J. theor. Biol. (1989) **137**, 375-395

Error Thresholds of Replication in Finite Populations Mutation Frequencies and the Onset of Muller's Ratchet†

MARTIN NOWAK AND PETER SCHUSTER‡

Institut für Theoretische Chemie, Universität Wien, Währingerstraße 17, *A*1090 *Wien, Austria*

(*Received* 30 *June* 1988, *and accepted in revised form* 6 *December* 1988)

The occurence of thresholds for error propagation in asexually replicating populations is investigated by means of a simple *birth* and *death* model as well as by numerical simulation. Previous results derived for infinite population sizes are extended to finite populations. Here, replication has to be more accurate than in infinitely large populations because the master sequence can be lost not only by accumulation of errors—similar to the loss of wildtype through the operation of *Muller's ratchet*—but also by natural fluctuations. An analytical expression is given which allows straight computation of highly accurate values of error thresholds. The error threshold can be expanded in a power series of the reciprocal square root of the population size and thus increases with $1/\sqrt{N}$ in sufficiently large populations.

1. The Concept of the Error Threshold

About 20 years ago Eigen (1971) conceived a dynamic theory of molecular evolution which is based on the kinetics of polynucleotide replication. He describes error-free replication and mutation as parallel reactions of the same class (Fig. 1) and hence, the model is able to handle all scenarios from rare to frequent mutations. The rare mutation case has been studied extensively in population genetics. Selection then leads to homogeneous populations unless neutral mutants introduce random drift. If mutations occur frequently, the goal of selection is no longer a single fittest type but rather an ensemble of types which centers around a most frequent type denoted as the *master sequence*—in the case of selective neutrality two or eventually more sequences may form the *inner core* of the mutant distribution (Schuster & Swetina, 1988). The stationary sequence distribution of a replicating ensemble has been characterized as *quasispecies* (Eigen & Schuster, 1977) in order to point at the analogy to the conventional notion of species in biology. (For recent updated reviews of quasispecies theory see Eigen *et al.*, 1988, 1989).

The nature of the quasispecies is illustrated best by considering its internal structures in the two limits of very accurate and very inaccurate replication. Some assumptions whose justifications will be discussed in Section 6 are used throughout this paper: we consider binary sequences rather than true polynucleotides. In addition, we assume uniform error frequencies at all positions of the polynucleotide sequences and then, replication accuracy is determined by only one parameter: the *single digit accuracy q*. The error rate per digit is simply given by $p = 1 - q$. Apparently,

† This work was supported by the Hochschuljubiläumsstiftung Wien and IBM Österreich.

‡ To whom all correspondence should be addressed.

0022-5193/89/080375+21 $03.00/0

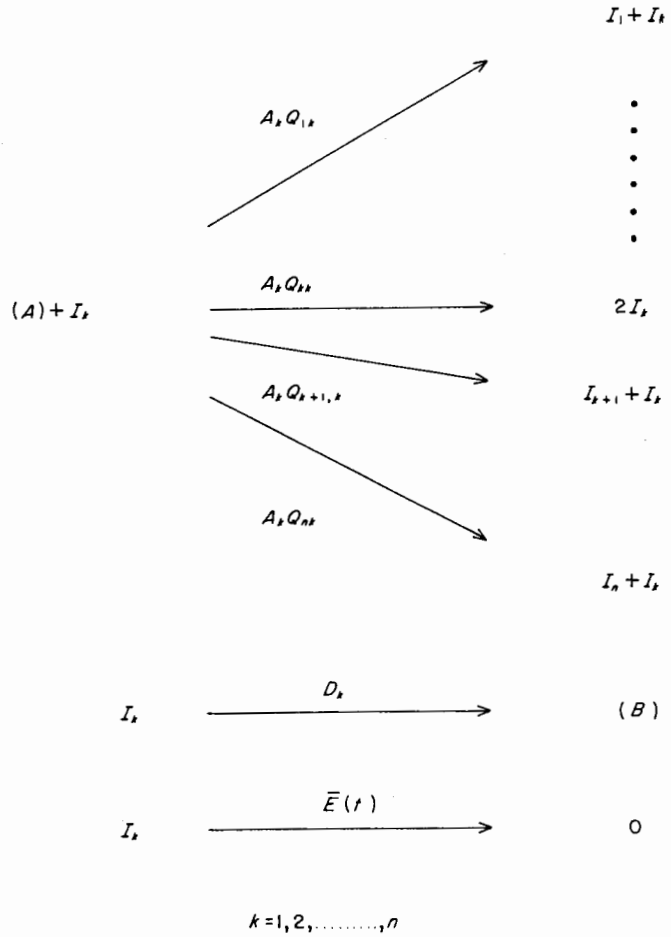

FIG. 1. Replication and mutation as parallel reactions of a uniform reaction mechanism. Replication rate constants are denoted by A_k, $k = 0, 1, \ldots, n$, degradation rate constants by D_k, $k = 0, 1, \ldots, n$. The elements of the mutation matrix $Q = \{Q_{ik}\}$ give the frequencies of mutations: Q_{ik} is the probability to obtain I_i as an error-copy of I_k. Accordingly, we have $\sum_{i=0}^{n} Q_{ik} = 1$. The compounds in parentheses, (A) and (B), do not enter the kinetic equations explicitly: (A) represents schematically the energy rich monomers of polynucleotide synthesis and its concentration is assumed constant, (B) is the degradation product of polynucleotide hydrolysis. The mean excess production, $\bar{E}(t) = \sum_{k=0}^{n} (A_k - D_k) x_k / \sum_{k=0}^{n} x_k$, is compensated by the (unspecific) dilution flux $\Phi = \bar{E}(t)$ shown in the last reaction.

the quasispecies converges towards a homogeneous population consisting exclusively of the fittest type in the limit of no errors, $\lim p \to 0$ or $\lim q \to 1$, respectively. With increasing error rates the mutant distribution spreads around this type as its center. Still the consensus sequence of the distribution coincides with the master sequence—with the exception of more complicated cases with selectively neutral or almost neutral types. At a sharply defined critical minimum single digit accuracy $q = q_{\min}$, however, the population loses stationarity. The quasispecies is no longer

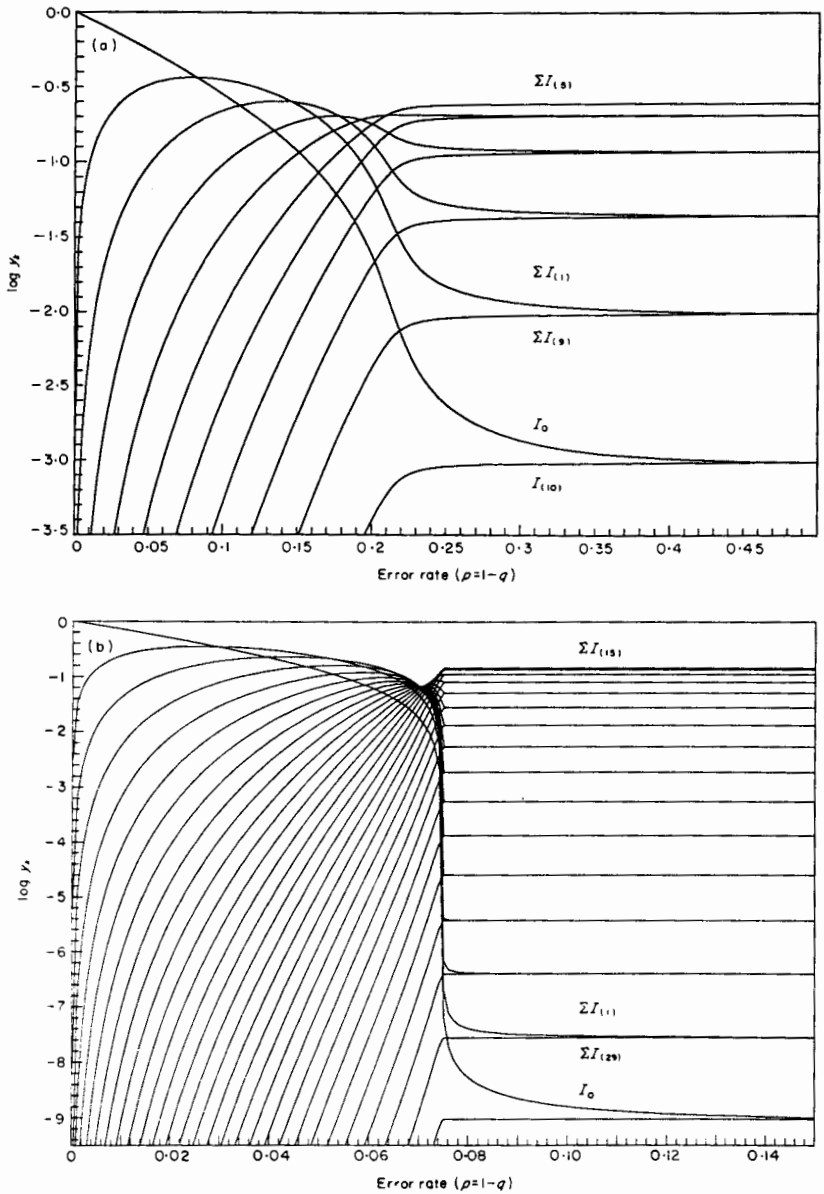

FIG. 2. The stationary distribution of master sequence and mutants, called the *quasispecies* as a function of the error rate $p = 1 - q$. We plot relative concentrations of error classes, y_k, $k = 0, 1, \ldots, \nu$; y_0 is the relative concentration of the master sequence I_0, y_1 the sum of concentration of all sequences in the one error class, $I_{(1)}$, i.e. all sequences with Hamming distance $d(0, k) = 1$ from the master sequence, and y_2, y_3 etc. defined analogously. Both computation were performed for a *single peak value landscape*: $A_0 = 10$ and $A_k = 1 \forall k \neq 0$. Chain length $\nu = 10$ was used in the upper plot, $\nu = 30$ in the lower diagram. Note, that the range of transition from the organized quasispecies to the uniform distribution, the *error threshold*, is very sharp at chain length $\nu = 30$ already.

localized and starts drifting randomly through sequence space (Fig. 2). The transition from a stationary quasispecies to a migrating ensemble of sequences was called the *error threshold.* Beyond error threshold the former master sequence will inevitably be lost. Then, second and third best sequences will dissappear until the fitness of the population degenerates to some average value.

Quasispecies delocalization is reminiscent of *Muller's ratchet* which describes degeneration of asexually multiplying populations through sequential loss of the fittest alleles. (For a discussion of *Muller's ratchet* see, e.g. Maynard Smith, 1978). There are, however, substantial differences: *Muller's ratchet* operates in small populations and is based on the neglect of back mutations from less efficient mutants to the wildtype. Error thresholds are effective in infinite populations too. Migration in sequence space, on the other hand, means that all sequences have a finite lifetime (Demetrius *et al.*, 1985). Thus, the metaphor of a ratchet with a single notch applies: at the threshold it ceases to click.

Error thresholds become very sharp with increasing chain lengths (Swetina & Schuster, 1982) and bear obvious resemblance to co-operative transitions. Indeed, Demetrius (1987) and Leuthäusser (1987) have shown that replication dynamics can be analyzed within the same general mathematical frame as spin statistics in the solid state. The error threshold in the limit of infinite chain lengths, $\nu \to \infty$, is equivalent to an order-disorder transition.

The biological relevance of the quasispecies concept has been established experimentally. Natural populations of RNA viruses were found to be highly heterogeneous (Domingo *et al.*, 1988). They consist of many individual types which are related by mutation. Measured (mean) single digit accuracies ($\bar{q}$) and genome lengths (ν) were compared (Eigen & Biebricher, 1988). Most of the viruses examined—examples are the Coliphage $Q\beta$, Vesicular stomatitis, Influenza *A* or Foot-and-mouth disease virus—were found to replicate under conditions close to the error thresholds. Then, mutant distributions and populations dynamics are described properly in terms of quasispecies only. Not enough data are presently available for bacterial populations and quantitative estimates on mutant distribution cannot be made yet. Presumably they are to be described in terms of quasispecies, too.

2. Diagnoses of Error Thresholds in Infinite Populations

Error thresholds were first discussed by means of the kinetic equations corresponding to the mechanism shown in Fig. 1. They refer—in a strict sense—to infinite populations only. There is no migration or random drift in infinite ensembles. Instead, delocalization leads to spreading over the entire sequence space. Ultimately, at very large error rates—$q \ll q_{\min}$—the infinite population approaches the uniform sequence distribution.

The reaction mechanism shown in Fig. 1 is properly modelled by the differential equation

$$\frac{dx_k}{dt} = \sum_{j=0}^{n} W_{kj} x_j - x_k \bar{E}(t), \qquad k = 0, \ldots, n, \tag{1}$$

which is essentially linear since the term containing the mean excess production

$$\bar{E}(t) = \sum_{k=0}^{n} (A_k - D_k)x_k \Big/ \sum_{k=1}^{n} x_k \tag{2}$$

can be removed by a non-linear transformation. The variables in eqn (1) are relative concentrations—$[I_k] = c_k$:

$$x_k = \frac{c_k}{\sum_{j=0}^{n} c_j} \quad \text{and} \quad \sum_{k=0}^{n} x_k = 1. \tag{3}$$

The coefficients W_{kj} are elements of a value matrix W which contains rate constants and mutation frequencies—δ_{kj} is Kronecker's delta symbol:

$$W_{kj} = A_j \,.\, Q_{kj} - D_k \,.\, \delta_{kj}. \tag{4}$$

Replication and degradation rate constants are denoted by A_k and D_k respectively. The mutation frequencies are given by

$$Q_{kj} = q^{\nu} \left(\frac{1-q}{q}\right)^{d(k,j)} \tag{5}$$

where $d(k, j)$ stands for the Hamming distance between the two sequences I_k and I_j. The Hamming distance of two sequences counts the number of digits in which they differ.

The quasispecies is obtained as the dominant eigenvector, ξ_0—corresponding to the largest eigenvalue λ_0—of the value matrix W. The dominant eigenvector describes the stationary mutant distribution of the replication-mutation system: $\xi_0 = (\bar{x}_0, \bar{x}_1, \ldots, \bar{x}_n)$. This distribution is commonly centered around a most efficient variant, called the *master sequence* I_0. Application of perturbation theory in lowest order to compute ξ_0 yields a critical minimum replication accuracy $q_{\min}$ which is defined by a vanishing stationary concentration of the master sequence, $\bar{x}_0(q_{\min}) = 0$:

$$Q_{\min} = q_{\min}^{\nu} = \sigma^{-1} \tag{6}$$

wherein we made use of a phenomenological superiority parameter

$$\sigma = \frac{A_0}{D_0 + \bar{E}_{k\neq 0}} \tag{7}$$

with a mean excess production of all sequences except the master

$$\bar{E}_{k\neq 0} = \sum_{k=1}^{n} (A_k - D_k)x_k \Big/ (1 - x_0).$$

The eigenvalue spectrum of W—as shown in Fig. 3—provides a useful tool to detect error thresholds: in the neighborhood of the threshold we observe a series of avoided crossings. The longer the chain, the narrower are the gaps between the individual crossings and, the closer approach each other the two eigenvalues. The rationale for this observation lies in the decrease of coupling terms with increasing chain lengths (Swetina & Schuster, 1982). The first avoided crossing—counted in

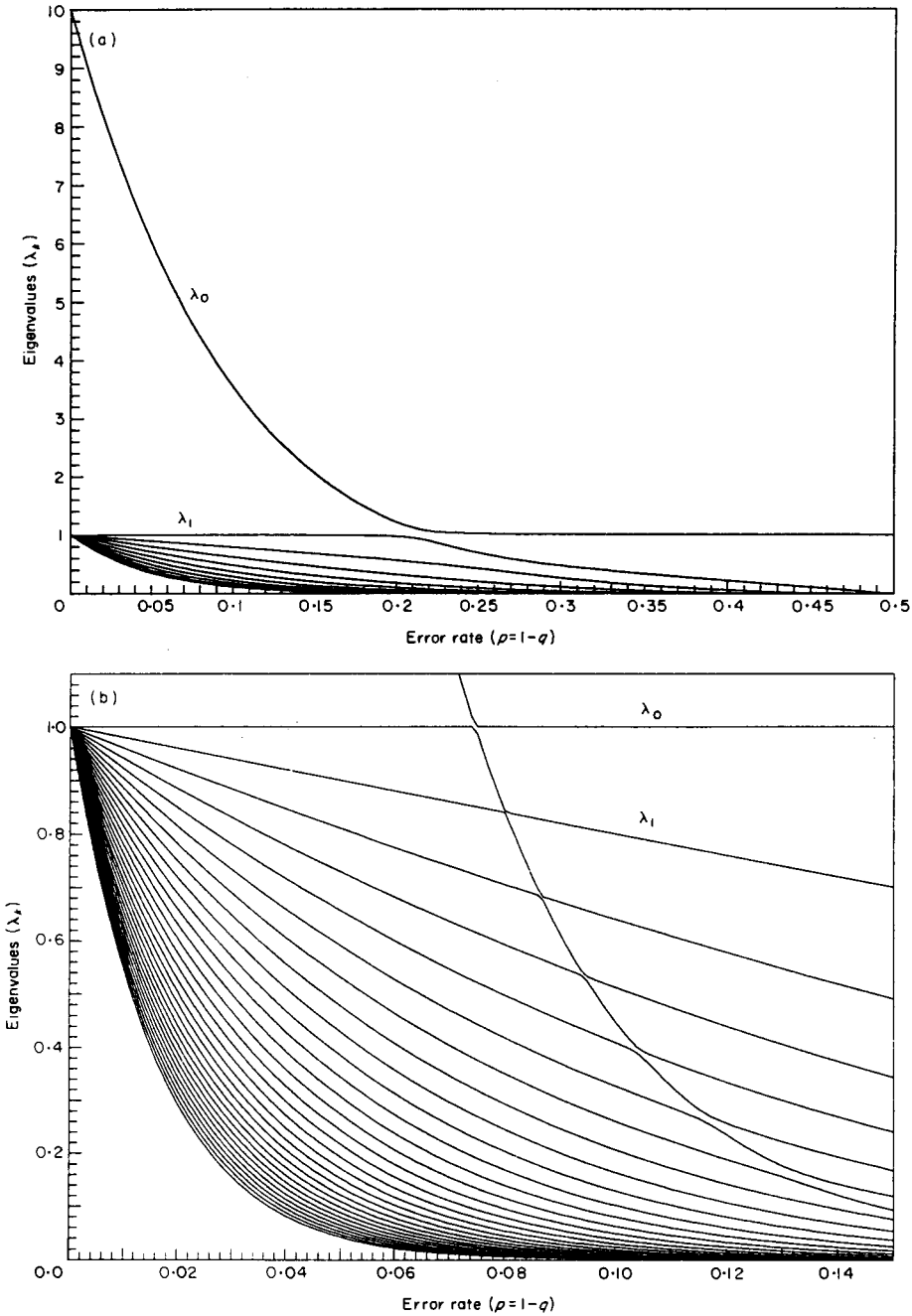

FIG. 3. The spectrum of eigenvalues (λ_k) of the value matrix W as a function of the error rate $p = 1 - q$. Value landscapes and chain lengths ($\nu = 10$ and 30) are the same as in Fig. 2. We observe *avoided crossing* of the two largest eigenvalues (λ_0 and λ_1) at the error threshold.

the direction of increasing error rates—is most important for an understanding of the nature of the error threshold since it involves the largest eigenvalue and the dominant eigenvector: in this narrow critical range an organized mutant distribution, which is centered around a most efficient master sequence, changes into the uniform distribution.

3. A Computer Model of Replication as a Stochastic Process

In order to discuss properties of small systems with respect to population size we have to leave the conventional—deterministic—kinetic differential equations and consider replication, mutation and degradation as stochastic processes. Although several attempts along this line have been made already none of them was really satisfactory. We mention here only one recent approach (Demetrius *et al.*, 1985) which is close to the numerical simulations reported in this contribution. Individual replication and mutation reactions are formulated as a multitype branching process. The major result of this analysis concerns the proof of the existence of a stochastic analogue to the deterministic error threshold. This *stochastic error threshold* is characterized by a vanishing probability of survival to infinite time of the master sequence. If replication is more accurate than the critical minimum value ($q > q_{\min}$) then the master sequence—and together with it the whole quasispecies—has a finite probability to survive to infinite time. The transition from certain extinction to a probability of survival close to one becomes very sharp with increasing chain lengths ν (Pichler & Schuster, in press).

The basic drawback of the multitype branching approach is to be seen in the difficulty to introduce a constraint which can be used to control population size. All results derived so far were obtained—and are valid therefore only—for unconstrained, and therefore randomly drifting population sizes.

Since we are essentially interested in small populations we may also use an alternative approach which is based on numerical simulations (Nowak, 1987). Gillespie (1976) suggested a simple and efficient algorithm which can be applied straightaway to chemical reaction networks. Particle numbers are considered as stochastic variables $X_k(t)$ with $k = 0, 1, \ldots, 2^\nu - 1$. The variables change stepwise by integers when elementary reactions occur at some time τ: $X_k(\tau + \delta) = X_k(\tau - \delta) \pm n$ with $n = 0, 1, 2, \ldots$, depending on the nature of the reaction step. The population size is given by

$$N(t) = \sum_{k=0}^{2^\nu - 1} X_k(t) \tag{8}$$

and may be either fluctuating or constant depending on the particular constraint applied.

The Gillespie algorithm simulates single trajectories which are understood as sequences of elementary reaction steps. The occurence of individual elementary reactions is controlled by two *Monte Carlo* steps. The first random number decides when the next reaction occurs, and the second random number chooses the particular kind of reaction step which is taking place. Probabilities of occurence for the

individual reactions are proportional to their rates which—as in conventional, deterministic reaction kinetics—are given by products of rate constants and particle numbers according to mass action kinetics.

All simulations approached soon a stationary state which represents the quasi-species. After the individual simulations had reached stationarity they were extended for further 20 000 time units. During the stationary period particle numbers, mean excess productions and mean Hamming distances were recorded every 100 time units. Expectation values and standard deviations of these quantities were computed from the samples taken. This procedure was repeated for every q-value. The computation time required for a population size of $N = 500$ and one q-value was about 1000 CPU sec on an IBM 3081 computer.

Error thresholds are easily detected on very simple fitness landscapes already. Therefore we shall apply here the most simple, conceivable model (Swetina & Schuster, 1982), which we characterize as *single peak* value landscape. Only two different replication rate constants are required: one for the master, A_0 and one for all other sequences which can be set $A_1 = \ldots = A_{2^\nu - 1} = A = 1$ without losing generality. The simple fitness landscape allows straight sampling of sequences into error classes—as it was applied by Swetina & Schuster:

$$Y_k(t) = \sum_{j \in \Gamma_k} X_j(t), \qquad k = 0, 1, \ldots, \nu. \tag{9}$$

Here Γ_k represent the *k-error class* of the master sequence comprising all sequences with Hamming distance $d(0, j) = k$ from the master. Then, the number of stochastic variables is sufficiently small to permit the application of conventional techniques to compute expectation values and variances by sampling of individual trajectories. The mutation matrix Q defined in eqn (5) has to be modified in order to account for all transitions between the individual error classes. For mutations from class Γ_l into class Γ_k we find

$$Q_{kl} = \sum_{i=l-\nu-k}^{\min(k,l)} \binom{k}{i} \binom{\nu-k}{l-i} q^\nu \left(\frac{1-q}{q}\right)^{k+l-2i}. \tag{5a}$$

In contrast to the conventional symmetrical mutation matrix (5) transitions between error classes lead to the asymmetrical matrix (5a).

The results obtained by Gillespie simulation resemble qualitatively those derived by integration of the corresponding kinetic differential equations. We are essentially interested in error thresholds which can be detected from stationary sequence distributions of the replicating ensembles. Therefore the computer simulations had to be carried on for sufficiently long times.

4. Diagnoses of Error Thresholds in Finite Populations

In order to detect error thresholds in stochastic simulations we use three quantities which depend characteristically on the error rate p:

(i) the stationary values of particle numbers

$$\bar{Y}_k(p) = \lim_{\tau \to \infty} \frac{1}{\tau} \int_0^\tau Y_k(t, p)\, \mathrm{d}t, \qquad k = 0, 1, \ldots, \nu, \tag{10}$$

(ii) the stationary value of the mean Hamming distance between the master sequence and the whole population

$$\bar{d}_0(p) = \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau d_0(t,p)\,\mathrm{d}t \tag{11}$$

with

$$d_0(t,p) = \frac{1}{N} \sum_{k=0}^{\nu} k.\, Y_k(t,p), \quad \text{and} \tag{12}$$

(iii) the stationary value of the mean excess production

$$\bar{E}(p) = \lim_{\tau\to\infty} \frac{1}{\tau} \int_0^\tau E(t,p)\,\mathrm{d}t, \tag{13}$$

with

$$E(t,p) = \frac{1}{N} \sum_{k=0}^{\nu} A_k Y_k(t,p) = \frac{1}{N} \{A_0 Y_0(t,p) + A[N - Y_0(t,p)]\}$$

$$= A + \frac{1}{N} (A_0 - A)\, Y_0(t,p). \tag{14}$$

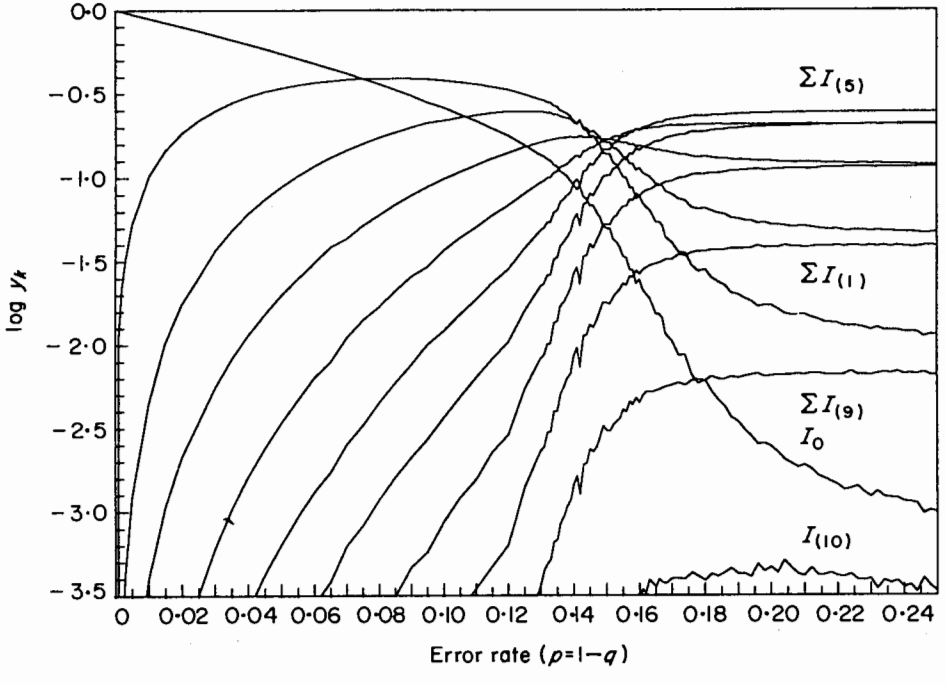

FIG. 4. The error threshold in finite populations. The plot is analogous to the upper diagram in Fig. 2 for a population size $N = 100$. The individual curves represent mean values of the relative particle numbers obtained as *long time* results of computer simulations. The curves are very similar to those shown in Fig. 2 but the error threshold is shifted towards lower error rates. Note the scatter of the results in the range of the error threshold.

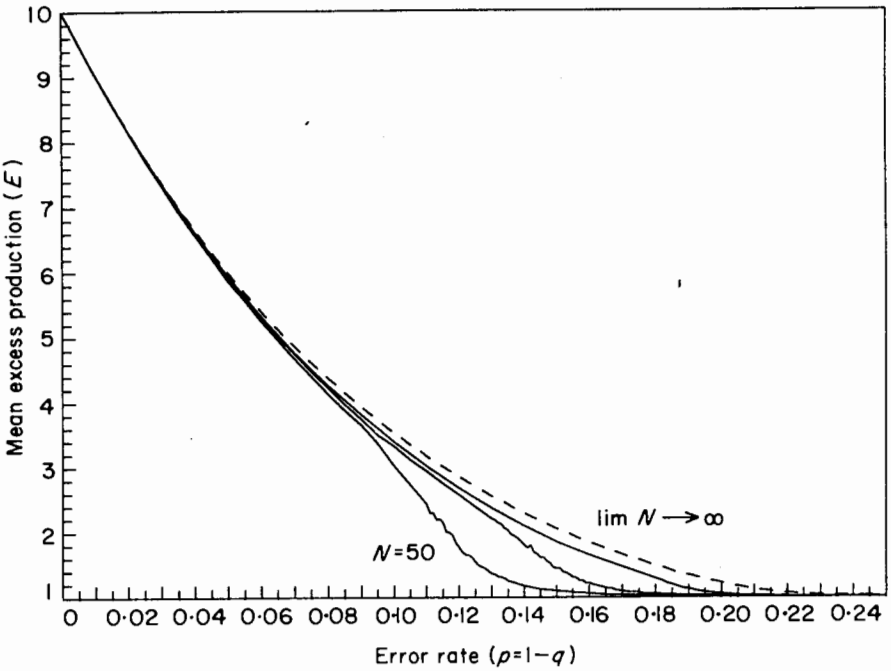

FIG. 5. The stationary mean excess production [$\lim_{t\to\infty} \bar{E}(t)$] as a function of the error rate $p = 1 - q$. Computation were performed for a *single peak* value landscape with $A_0 = 10$ and a chain length $\nu = 10$. Curves are given for $N = 50, 10, 500$ and for the deterministic case, lim $N \to \infty$ (broken curve).

Since no integration in reality can be extended to infinite time, it matters whether the simulation is carried on for longer times or whether several trajectories are sampled. The latter was found to be more efficient in this particular case. Characteristic plots of the three quantities proposed for the detection of error thresholds in finite populations—relative particle numbers, mean excess production and mean Hamming distance—are presented in Figs 4-6. All three quantities show appreciable scatter of computed points in the neighborhood of the error thresholds.

5. A Birth and Death Model of Error Thresholds

The numerical data obtained by computer simulations were put on a firm basis by means of an analytical study. In order to derive an expression for error thresholds in finite populations we use a *birth* and *death* process as a stochastic model which comes as close as possible to the reaction mechanism of Fig. 1. Simplifications are inevitable; we restrict the model therefore to two types only, which live on the *single peak landscape*. One type is the master sequence I_0, the second type—called the *error tail*, I_E—comprises all other sequences [for the definitions of variables see eqn (9)]:

$$I_0: \quad Z_0 = Y_0 = X_0 \tag{15}$$

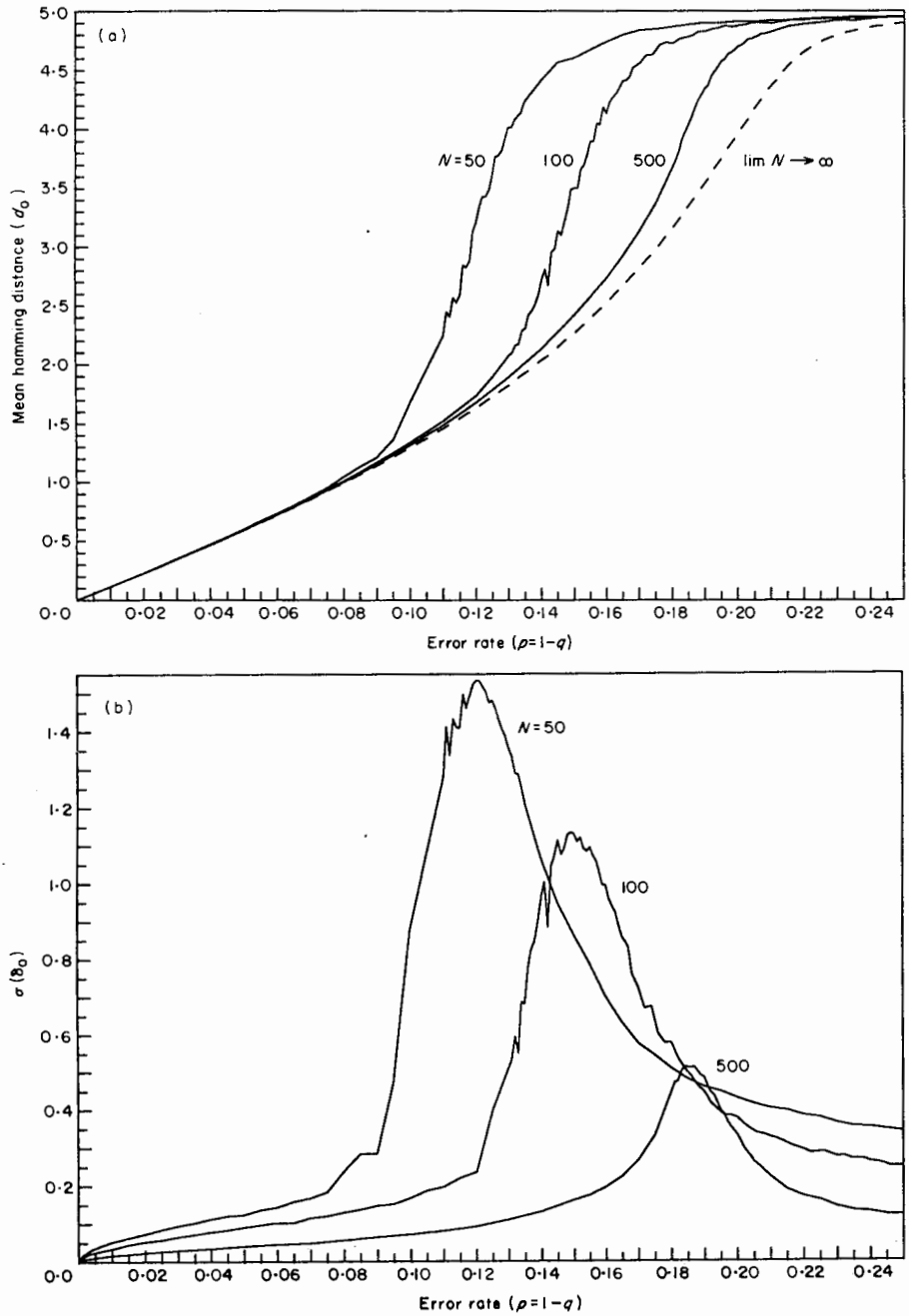

FIG. 6. Mean Hamming distances (d_0) in stationary mutant distributions as functions of the error rate $p = 1 - q$. The same parameters as in Fig. 5 were used in the computations. The lower plot presents standard deviations, $\sigma(d_0)$. Note, that the increase in mean Hamming distance—indicating spreading of populations—is steepest and the standard deviation is largest around the error threshold.

$$I_E: \quad Z_E = \sum_{k=1}^{\nu} Y_k = \sum_{j=1}^{2^\nu - 1} X_j. \tag{16}$$

The *birth* and *death* model applied here is originally due to Moran (1958) who worked in formal population genetics. The constraint of constant total particle number N is realized in the following way: two molecules are chosen at random, one enters the replication-mutation mechanism and the other one is eliminated. The two molecules need not be different—the same molecule may be replicated and eliminated after replication. Accordingly, we are dealing with a stochastic process in a single, discrete variable: $Z = Z_0 = 0, 1, \ldots, N$. The time is assumed to be continuous. Then, the transition probabilities $P_{i,i\pm1}$ of the infinitesimal generator of the *birth* and *death* process can be written as (Karlin & Taylor, 1975):

$$\lambda_i = P_{i,i+1} = \frac{N-i}{N}\left(W_{MM}\frac{i}{N} + W_{ME}\frac{N-i}{N}\right) \tag{17}$$

$$\mu_i = P_{i,i-1} = \frac{i}{N}\left(W_{EM}\frac{i}{N} + W_{EE}\frac{N-i}{N}\right). \tag{18}$$

Now we introduce the molecular mutation mechanism into the *birth* and *death* model. Two elements of the reduced value matrix W are obtained straightway:

$$W_{MM} = A_0 \cdot q^\nu, \tag{17a}$$

$$W_{EM} = A_0 \cdot (1 - q^\nu). \tag{18a}$$

The other two elements are more tricky and cannot be computed exactly without a knowledge of the numbers of sequences in the individual error classes (Y_k). The simplest approximation possible assumes uniform distribution of all sequences in the error tail:

$$W_{ME} = \sum_{j=1}^{\nu} y_j q^{\nu-j}(1-q)^j = \sum_{j=1}^{\nu} \frac{\binom{\nu}{j}}{2^\nu - 1} q^{\nu-j}(1-q)^j = \frac{1-q^\nu}{2^\nu - 1} \tag{17b}$$

$$W_{EE} = 1 - W_{ME}. \tag{18b}$$

Here and in Fig. 4 we use normalized—or *relative*—stochastic variables for the error classes in the error tail:

$$y_j = Y_j \Big/ \sum_{k=1}^{\nu} Y_k; \qquad j = 1, \ldots, n.$$

Following Karlin & Taylor (1975) we obtain the stationary probability distribution $p_k (k = 0, \ldots, N)$:

$$\pi_k = \frac{\lambda_{k-1}}{\mu_k}\pi_{k-1}; \qquad k = 1, \ldots, N \quad \text{with} \quad \pi_0 = 1, \tag{19}$$

$$p_k = \frac{\pi_k}{\sum_{i=0}^{N} \pi_i}; \qquad k = 0, \ldots, N. \tag{20}$$

This recursion is easily computed numerically.

In order to test the reliability of the *birth* and *death* model we computed the solutions in the limit $N \to \infty$ and compared them with the exact solution curves of the differential eqn (1). Because of the assumption of a uniform distribution of mutants the model underestimates the stationary concentrations of the master sequence:

$$\bar{x}_0 \geq \lim_{N\to\infty} \frac{\bar{Z}_0}{N}.$$

The error is largest around the error threshold. Nevertheless, the approximation turned out to be sufficiently accurate, and highly useful for our purpose. In order to analyze the stationary probability of the *birth* and *death* process we approximate the discrete distributions by continuous functions. The difference of birth and death rates is then expressed by

$$P_{i-1,i} - P_{i,i-1} = \lambda_{i-1} - \mu_i \to \tilde{\lambda}(x) - \tilde{\mu}(x) = \zeta(x)$$

and

$$\zeta(x) = (1 - A_0)x^2 + (\gamma_3 Q + \gamma_2)x + \gamma_1 Q + \gamma_0, \tag{21}$$

where x replaces i/N, $\beta = 1/N$, $\kappa = (2^\nu - 1)^{-1}$, $Q = q^\nu$, and

$$\gamma_0 = \kappa(1+\beta)^2$$

$$\gamma_1 = -(1+\beta)[\kappa(1+\beta) + A_0\beta]$$

$$\gamma_2 = -\kappa(1+2\beta) - 1$$

$$\gamma_3 = (A_0 + \kappa)(1+2\beta).$$

The curvature of eqn (21) is negative everywhere—$\zeta''(x) = 2(1 - A_0) < 0$. Depending on the choice of the single digit accuracy q, eqn (21) has two roots, one root or no root in the range of interest: $0 \leq x \leq 1$. Extreme values of the stationary probability distribution p_i according to eqn (20) lie around the points which fulfil the equation $\zeta(\bar{x}) = 0$.

(i) For $q = 1$ we have two absorbing states at $x = 0$ and $x = 1$. The root near $\zeta = 0$ has no physical meaning since the probability distribution vanishes everywhere between the two δ-functions.

(ii) At very small error rates, $1 - q = p = \delta$, the stationary distribution has two peaks at $x = 0$ and $x = 1$ and the root of the quadratic equation corresponds to the minimum.

(iii) At higher error rates the distribution has a maximum near $x = 1$ and a minimum near $x = 0$ corresponding to two roots of the quadratic equation.

(iv) There exists a critical replication accuracy q_{cr} at which the maximum and the minimum coalesce.

(v) Finally, at values $q < q_{cr}$ we find no root of the quadratic equation and the distribution p_i decreases monotonously from $x = 0$ to $x = 1$.

In order to derive a criterion for the error threshold we assume first $N \ll 2^\nu$, a condition which is fulfilled for all realistic population sizes already at small chain lengths ν. We recall that in this case populations are drifting through sequence space if the replication accuracy is below the critical value. Then the master sequence—like any other particular sequence—is bound to vanish and hence, zero particle number ($i=0$) is the most probable state. The probability distribution is expected to decrease monotonously with increasing i—or x, respectively—in the entire domain:

$$p'(x) \leq 0 \quad \text{and} \quad \zeta(x) \leq 0 \quad \forall x: 0 \leq x \leq 1.$$

This condition is fulfilled in case (v) and it is straightway to identify the critical accuracy q_{cr} defined in (iv) with the error threshold. Indeed, below the critical accuracy the death rate of the master sequence, $\tilde{\mu}(x)$, is larger than the birth rate $\tilde{\lambda}(x)$ in the whole range of physically meaningful x-values: $0 \leq x \leq 1$.

The condition of coincidence of the two roots of eqn (21) is given by

$$\gamma_3^2 Q_{cr}^2 + [2\gamma_2\gamma_3 - 4(A_0-1)\gamma_1]Q_{cr} + \gamma_2^2 - 4(A_0-1)\gamma_0 = 0,$$

where $Q_{cr} = q_{cr}^\nu$. The error threshold ($q_{\min}$) is given by the larger root of this quadratic equation. A fairly simple analytic equation for the population size dependence of $q_{\min}$ is obtained if we put $\kappa = 0$:

$$q_{\min}^\nu = \frac{1}{A_0(N+2)^2}\left\{N(N+2) + 2(A_0-1)(N+1)\left[1+\sqrt{1+\frac{N(N+2)}{(A_0-1)(N+1)}}\right]\right\} \quad \text{with} \quad N < 2^\nu. \tag{22}$$

In order to discuss the limit of large populations ($N \to \infty$) we have to consider also the biologically, and also physically insignificant case in which the population size is larger than the possible number of binary sequences: $N > 2^\nu$. We cannot assume then, that the zero state ($i=0$) is the most probable state below the minimum replication accuracy $q_{\min}$. In contrary, the condition $p_0 < p_1$ is fulfilled for all q-values. The criterion for the error threshold—which was coincidence of the two roots of eqn (21)—is obsolete for large populations with $N > 2^\nu + \delta$ with $\delta > (2^\nu - A_0)/(A_0-1)$. Instead, we choose the minimum distance of both roots as criterion of the error threshold and find:

$$q_{\min}^\nu = \frac{2(1-A_0)\gamma_1 - \gamma_2\gamma_3}{\gamma_3^2}; \qquad N > 2^\nu + \delta. \tag{23}$$

Both critera yield the same values in numerical computations at the critical population size $N_{cr} = 2^\nu + \delta$. The slopes of the two curves $q_{\min}(N)$ for $N < N_{cr}$ and $N > N_{cr}$ are different (see Fig. 7).

It is interesting to note that the single digit accuracy $q_{\min}$ obtained from the *birth* and *death* model by eqn (23) in the limit $N \to \infty$ coincides with the q-value at which

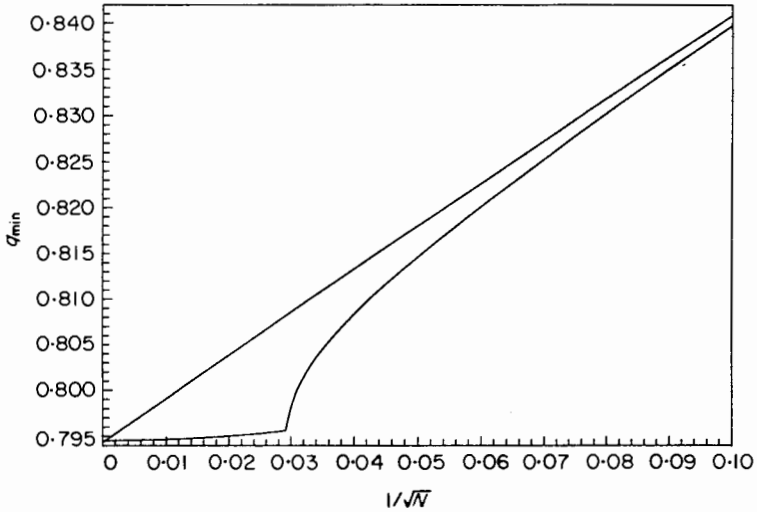

FIG. 7. The error threshold as a function of the reciprocal square root of the population size ($1/\sqrt{N}$). The upper (almost) straight line corresponds to the approximation $\kappa = 0$. The lower curve uses the correct value $\kappa = (2^\nu - 1)^{-1}$ and consists of two parts which are joined in one point with discontinuous first derivative. The right hand side represents critical coincidence of the two roots of eqn (21)—as it was used to draw the curve for $\kappa = 0$—whereas the curve in the left hand part is defined by minimum distance of both roots. Parameters: $\nu = 10$ and $A_0 = 10$.

the difference of two eigenvalues of the value matrix (17a, b–18a, b)

$$\begin{bmatrix} A_0 q^\nu & \kappa(1-q^\nu) \\ A_0(1-q^\nu) & 1-\kappa(1-q^\nu) \end{bmatrix}$$

has the minimum value:

$$[q|_{(|\lambda_0-\lambda_1|=\min)}]^\nu = \frac{A_0 + 3A_0\kappa + \kappa^2 - \kappa}{(A_0+\kappa)^2}. \tag{24}$$

The critical q_{min}-value determined in this way is always larger than the conventional threshold value

$$q_{min}^\nu = \frac{1}{A_0}$$

obtained from perturbation theory in lowest order (Eigen & Schuster, 1977). This value is obtained for $\kappa = 0$ which corresponds to the limit of large chain lengths ($\lim \nu \to \infty$) in which both values become identical. The difference between the two critical q_{min} values, however, is very small for short chains already.

In Table 1 we compare results obtained within the frame of the various approximations made in the *birth* and *death* model with the accurate numerical values. The critical q_{min}-values agree well with the error thresholds determined in Section 4, in particular, if one takes into account the principal differences between the replication-mutation model and the *birth* and *death* process.

TABLE 1

Comparison of error thresholds (q_{min}) obtained from computer simulation and from the birth and death model. Chain length: $\nu = 10$, Superiority: $\sigma = A_0 = 10$

Population size N	Computer simulation†		*Birth* and *death* model	
	Fluctuating N	Constant N	$\kappa \neq 0$	$\kappa = 0$
100	0·851	0·847	0·840	0·841
150	0·839	0·837	0·831	0·833
200	0·831	0·828	0·826	0·828
300	0·822	0·821	0·819	0·822
500	0·816	0·817	0·811	0·816
1000	0·810	0·807	0·801	0·809
2000	0·804	—	—	0·805
3000	0·803	—	—	0·803

† The computer simulations were carried out under two different constraints:
(i) particle numbers fluctuating like in a flow reactor with constant mean [$N = \lim_{t\to\infty} t^{-1} \int_0^t N(\tau)\, d\tau$] and standard mean deviation in the order of ($\sqrt{N}$), and
(ii) constant partical numbers corresponding to the constraint applied in the *birth* and *death* process.

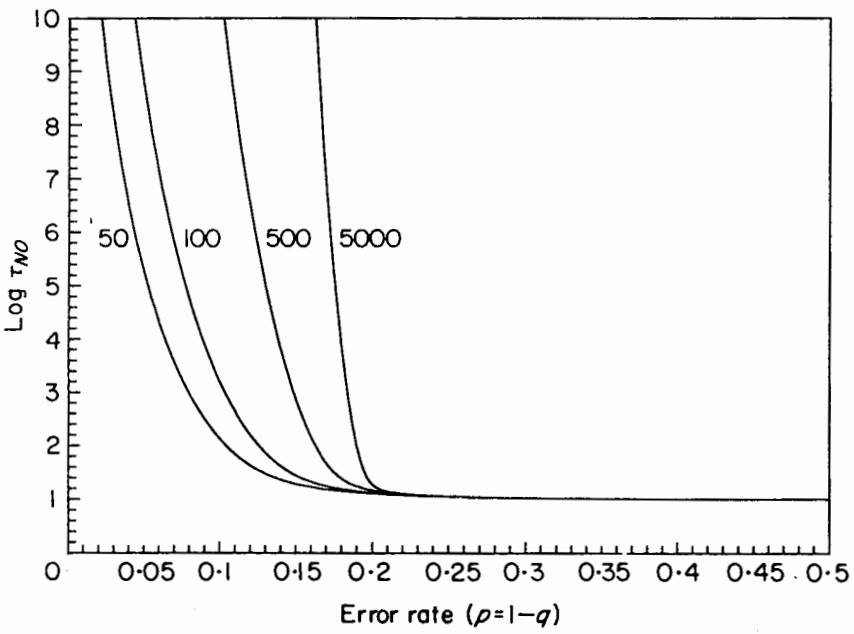

FIG. 8. First passage time τ_{N0} as a function of the error rate $p = 1 - q$. This time represents the average time which is required to lose the master sequence in an initially homogeneous population of N master sequences. We realize a sharp increase of the first passage time with decreasing error rates which starts off at the error threshold. The larger the population size (N) is, the sharper is the transition from steep decrease to an almost constant value of τ_{N0}. The curves $\tau_{N0}(p)$ in the plot were scaled such that $\log \tau_{N0}(0{\cdot}5) = 1$. Parameters: $\nu = 10$, $A_0 = 10$ and $N = 50$, 100, 500 and 5000.

In addition to the stationary probability distributions we compute quantities which characterize the degree of localization of mutant distributions. We use mean first passage times from state $Z = i$ to state $Z = 0$, τ_{i0}—these are the mean values of time the systems require to reach for the first time $Z = 0$ after having started at time $t = 0$ from $Z = i$ (Hunter, 1983). Mean first passage times fulfil the equations

$$\tau_{i0} = \frac{1}{\lambda_i + \mu_i}(1 + \lambda_i \tau_{i+1,0} + \mu_i \tau_{i-1,0}); \qquad i = 1, \ldots, N-1. \tag{25}$$

They can be computed by recursion after the substitution $\Delta_i = \tau_{i,0} - \tau_{i-1,0}$:

$$\Delta_i = \frac{1 + \lambda_i \Delta_{i+1}}{\mu_i} \tag{26}$$

with $\Delta_N = 1/\mu_N$ and $\Delta_1 = \tau_{10}$.

In Fig. 8 we show plots of mean first passage times τ_{N0} as functions of the error rate $p = 1 - q$. We observe a steep decrease in $\tau(p)$ in the neighborhood of the error threshold: the master sequence is lost and the mutant distribution starts to drift randomly through sequence space.

6. Discussion of Results

Comparison of the results obtained by numerical simulation and from the *birth* and *death* model reveals, in essence, two major results:

(1) The concept of the error threshold has been put on a new formal basis which allows to diagnose the occurrence of this phenomenon also in finite populations. At error rates above the critical value the quasispecies ceases to be localized in sequence space and starts to drift randomly. This interpretation of the results derived from probability distributions is supported strongly by the behavior of mean first passage times around the critical q-value.

(2) From the critical single digit accuracy of the *birth* and *death* model eqn (22) we derive an approximate analytical expression for the population size dependence of the error threshold:

$$\begin{aligned}[q_{\min}(N)]^{\nu} &\approx \frac{1}{A_0}\left[1 + \frac{2\alpha^2}{N}\left(1 + \frac{\sqrt{N}}{\alpha}\sqrt{1 + \frac{\alpha^2}{N}}\right)\right] \\ &= \frac{1}{A_0}\left[1 + \frac{2\alpha^2}{N} + \frac{2\alpha}{\sqrt{N}}\left(1 + \frac{\alpha^2}{2N} + \ldots\right)\right] \\ &= \frac{1}{A_0}\left[1 + \frac{2\alpha}{\sqrt{N}} + \frac{2\alpha^2}{N} + \frac{\alpha^3}{(\sqrt{N})^3} + \ldots\right]\end{aligned} \tag{27}$$

with $\alpha = \sqrt{A_0 - 1}$. The approximation is valid in the limit of large population sizes, $N \gg 1, 2, \ldots, A_0$. This condition is fulfilled in all populations of practical interest, because superiorities σ commonly lie just above 1, and very rarely exceed values of ten. Note, that $\sigma = A_0$ holds on the single peak value landscape.

(3) In the limit of large population sizes, $N \to \infty$, the simple *birth* and *death* model showed in addition that the single digit accuracy q, at which the difference of the two largest eigenvalues of the value matrix W assumes its minimum value ($\min |\lambda_0 - \lambda_1|$), is analytically identical to the q-value at which the critical change in the probability distribution of the master sequence's particle number occurs. This critical change is tantamount to quasispecies delocalization.

Within the limits of the phenomenological approach to the error threshold relation (Eigen *et al.*, 1988) eqn (27) can be extended to general value landscapes if we make use of the conventional superiority parameter σ according to eqn (6). Then we find

$$q_{\min}(N) = q_{\min}(\infty)\sqrt[\nu]{1+\frac{2\alpha}{\sqrt{N}}+\frac{2\alpha^2}{N}+\frac{\alpha^3}{(\sqrt{N})^3}+\dots}$$

$$\approx q_{\min}(\infty)\left[1+\frac{2\alpha}{\nu\sqrt{N}}+\frac{2\alpha^2}{\nu N}+\frac{\alpha^3}{\nu(\sqrt{N})^3}+\dots\right], \tag{27a}$$

where the approximation holds for sufficiently large N, $q_{\min}(\infty) = \sigma^{-1}$ and $\alpha = \sqrt{\sigma - 1}$.

According to eqn (27a), plots of $q_{\min}$ against $1/\sqrt{N}$ are expected to yield straight lines in the limit of small values of $1/\sqrt{N}$, or by the same token for $\lim N \to \infty$. At larger values of $1/\sqrt{N}$ we expect a deviation from the straight line towards higher values of $q_{\min}$ since the $1/N$ term of the expansion becomes dominant. Figure 9 shows such a plot together with data from the numerical simulations reported in Section 4. In sufficiently large populations we observe excellent agreement between

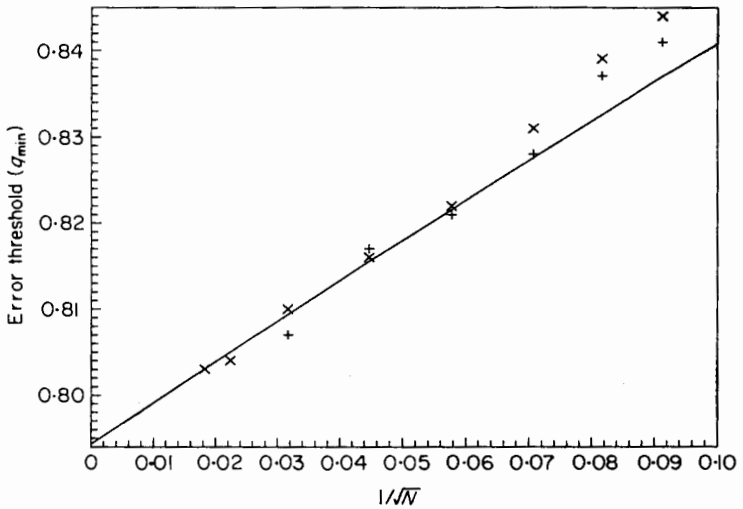

FIG. 9. The error threshold as a function of the reciprocal square root of the population size ($1/\sqrt{N}$). The full line was computed from eqn (22). The individual points were obtained from computer simulations using a constraint with fluctuating (×) and constant (+) total particle numbers N. Parameters: $\nu = 10$ and $A_0 = 10$.

eqn (27a) and the computer simulations. Deviations at small population sizes occur partly, because the approximation made in the derivation of eqn (27a) are no longer valid, partly, because the numerical detection of the error threshold suffers from enormous scatter and becomes obscure at populations of $N < 100$.

The only previous derivation of an error threshold in finite populations (Eigen *et al.*, 1989) makes use of an inequality derived by McCaskill (1984) for the fraction of the master sequence:

$$\frac{\sigma q^{\nu}}{\sigma q^{\nu} - 1} = \varepsilon \bar{x}_0, \qquad \varepsilon \ll 1.$$

The parameter ε is adjusted to the underlying value landscape. A choice of $\varepsilon = 0{\cdot}1$ was used in Eigen *et al.* (1989).

Combination of the inequality with the conventional expression from perturbation theory

$$\bar{x}_0 = \frac{q^{\nu} - \sigma^{-1}}{1 - \sigma^{-1}} N$$

yields a quadratic equation for the critical replication accuracy which we transform such that it can be compared to eqn (27a)—we use again $q_{\min}(\infty) = \sigma^{-1}$ and $\alpha = \sqrt{\sigma - 1}$:

$$q_{\min}(N) = q_{\min}(\infty) \sqrt[\nu]{1 + \frac{\alpha^2}{2\varepsilon N} + \sqrt{\frac{\alpha^2}{\varepsilon N} + \left(\frac{\alpha^2}{2\varepsilon N}\right)^2}}$$

$$\approx q_{\min}(\infty) \sqrt[\nu]{1 + \frac{\alpha}{\sqrt{\varepsilon N}} + \frac{\alpha^2}{2\varepsilon N} + \frac{\alpha^3}{8(\sqrt{\varepsilon N})^3} \cdots} \qquad (27b)$$

The error threshold thus is obtained as a series expansion in $\alpha/\sqrt{N}$. Equations (27a) and (27b) become identical in the first three terms of the series expansion within the root for a choice: $\varepsilon = 0{\cdot}25$. The remaining term of order $1/(\sqrt{N})^3$ is different as far as it requires a choice of $\varepsilon = 1/(2\sqrt{2})$ in order to become identical in both expansions. In Table 2 we compare numerical values for the different approximations made in the derivation of eqn (27a) with the values computed from eqn (27b) with the conventional choice $\varepsilon = 0{\cdot}1$ and with $\varepsilon = 0{\cdot}25$.

In essence, the results boil down to a *reciprocal square root N law* for the error threshold in the limit of large populations:

$$q_{\min}(N) = q_{\min}(\infty) \left(1 + \frac{2\sqrt{\sigma - 1}}{\nu\sqrt{N}} + \ldots\right).$$

The slope of the limiting straight line is proportional to the square root of a *superiority* increment $(\sigma - 1)$ and inversely proportional to the chain length ν of the polynucleotide sequences under consideration. This simple analytical expression is in good agreement with computer simulation data down to population sizes of a few hundred particles.

TABLE 2

Different approximations in the computation of error thresholds (q_{min}) derived from the birth and death model presented in this work and from perturbation theory (McCaskill, 1984; Eigen et al., 1989)

Parameters			*Birth* and *death* model ($\kappa = 0$)			Perturbation theory eqn (27b)	
				Eqn (27a)†			
ν	σ	N	Eqn (22)	E	A	$\varepsilon = 0{\cdot}1$	$\varepsilon = 0{\cdot}25$
10	10	100	0·8408	0·8427	0·8584	0·8706	0·8427
		200	0·8276	0·8285	0·8359	0·8484	0·8285
		500	0·8155	0·8159	0·8187	0·8285	0·8159
		1000	0·8094	0·8095	0·8109	0·8184	0·8095
		10 000	0·7991	0·7991	0·7992	0·8019	0·7991
		100 000	0·7958	0·7958	0·7959	0·7967	0·7958
		$N \to \infty$	0·7943	0·7943	0·7943	0·7943	0·7943
50	10	100	0·9659	0·9664	0·9704	0·9727	0·9664
		200	0·9629	0·9631	0·9650	0·9677	0·9631
		500	0·9600	0·9601	0·9609	0·9631	0·9601
		1000	0·9586	0·9586	0·9590	0·9607	0·9586
		10 000	0·9561	0·9561	0·9562	0·9568	0·9561
		$N \to \infty$	0·9550	0·9550	0·9550	0·9550	0·9550

† Two values are given: the first one (E) was obtained by exact evaluation of the root, the second one (A) represents the approximation.

Finally, we ask whether or not the assumptions made initially are likely to change the results. The replacement of real polynucleotides by binary sequences has important consequences for the richness of secondary structures but is of no substantial influence on the occurence of error thresholds. Polynucleotides are characterized by three mutation rates for every digit and hence the computation of elements of the mutation matrix, Q_{ik}, is more involved. The phenomenon of the error threshold, however, is mainly caused by their exponential dependence on chain lengths (q^ν) which in essence is unaffected by some variation of the single digit accuracies. Thus, the assumption of a uniform error rate per digit ($p = 1 - q$) is consistent with the consideration of binary sequences. The choice of a simple *single peak* value landscape can be interpreted in terms of the *phenomenological* approach to the error propagation problem (Eigen *et al.*, 1989) where only the master sequence (I_0) and a kind of mean of the mutant distribution is distinguished. This approach fails only if two or more sequences have the same—or almost the same—selective values. In such cases of kinetic degeneracy the *error tail* develops around a group of two or more sequences (Schuster & Swetina, 1988) and the phenomenological approach has to be replaced by a more sophisticated treatment.

This work was supported financially by the Hochschuljubiläumsstiftung Wien. IBM Austria provided personal computer equipment for this study. Generous supply with computer time on the IBM 3081 mainframe by the EDV Zentrum, Universität Wien is gratefully acknowledged.

REFERENCES

DEMETRIUS, L. (1987). *J. Chem. Phys.* **87**, 6939.

DEMETRIUS, L., SCHUSTER, P. & SIGMUND, K. (1985). *Bull. math. Biol.* **47**, 239.

DOMINGO, E., HOLLAND, J. J. & AHLQUIST, P. (eds) (1988). *RNA Genetics. Vol. III: Variability of Virus Genomes.* Boca Raton, FL.: CRC Press.

EIGEN, M. (1971). *Naturwissenschaften* **58**, 465.

EIGEN, M. & BIEBRICHER, C. K. (1988). Sequence space and Quasispecies Distribution. In: *RNA Genetics. Vol. III: Variability of Virus Genomes.* (Domingo, E., Holland, J. J. & Ahlquist, P., eds.), pp. 212–245. Boca Raton, FL.: CRC Press.

EIGEN, M., MCCASKILL, J. & SCHUSTER, P. (1988). *J. phys. Chem.* **92**, 6881.

EIGEN, M., MCCASKILL, J. & SCHUSTER, P. (1989) *Adv. Chem. Phys.* **75**, 149.

EIGEN, M. & SCHUSTER, P. (1977). *Naturwissenschaften* **64**, 541.

GILLESPIE, D. T. (1976). *J. Comp. Phys.* **22**, 403.

HUNTER, J. (1983). *Mathematical Techniques of Applied Probability.* Vol. I and Vol. II. New York: Academic Press.

KARLIN, S. & TAYLOR, H. M. (1975). *A First Course in Stochastic Processes* (2nd ed.) pp. 131–137. New York: Academic Press.

LEUTHÄUSSER, I. (1987). *J. statist. Phys.* **48**, 343.

MAYNARD SMITH, J. (1978). *The Evolution of Sex.* pp. 33–36. Cambridge: Cambridge University Press.

MCCASKILL, J. (1984). *Biol. Cybern.* **50**, 63.

MORAN, P. A. P. (1958). *Proc. Camb. Phil. Soc.* **54**, 60, 463.

NOWAK, M. (1987). *Eine numerische Simulation der RNA Replikation zur Berechnung der stochastischen Error Threshold.* Diploma Thesis. Universität Wien.

PICHLER, E. & SCHUSTER, P. *Replication accuracy and extinction probabilities in unconstrained populations.* (in press).

SCHUSTER, P. & SWETINA, J. (1988). *Bull. math. Biol.* **50**, 635.

SWETINA, J. & SCHUSTER, P. (1982). *Biophys. Chem.* **16**, 329.