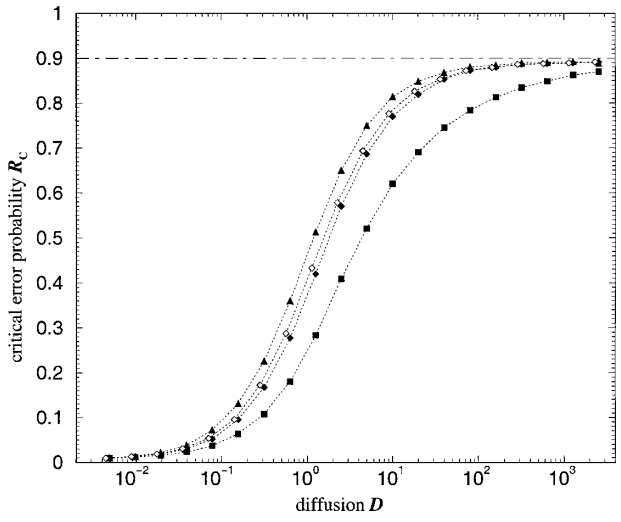

FIG. 1. Spatial dependence of critical error probability in quasispecies model: The graph shows the simulation results for the critical error probability R, sequence length $\nu = 20$, and superiority $\sigma = 10$ as a function of diffusion rate D for various spatial dimensions $d = 1$ (■), $d = 2$ (◆), $d = 3$ (◇), and $d = \infty$ (▲). Observe that R is a monotonic function of D, independent of the spatial dimension, and that the simplex case does indeed behave as an ∞-dimensional limit. For high diffusion, one obtains the results for the well-stirred case $R \approx 1 - 1/\sigma$.

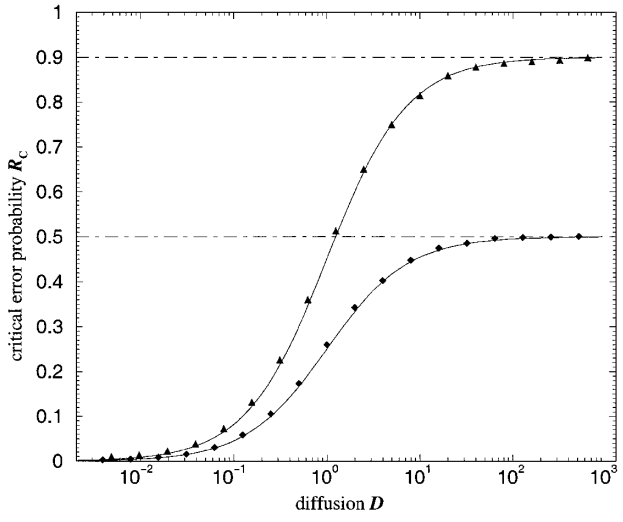

FIG. 2. Critical error probability as a function of diffusion coefficient for two different selection strengths: The graph shows the result of the analytical calculation (solid lines) for the ∞-dimensional case to be in good agreement with the results of the numerical simulation for $\sigma = 10$ (▲) and $\sigma = 2$ (◆). As the diffusion coefficient gets higher, the accessible space for single molecules also becomes larger and, in the limit of infinitely large diffusion, the error probability reaches the value for the homogeneous quasispecies (marked as a horizontal dashed line).

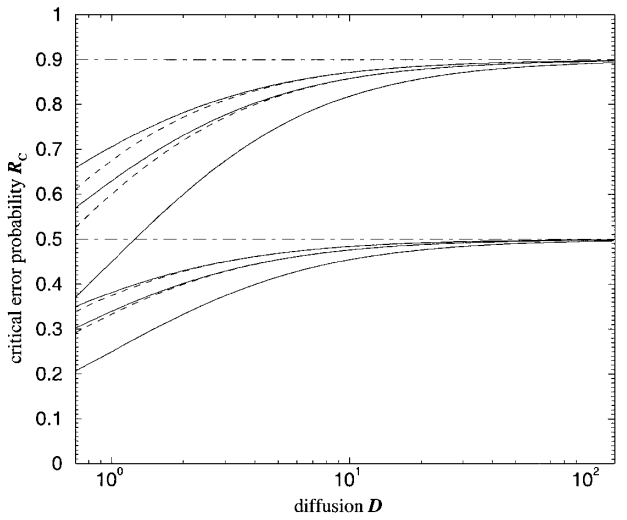

FIG. 3. Population size dependence and scaling behavior of the critical error probability for higher diffusion rates: The graph shows the results of the analytical calculation (solid lines) for $n = 2, 3, 4$ (from bottom to top) for $\sigma = 10$ (upper half, converging to 0.9) and $\sigma = 2$ (lower half, converging to 0.5). For the diffusion range shown, the scaling results of Eq. (12) (dashed lines) from $n = 2$ to $n = 3$, and to $n = 4$ provide a good approximation becoming exact in the high D limit.

derive the scaling approximation given by Eq. (12). In the limit of zero diffusion, the two state assumption given by Eqs. (3) and (4) with neglect of mutational backflow from error copies makes $P_0 = 1$ an absorbing state for isolated sites. However, since the rate of mutational backflow is small, proportional to $\nu^{-1}\kappa^{-\nu}$, this recreation by mutation is much more seldom than repopulation by diffusion for physically reasonable D (see also below), and this explains why simulations with the full quasispecies model including mutational backflow agreed qualitatively well with the two state assumption for finite D.

What are the experimental conditions necessary for diffusion limitation of the error threshold? In continuous aqueous solutions in three-dimensional space, typical diffusion times for informational macromolecules, such as RNA with typical lengths of about 100 bases, are 2×10^{-11} m^2/s. On the time scale τ of replication, several seconds in *in vitro* experiments with RNA [21], a characteristic length scale $L = \sqrt{D\tau}$ of 4×10^{-5} m results. A cube of this dimension has volume ca. 6×10^{-11} liters, so that very small concentrations, around 100 fM, are necessary to sufficiently isolate individuals to see the spatial effects on the error threshold. However, in gels or porous media with microscopic chambers on the spatial scale of 1 μm and below, more readily detectable concentrations up to nanomolar range may be employed. One of the main advantages of small local population sizes is that their use allows cooperative interactions between sequences to be selected for, as in cells. Achieving understanding in this regime has been the prime motivation for this work.

For moderate values of the diffusion rate, in which individuals enter several sites on the time scale of selection, we have shown that the results become independent of the number of particles per site, when the diffusion constant is rescaled to preserve the number of accessible individuals in a given time. For high diffusion rates, where the effects of spatial structure are limited, the error probability reaches asymptotically the limit of the well-stirred case. The conclusion is that the effect of spatial correlations or compartmentation is universally negative on the amount of information which can be generated by selection in simple noninteracting Darwinian populations. In a forthcoming paper we will show how this result is lifted in the case of interacting populations of sequences.

The authors thank Dr. R. Füchslin and Dr. J. Ackermann for fruitful discussions and helpful comments.

[1] M. Eigen, Naturwissenschaften **58**, 465 (1971).
[2] M. Eigen, J. S. McCaskill, and P. Schuster, Adv. Chem. Phys. **75**, 149 (1989).
[3] J. S. McCaskill, Biol. Cybernet. **50**, 63 (1984).
[4] M. Nowak and P. Schuster, J. Theor. Biol. **137**, 375 (1989).
[5] J. S. McCaskill, J. Chem. Phys. **80**, 5194 (1984).
[6] P. W. Anderson, Phys. Rev. **109**, 1492 (1958).
[7] E. Baake, M. Baake, and H. Wagner, Phys. Rev. Lett. **78**, 559 (1997).
[8] M. Nilsson and N. Snoad, Phys. Rev. Lett. **84**, 191 (2000).
[9] M. Boerlijst and P. Hogeweg, Physica (Amsterdam) **48D**, 17 (1992).
[10] R. Durrett and S. Levin, Theor. Popul. Biol. **46**, 363 (1994).
[11] J. Swetina and P. Schuster, Biophys. Chem. **16**, 329 (1982).
[12] J. F. Crow and M. Kimura, *Introduction to Population Genetics Theory* (Burgess Publishing Company, Edina, MN, 1970).
[13] I. G. Darvey, B. W. Ninham, and P. J. Staff, J. Chem. Phys. **45**, 2145 (1966).
[14] S. Wright, Genetics **16**, 97 (1931).
[15] D. Gillespie, J. Phys. Chem. **81**, 2340 (1977).
[16] P. R. A. Campos and J. F. Fontanari, J. Phys. A **32**, L1 (1999).
[17] M. Nei and N.Takahata, J. Mol. Evol. **37**, 240 (1993).
[18] A. Caballero, Genetics **139**, 1007 (1995).
[19] J. Wang, Theor. Popul. Biol. **55**, 176 (1999).
[20] M. Kimura, Proc. Natl. Acad. Sci. U.S.A. **80**, 6317 (1983).
[21] J. S. McCaskill and G. J. Bauer, Proc. Natl. Acad. Sci. U.S.A. **90**, 4191 (1993).