Increase in error threshold for quasispecies by heterogeneous replication accuracy

Kazuhiro Aoki
White-Bird Institute, 1-14-9 Hanabatake, Tsukuba, Ibaraki 300-3261, Japan

Mitsuru Furusawa*
Neo-Morgan Laboratory Inc., Imperial Tower 15F, 1-1-1 Uchisaiwai-cho, Chiyoda-ku, Tokyo 100-0011, Japan
and Discovery Research Laboratory, Daiichi Pharmaceutical Co. Ltd., 1-16-13 Kita-Kasai, Edogawa-ku, Tokyo 134-8630, Japan
(Received 26 March 2003; revised manuscript received 29 May 2003; published 10 September 2003)

In this paper we investigate the error threshold for quasispecies with heterogeneous replication accuracy. We show that the coexistence of error-free and error-prone polymerases can greatly increase the error threshold without a catastrophic loss of genetic information. We also show that the error threshold is influenced by the number of replicores. Our research suggests that quasispecies with heterogeneous replication accuracy can reduce the genetic cost of selective evolution while still producing a variety of mutants.

DOI: 10.1103/PhysRevE.68.031904 PACS number(s): 87.10.+e, 64.60.Cn

I. INTRODUCTION

Evolution requires both genetic diversity and stable reproduction of advantageous mutants. Accurate replication of the genome guarantees stable reproduction, while errors during replication produce genetic diversity. One key to evolution is thus inherent in replication accuracy. Replication accuracy depends on nucleotide polymerases. It was believed that intracellular polymerases have homogeneous replication accuracy. Most studies of evolutionary models have also been based on homogeneous replication accuracy. Recent investigations, however, have demonstrated that organisms replicate with heterogeneous replication accuracy, i.e., error-free and error-prone polymerases coexist in the same cell.

A. Error-free and error-prone polymerases

Many error-prone DNA polymerases were recently discovered in succession, from bacteria to humans [1–7]. The major replicative DNA polymerases have a proofreading function, which eliminates errors by $3' \rightarrow 5'$ exonuclease activity [8–10], with resultant error-free replication [11,12]. The error-prone polymerases, by contrast, have no proofreading function and bypass DNA damage, thereby engendering mutagenic activity [1–6]. Induction of error-prone polymerases suggests their participation in evolution, carcinogenesis, and diversification of antibodies [13].

The major replicative polymerases could potentially change into error-prone ones by a down regulating of their proofreading activities [14]. In this context, it is especially interesting that the proofreading and DNA synthesis activities reside on distinct subunits in the DNA polymerase III holoenzyme of Gram-negative bacteria [9]. The replication accuracy in nature is thus thought to be variable and heterogeneous.

B. Disparity model: Promotion of evolution by coexistence of error-free and error-prone polymerases

We expect that heterogeneous replication accuracy should influence evolution advantageously. There is a limit to homogeneous replication accuracy (parity model) for realization of both genetic diversity and stable reproduction. If different kinds of polymerases with and without proofreading coexist, an error-prone polymerase would extend genetic diversity and an error-free polymerase would replicate an advantageous mutant (disparity model). The disparity model of a population could cause evolution to continue without losing fitness once it is acquired. This model is an extension of the original disparity hypothesis [15–17] with respect to the evolution of bacteria or higher organisms. We have demonstrated rapid evolution of the disparity model of a population in a stochastic simulation [14]. In this paper, we would like to show the evolutionary advantages of the disparity model in the context of the quasispecies theory.

C. Quasispecies and error threshold

Quasispecies is a model of evolution with error-prone replication, which Eigen and his co-workers introduced and developed [18–22]. Many modifications of the quasispecies model have been studied (for example, finite population size [23,24], spatially resolved systems [25], maternal effects [26], dynamic fitness landscapes [27–30], or other various fitness landscapes [31–35]).

Quasispecies can be defined as a stable ensemble of the fittest sequence (or master sequence) and its mutants distributed around the master sequence in sequence space with selection. The target of natural selection appears to be not a single sequence but rather an entire quasispecies distribution. The evolution of quasispecies occurs as follows: a mutant with a higher fitness than the master sequence appears in the quasispecies, this mutant replaces the old master sequence with selection, and then a new quasispecies distribution organizes around the mutant.

Studies of quasispecies have led to the conclusion that there exists an error threshold for maintaining genetic information and that quasispecies can only evolve below this threshold [18–22]. This means that the upper limit of evolution rate is determined by the error threshold. The quasispecies theory proved to be successful in studies of RNA viruses, which evolve at a high rate near the error threshold.

*Email address: furusm0q@daiichipharm.co.jp

This theory is also expected to provide a framework to examine the features of bacterial evolution and carcinogenesis, because recent studies have found evidence that mutator phenotypes with an increased error rate play an important role in these processes [36–39].

D. Purpose of this study

Although a great deal of research has been carried out regarding the genetics and biochemistry of the error-prone polymerases, little is known about their influence on evolution. Most studies of evolutionary models, including those of quasispecies, have not focused on heterogeneous replication accuracy. In this paper we consider the application of the disparity model to quasispecies theory and its influence on the error threshold.

The bacterial genome is replicated bidirectionally from a single origin of replication, and eukaryotes have multiple origins of replication in the genome [10]. This means that the genome sequence is partitioned into more than one replication unit (replicore), and thus more than one polymerase can participate in genome replication simultaneously. In this paper we also consider the influence of the number of replicores on the error threshold.

II. DISPARITY-QUASISPECIES HYBRID MODEL

A. Mutant distribution of quasispecies with heterogeneous replication accuracy

In the present study, a quasispecies consists of a population of genomes, each represented by a binary base sequence of length n, which has 2^n possible genotypes (or sequence space). A sequence with the best fitness is called the master sequence. The population size is assumed to be very large and constant. The replication of one template sequence produces one direct copy sequence, and thus the replication error is fixed to a mutation by one step. Only base substitutions occur, and hence the sequence length is constant. Sequence degradation is neglected. For easy handling, we classify the sum of all i-error mutants of the master sequence (I_0) into a mutant class $I_i (i=0,1,\ldots,n)$. The corresponding sum of relative concentrations is denoted by x_i. The rate of change in x_i is then

$$\dot{x}_i = (A_i Q_{ii} - f) x_i + \sum_{j \neq i} A_j Q_{ij} x_j , \tag{1}$$

where A_i is the replication rate constant (or fitness) of the mutant class I_i; f keeps the total concentration constant, and is then $\Sigma_i A_i x_i$; Q_{ii} is the replication accuracy or the probability of producing I_i by complete error-free copying of I_i; Q_{ij} is the probability of producing I_i by miscopying of I_j.

The genome sequence is replicated by a polymerase. By E_k we denote p kinds of polymerases with different accuracies ($k=1,2,\ldots,p$). The relative concentration of E_k is denoted by c_k. Single-base accuracy of polymerase E_k is $0 \leqslant q_k \leqslant 1$, so that the per base error rate is $1-q_k$. Because of the consistent replication of one sequence by the same polymerase, the per genome error rate of E_k is $n(1-q_k)$. The per genome mean error rate of the quasispecies is then $n\Sigma_k c_k(1-q_k)=m$. By transforming the homogeneous replication accuracy [20,22], we have a heterogeneous one:

$$Q_{ij} = \sum_k c_k q_k^n \sum_{h=0}^{l} \left(\frac{1-q_k}{q_k} \right)^{2h+|j-i|} \binom{n-j}{h+\frac{1}{2}(|j-i|-j+i)} \times \binom{j}{h+\frac{1}{2}(|j-i|+j-i)},$$

$$\text{with} \quad l = \left[\tfrac{1}{2} (\min\{j+i, 2n-(j+i)\} - |j-i|) \right].$$

The stationary mutant distribution, $\lim_{t\to\infty} x_i = y_i$, is a quasispecies. This is obtained from the eigenvectors of matrix $\mathbf{W}=\{A_j Q_{ij}\}$ [20–22]. Figure 1 shows the examples of the quasispecies with homogeneous and heterogeneous replication accuracies. We used a simple single-peaked fitness landscape for easy calculations. A replication rate constant, A_0 is assigned to the master sequence, and all other mutant classes have the same fitness ($A_1=A_2=\ldots=A_n<A_0$).

Parity quasispecies with a homogeneous replication accuracy below the error threshold localizes around the master sequence [Fig. 1(a)]. At the error threshold near $m=2.3$, the transition is very sharp, and the relative concentration of the master sequence decreases over ten orders of magnitude (at $c=0$ in Fig. 2). Such a phenomenon is called an error catastrophe. Above the error threshold, quasispecies localization is replaced by a uniform distribution, in which individual concentrations are extremely small: $y_i = 8.88 \times 10^{-16}$. In a real, finite population, the genetic information of the master sequence can no longer be maintained by selection due to error accumulation. Only below the error threshold can the quasispecies evolve, and the rate of evolution appears to reach its maximum near the error threshold.

The disparity quasispecies [Figs. 1(b)–1(d)] have two kinds of polymerases, each with different accuracy. Polymerase E_1 is error free, $q_1=1$, and E_2 is error prone, $0 \leqslant q_2 \leqslant 1$; each present at a relative concentration of c and $1-c$. Of course, the assumption of a complete error-free polymerase is not realistic; however, the error rate of the proofreading polymerase in DNA-based microbes is very small, 0.003 errors per genome per replication [11], thus it is negligible in this case.

When the relative concentration of error-free polymerase is low, $0<c<0.1$, the error threshold is shifted to a higher mean error rate with increasing c, and the magnitude of the error catastrophe decreases [Figs. 1(b) and 2]. At $c=0.1$ the error threshold vanishes [Fig. 1(c)]. The relative concentration of the master sequence gradually decreases and finally levels off at a 10^7 times higher concentration than the parity uniform distribution (at $c=0.1$ in Fig. 2). When $c>0.1$, independent of the mean error rate, the master sequence is present in sufficient concentration [Figs. 1(d) and 2]. Figure 2 shows the dramatic change of the quasispecies dynamics near $c_{\text{crit}}=0.1$. In the disparity quasispecies, mutants far distant from the master sequence can be present without incur-

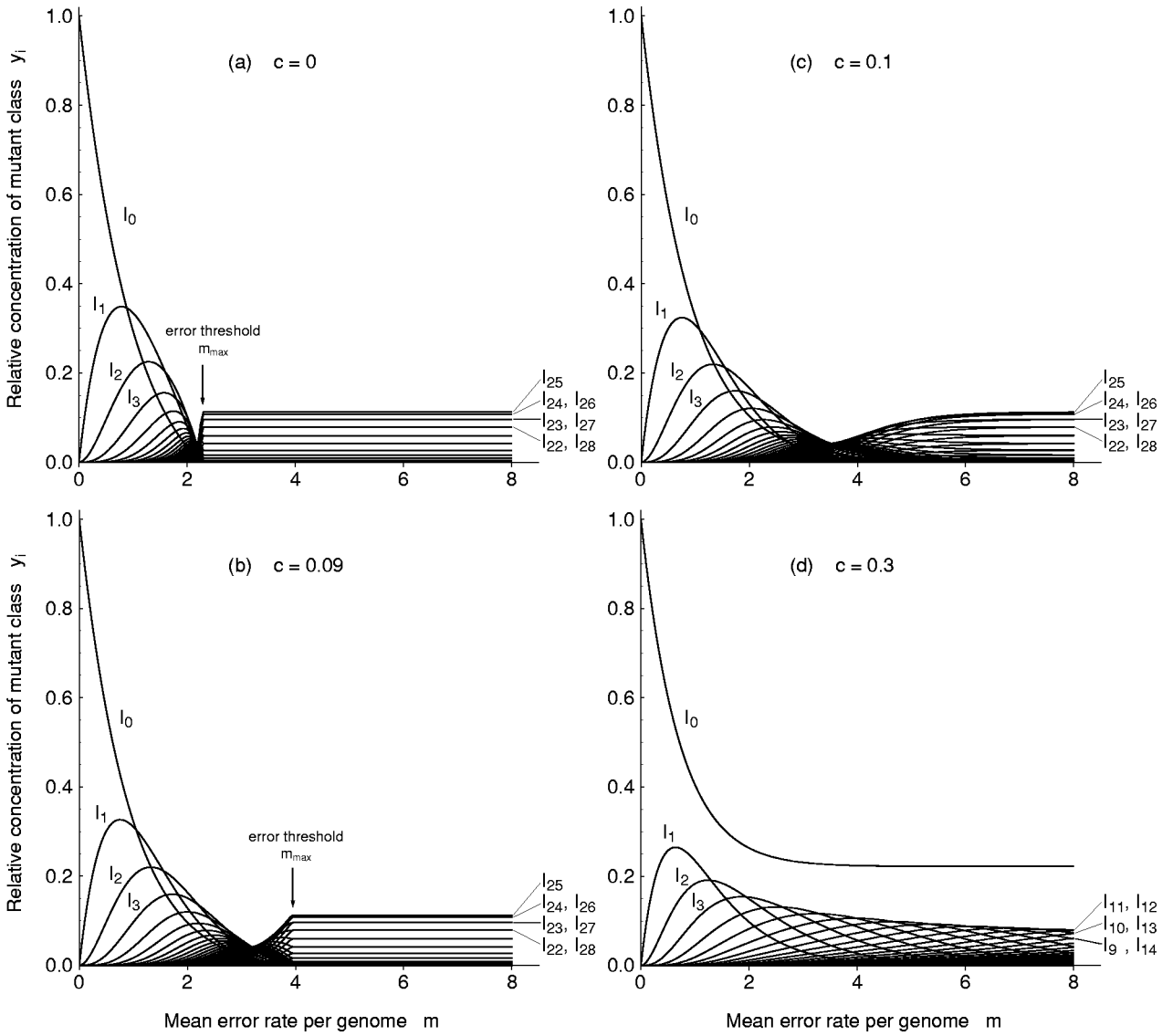

FIG. 1. The mutant distribution in quasispecies as a function of the mean error rate per genome ($n=50$), where c is the relative concentration of error-free polymerase. We plot the relative stationary concentration of the master sequence (I_0), the sum of the relative stationary concentration of all one error mutants (I_1), of all two error mutants (I_2), etc. (a) The parity model with homogeneous replication accuracy: $c=0$. (b)–(d) Typical examples of the disparity model: $c>0$. Error-free and error-prone polymerases coexist. The following selective values were used in all of the examples: $A_0=10$, $A_i=1$ for all $i\neq 0$.

ring the loss of quasispecies localization. This means that the rate of evolution can increase without error catastrophe.

B. Error threshold for quasispecies with multiple replicores

Considering the error threshold for the disparity model, we encounter the following two difficulties: (i) the genome size in nature is too large, virus: $n>10^3$, bacteria: $n>10^6$, to do exact calculations; and (ii) the genome replication in nature is partitioned into more than one unit (replicore) and more than one polymerase participates at the same time. The multiple replicores appear to influence the error threshold. Therefore, we approach the error threshold by using an approximation of the relative stationary concentration of the master sequence [18–22]:

$$y_0 \approx \frac{A_0 Q_{00} - A_{i\neq 0}}{A_0 - A_{i\neq 0}},$$

where A_0 is the replication rate constant of the master sequence and $A_{i\neq 0}$ is the overall average of other mutant sequences; Q_{00} is the replication accuracy for complete error-free copying of the master sequence. This approximation relies on the negligence of considering back mutations from mutants to the master sequence in Eq. (1). Agreement with the exact solution increases with increasing genome size [20]. The relative stationary concentration of the master sequence vanishes for a critical error rate that fulfills

$$(Q_{00})_{\min} = \frac{A_{i\neq 0}}{A_0} = s^{-1}, \tag{2}$$

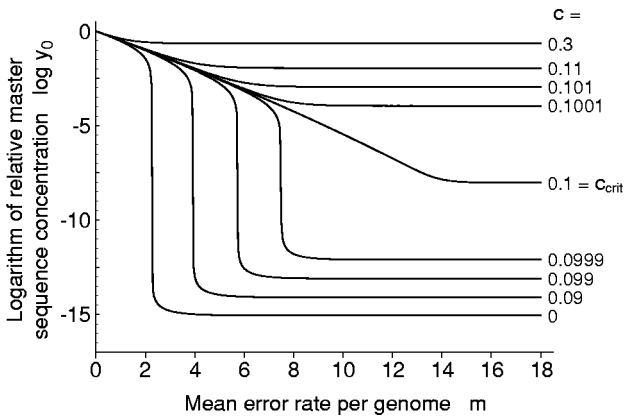

FIG. 2. Log_{10} plot of the relative stationary concentration of the master sequence as a function of the mean error rate at various relative concentrations of error-free polymerase (c). Parity model: $c=0$. Disparity mode: $c>0$. The following selective values and parameter were applied: $A_0=10$, $A_i=1$ for all $i\neq 0$ and $n=50$.

where s is the selective superiority of the master sequence. To obtain Q_{00} for the disparity model with multiple replicores, we assume that there are two kinds of polymerases E_1 and E_2, each present at a relative concentration of c and $1-c$. The error rate of the proofreading polymerases is very small and negligible. Thus, polymerase E_1 is error free, $q_1=1$, and E_2 is error prone, $0\leqslant q_2\leqslant 1$. The per genome mean error rate is then

$$m=n(1-c)(1-q_2). \tag{3}$$

The probability of replicating the genome by error-prone polymerase E_2 is obtained from a binomial distribution. The nonerror probability by the error-prone polymerase E_2 is obtained from a Poisson approximation, in which the genome size is assumed to be very large compared with the number of replicores. Multiplying them we have

$$\begin{aligned} Q_{00}&=\sum_{b=0}^{a}\binom{a}{b}c^{a-b}(1-c)^b e^{-mb/a(1-c)} \\ &=[c+(1-c)e^{-m/a(1-c)}]^a, \end{aligned} \tag{4}$$

where a is the number of all replicores in the genome. Combining Eqs. (2) and (4) we have the error threshold for the disparity model:

$$m_{\max}=a(1-c)\ln\left(\frac{1-c}{s^{-1/a}-c}\right). \tag{5}$$

Figure 3 shows the error threshold as a function of the relative concentration of error-free polymerase at various numbers of replicores. The error threshold for the parity model, $c=0$, is not influenced by the number of replicores. In the disparity model, $c>0$, the singularity occurring at the critical concentration of the error-free polymerase,

$$c_{\mathrm{crit}}=s^{-1/a},$$

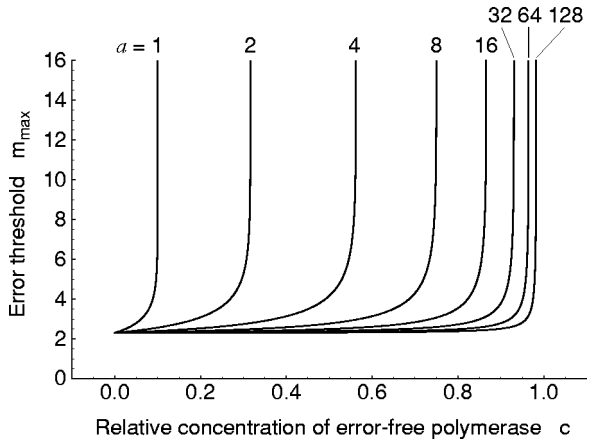

FIG. 3. Error threshold for the disparity quasispecies as a function of the relative concentration of error-free polymerase at various numbers of replicores (a). The genome size was assumed to be infinity. The following selective values were applied: $A_0=10$, $A_i=1$ for all $i\neq 0$.

leads to a very sharp increase of error threshold. This means that in $c\geqslant c_{\mathrm{crit}}$, the error threshold vanishes. The result at $a=1$ agrees with the exact solution for $n=50$ (see Fig. 2). c_{crit} increases with increasing number of replicores.

The permissible error rate is thus obtained from Eqs. (3) and (5):

$$m_{\mathrm{pms}}\begin{cases} <a(1-c)\ln\left(\dfrac{1-c}{s^{-1/a}-c}\right), & c<z \\ \leqslant n(1-c)(1-q_{\min}), & c\geqslant z, \end{cases}$$

$$z=\frac{\exp(nq_{\min}/a)-\exp(n/a)s^{-1/a}}{\exp(nq_{\min}/a)-\exp(n/a)}\approx s^{-1/a}$$

assuming $s^{-1/a}\ll 1$. When $c\geqslant z$ there are two constraints: (i) the genome size n is finite and (ii) the error-prone polymerase has a nonzero accuracy $q_{\min}$ in real organisms. The error rate of the complete proofreading-free DNA poly-

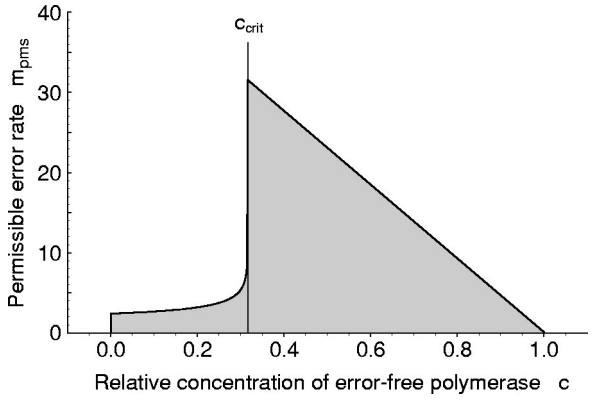

FIG. 4. Permissible error rate of the disparity quasispecies as a function of the relative concentration of error-free polymerase. The permissible region is the shaded one. The following selective values and parameters were applied: $A_0=10$, $A_i=1$ for all $i\neq 0$, $n=4.6\times 10^6$, $a=2$, and $1-q_{\min}=10^{-5}$.

merase of *Escherichia coli* (*E. coli*) is assumed to be $1-q_{\min}=10^{-5}$ [40]. Figure 4 shows an example of the permissible error rate based on the parameters of *E. coli*. The plot resembles a λ transition in shape. For $s=10$, the maximum of m_{pms} of *E. coli* becomes 31 errors per genome per replication. This error rate is sufficiently high compared with the error threshold of the parity model: $\ln(s)=2.3$.

III. CONCLUSION

In this paper, we analyzed a disparity-quasispecies hybrid model in which both error-free and error-prone polymerases coexist. The results show that the dynamics of a quasispecies are determined not only by the error rate but also by the proportion of polymerases with different accuracies and by the number of replicores partitioning the genome. One notable finding to emerge was that the coexistence of the error-free and error-prone polymerases could greatly increase the error threshold for quasispecies compared with the traditional parity model.

Many organisms in nature live in a continuously changing environment [27–30,41]. This is especially true for microbial pathogens and tumor cells dodging the host immune system. The chance of finding an advantageous mutant will increase with increasing Hamming distance from the master sequence, because of the large increase in the number of mutants, and hence possible candidates, with increasing distance [22].

A simple homogeneous increase in the error rate would incur a considerable cost of deleterious mutations, even if it were transient. So small is the error threshold of the parity quasispecies that the distribution range of mutants is limited to a short distance from the master sequence. In the ''hill-climbing'' metaphor [42,43] of adaptive evolution, the parity quasispecies would be trapped in a local low peak on the rugged fitness landscape and could never reach the higher peaks far from the master sequence. The disparity quasispecies, on the other hand, could increase the error threshold without losing genetic information, and hence produce a large number of advantageous mutants with increasing distance from the master sequence. The disparity quasispecies could search long distances across the sequence space and finally find a higher peak.

The processivity of the error-prone polymerases seems to be much lower than that of the major replicative polymerases with proofreading [44]. The disparity model with multiple replicores takes this observation into account. In this model, errors are concentrated within regions of the replicores in which error-prone polymerases participate. If an error-prone replication is restricted within a specific gene region, the error rate of the region greatly increases as the costs for the other genes keep to a minimum.

ACKNOWLEDGMENT

We thank Dr. Mitsuoki Kawano for useful comments and discussions.

[1] J. Wagner *et al.*, Mol. Cell **4**, 281 (1999).
[2] M. Tang *et al.*, Proc. Natl. Acad. Sci. U.S.A. **96**, 8919 (1999).
[3] N.B. Reuven, G. Arad, S.A. Maor, and Z. Livneh, J. Biol. Chem. **274**, 31 763 (1999).
[4] V.L. Gerlach *et al.*, Proc. Natl. Acad. Sci. U.S.A. **96**, 11 922 (1999).
[5] E.C. Friedberg, R. Wagner, and M. Radman, Science **296**, 1627 (2002).
[6] M.F. Goodman, Annu. Rev. Biochem. **71**, 17 (2002).
[7] H.I. Boshoff *et al.*, Cell **113**, 183 (2003).
[8] H. Echols, C. Lu, and P.M. Burgers, Proc. Natl. Acad. Sci. U.S.A. **80**, 2189 (1983).
[9] R.H. Scheuermann and H. Echols, Proc. Natl. Acad. Sci. U.S.A. **81**, 7747 (1984).
[10] A. Kornberg and T. A. Baker, *DNA Replication*, 2nd ed. (Freeman, New York, 1992).
[11] J.W. Drake, Proc. Natl. Acad. Sci. U.S.A. **88**, 7160 (1991).
[12] R.M. Schaaper, J. Biol. Chem. **268**, 23762 (1993).
[13] M. Radman, Nature (London) **401**, 866 (1999).
[14] K. Aoki and M. Furusawa, J. Theor. Biol. **209**, 213 (2001).
[15] M. Furusawa and H. Doi, J. Theor. Biol. **157**, 127 (1992).
[16] K.N. Wada *et al.*, Proc. Natl. Acad. Sci. U.S.A. **90**, 11 934 (1993).
[17] M. Furusawa and H. Doi, Genetica (Dordrecht, Neth.) **102/103**, 333 (1998).
[18] M. Eigen, Naturwissenschaften **58**, 465 (1971).
[19] M. Eigen and P. Schuster, Naturwissenschaften **64**, 541 (1977).
[20] J. Swetina and P. Schuster, Biophys. Chem. **16**, 329 (1982).
[21] M. Eigen, J. McCaskill, and P. Schuster, J. Phys. Chem. **92**, 6881 (1988).
[22] M. Eigen, J. McCaskill, and P. Schuster, Adv. Chem. Phys. **75**, 149 (1989).
[23] M. Nowak and P. Schuster, J. Theor. Biol. **137**, 375 (1989).
[24] P.R.A. Campos and J.F. Fontanari, J. Phys. A **32**, L1 (1999).
[25] S. Altmeyer and J.S. McCaskill, Phys. Rev. Lett. **86**, 5819 (2001).
[26] C.O. Wilke, Phys. Rev. Lett. **88**, 078101 (2002).
[27] M. Nilsson and N. Snoad, Phys. Rev. Lett. **84**, 191 (2000).
[28] C.O. Wilke, C. Ronnewinkel, and T. Martinetz, Phys. Rep. **349**, 395 (2001).
[29] C. Kamp and S. Bornholdt, Phys. Rev. Lett. **88**, 068104 (2002).
[30] C. Kamp *et al.*, Complexity **8**, 28 (2002).
[31] G. Woodcock and P.G. Higgs, J. Theor. Biol. **179**, 61 (1996).
[32] E. van Nimwegen, J.P. Crutchfield, and M. Huynen, Proc. Natl. Acad. Sci. U.S.A. **96**, 9716 (1999).
[33] C.O. Wilke, Bull. Math. Biol. **63**, 715 (2001).
[34] C.O. Wilke *et al.*, Nature (London) **412**, 331 (2001).
[35] D.C. Krakauer and J.B. Plotkin, Proc. Natl. Acad. Sci. U.S.A. **99**, 1405 (2002).
[36] J.E. LeClerc, B. Li, W.L. Payne, and T.A. Cebula, Science **274**, 1208 (1996).

[37] I. Matic *et al.*, Science **277**, 1833 (1997).
[38] P.D. Sniegowski, P.J. Gerrish, and R.E. Lenski, Nature (London) **387**, 703 (1997).
[39] L.A. Loeb, Cancer Res. **61**, 3230 (2001).
[40] I.J. Fijalkowska and R.M. Schaaper, Proc. Natl. Acad. Sci. U.S.A. **93**, 2856 (1996).
[41] L. Van Valen, Evol. Theory **1**, 1 (1973).
[42] S. Wright, Genetics **1**, 356 (1932).
[43] S. A. Kauffman, *The Origins of Order. Self-Organization and Selection in Evolution* (Oxford University Press, New York, 1993).
[44] J. Wagner *et al.*, EMBO Rep. **1**, 484 (2000).