J Mol Evol (1994) 38:468–475

Journal
of Molecular
Evolution

A Model for the Correlation of Mutation Rate with *GC* Content and the Origin of *GC*-Rich Isochores

X. Gu, W.-H. Li

Center for Demographic and Population Genetics, University of Texas, P.O. Box 20334, Houston, TX 77225, USA

Received: 10 September 1993 / Revised: 17 November 1993

Abstract. Based on the biochemical kinetics of DNA replication and mutagenesis, including misincorporation and correction, a model has been developed for studying the relationships among the mutation rate (u), the $G + C$ content of the sequence (f), and the $G + C$ proportion in the nucleotide precursor pool (N). Also a measure for the next-nucleotide effect, called the maximum capacity of the next-nucleotide effect (MC), has been proposed. Under the normal physiological conditions of mammalian germ cells, our results indicate: (1) the equilibrium $G + C$ content in a sequence is approximately equal to the $G + C$ proportion in the nucleotide precursor pool, i.e., $f \approx N$, which is independent of the next-nucleotide effect; (2) an inverted-V-shaped distribution of mutation rates with respect to $G + C$ contents is predicted, when the next-nucleotide effect is week, i.e., $MC \approx 1$; (3) the distribution becomes flatter (i.e., inverted-U-shaped) as MC increases, but the peak at 50% GC is still observed when $MC < 2$; and (4) the peak disappears when $MC > 2.8$, that is, when the next-nucleotide effect becomes strong. Our results suggest that changes in the relative concentrations of nucleotide precursors can cause variations among genes both in mutation rate and in $G + C$ content and that compositional isochores (DNA segments with a homogeneous $G + C$ content) can arise in a genome due to differences in replication times of DNA segments.

Key words: DNA replication — Misincorporation — Correction — Nucleotide precursors — Variation in mutation rate — Variation in $G + C$ content

Introduction

Wolfe et al. (1989) found that the rate of synonymous substitution between rat and mouse varies among genes, leading to an inverted-V-shaped distribution of substitution rates with respect to $G + C$ contents of sequences. Since synonymous changes are likely to be nearly neutral, the variation in synonymous rate among genes points to the existence of variation in mutation rate among genes. To explain the inverted-V-shaped distribution of mutation rates with respect to $G + C$ contents, Wolfe et al. (1989) proposed that both the mutation rate and the $G + C$ content of a neutral sequence are determined by the relative concentration of nucleotides (dNTPs) in the precursor pool during DNA replication in mammalian germ cells. Their hypothesis, which will be called the mutationist hypothesis, was based on the observations that the mutation rate depends on the relative concentrations of nucleotide precursors (see Kunz and Kohalmi 1991, for review) and that the relative concentrations of nucleotide precursors vary through the cell cycle and different sequences in a mammalian genome are replicated at different times. (See Leed et al. 1985; Holmquist 1987; Goldman 1988.)

Wolfe et al. (1989) also suggested that the mutationist hypothesis can explain the origin of *GC*-rich isochores in the genome of mammals and other vertebrates. Compositional isochores are long DNA segment (30 kb or longer) with a homogeneous $G + C$ content (Bernardi et al. 1985). There is some evidence that compositional isochores are related to the types of chromosomal bands and to their replication time in cell division (Ikemura and Aota 1988; Holmquist 1987, 1992). Generally speaking, *GC*-rich isochores are early repli-

Correspondence to: W.-H. Li

cating in mammalian genomes whereas GC-poor isochores are late-replicating (Holmquist 1987, 1992; Bernadi 1989). Recently, this conclusion was challenged by Eyre-Walker (1992a), who provided evidence that both GC-rich and GC-poor isochores are replicated early and late in the somatic cell cycle. However, his data were from somatic cell lines and so may not be applicable to germ cells. Thus, this issue remains to be settled. Based on the traditional view of the relationship between $G + C$ content and replication time, Wolfe et al. (1989) suggested that compositional isochores arose because of differences in replication-timing of different replicons in the germ cells. In contrast, Bernardi et al. (1985, 1988) proposed that GC-rich isochores arose because of selective advantages. Their main argument is that in warm-blooded vertebrates an increase in $G + C$ content can protect DNA, RNA, and proteins from degradation by heat, because G-C bonds are stronger than A-T bonds.

The same distribution pattern of synonymous rates was also observed recently in a study of more than 50 genes in humans and rodents (Bulmer et al. 1991) and in a study of more than 300 genes in rat and mouse (Wolfe and Sharp 1993), although the correlation between the synonymous substitution rate and the $G + C$ content was considerably weaker than that in Wolfe et al. (1989) i.e., the distribution became more like an inverted U rather than an inverted V. Therefore, it is interesting to see whether one can explain the inverted-V-(or U-)shaped distribution and the origin of GC-rich isochores on the basis of the biochemical kinetics of DNA replication and mutagenesis. Two papers (Wolfe 1991; Eyre-Walker 1992b) were published on this issue but were unable to explain the inverted-V-(or U-)shaped distribution. However, we think that there are some biologically unrealistic assumptions in their models.

It is commonly believed that error during DNA replication is one of the major sources of mutation in germ cells. *Figure 1*, which is from Kunkel (1992a), illustrates the ordered sequential DNA polymerization reaction in prokaryotes. There are several coupled mechanisms for the fidelity of DNA replication. The first is the kinetic discrimination of a polymerase against nucleotide misincorporation. It has been suggested that the sequential change in the conformation of the polymerase-template-precursor complex (i.e., from step B to D in Fig. 1) has a slower rate if the incorporated nucleotide is incorrect (Kunkel 1992a). This delay may provide an opportunity for an associated exonuclease to correct the error via the E_1 pathway (Fig. 1). Another opportunity for correction is via the E_2 pathway (Fig. 1); a misincorporation results in a slower rate for the incorporation of the next correct nucleotide because a mismatched nucleotide distorts the conformation of the polymerase-template-precursor complex and retards the entry into the next cycle of polymerization. Therefore, a mutation is generated only when a mismatched nucleotide is not corrected by either the E_1 pathway or the E_2 pathway during DNA replication.

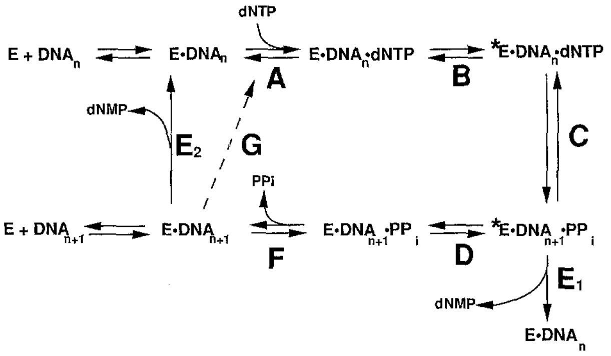

Fig. 1. Reaction pathway for DNA polymerization. Asterisks (*) represent enzyme in a different conformation. Entry into the next cycle of polymerization is indicated by G. Taken from Kunkel (1992a).

In this paper, a model will be proposed based on the process of DNA replication shown in Fig. 1. We mention here that the scheme in Fig. 1 is based mainly on studies in prokaryotes (primarily for the Klenow polymerase). In eukaryotes, DNA replication is complicated (Kunkel 1992b), but not necessarily more than in prokaryotes with respect to Fig. 1, because some basic features illustrated in Fig. 1 should be the same in both prokaryotes and eukaryotes. Therefore, it can be used as a working hypothesis for the study of DNA replication fidelity. After discussing data on the normal physiological conditions in mammalian cells, the relations among the mutation rate (u), the $G + C$ content in a sequence (f), and the $G + C$ proportion in the pool of precursors (N) will be investigated extensively.

Methods

Misincorporation of Nucleotides. At a given site with the correct nucleotide k, let v_k and v'_k be, respectively, the rates of correct and incorrect incorporation during DNA replication. The two rates are determined by the following Michaelis enzymic kinetic equations

$$v_k = \frac{V_{max} C_k}{K_m + C_k} \tag{1}$$

$$v'_k = \frac{V'_{max} C'_k}{K'_m + C'_k} \tag{2}$$

where V_{max} and V'_{max} are the maximum rates of polymerization for correct and incorrect incorporation, respectively, and K_m and K'_m are the Michaelis constants for correct and incorrect incorporation, respectively. In Eq. (1) and Eq. (2), C_k is the concentration of the correct nucleotide k in the nucleotide precursor pool ($k = 1, 2, 3, 4$ for A, G, C, T), and $C'_k = \Sigma_{i \neq k} C_i$ is the total concentration of the incorrect nucleotides. Furthermore, let $v'_k(i)$ be the rate of incorrect incorporation by nucleotide i, given the correct one is k, which is defined by

$$v'_k(i) = \frac{V'_{max} C_i}{K'_m + C'_k} \tag{3}$$

By definition, it is obvious that $v'_k = \Sigma_{i \neq k} v'_k(i)$. These equations are the simplest forms of DNA replication kinetics (Echols and Goodman 1991).

Let $\sigma_m(k)$ be the probability of misincorporation during DNA replication for a given site at which the correct nucleotide is *k;* the subscript *m* stands for *misincorporation.* Further, let $\sigma_m(k \rightarrow i)$ be the probability that the correct nucleotide *k* is misincorporated as nucleotide *i.* These two probabilities are defined by the relative rates of misincorporation—that is,

$$\sigma_m(k) = \frac{v'_k}{v_k + v'_k}$$

$$\sigma_m(k \rightarrow i) = \frac{v'_k(i)}{v_k + v'_k} \tag{4}$$

so that $\sigma_m(k) = \Sigma_{i \neq k} \sigma_m(k \rightarrow i)$ holds.

Obviously, the rate of misincorporation is $v'_k = 0$ if the nucleotide precursor pool consists of only nucleotide *k*, for the concentrations of all other nucleotides are zero, i.e., $C_i = 0$, $(i \neq k)$. In this case, $\sigma_m(k) = 0$ and $\sigma_m(k \rightarrow i) = 0$. Similarly, if there is no nucleotide *k* in the precursor pool, the rate of correct incorporation v_k is zero and therefore misincorporation will occur with certainty, i.e., $\sigma_m(k) = 1$.

In human somatic cells, the value of K_m for the replicative DNA polymerase ranges from 1 to 5 μM, whether measured with purified enzymes or permeabilized cells (Dresler et al. 1988). Wong et al. (1991) estimated the value of K'_m for the *T7* DNA polymerase to be from 6,000 to 8,000 μM, whereas the value of K_m is estimated to be only 20 μM. Since the typical concentration of nucleotide precursors under physiological conditions is about 100 μM (Wong et al. 1991), these data indicate the occurrence of K_m discrimination of the DNA polymerase against misincorporation, causing the enzyme to stall for the following correction during DNA replication. Although one should be cautious to apply these data directly to mammals, it is reasonable to assume that K_m discrimination plays a key role in correct incorporation during DNA replication in mammalian germ cells, for K_m discrimination has been recognized widely as one of the general mechanisms for the fidelity of DNA replication (e.g., Echols and Goodman 1991; Wong et al. 1991; Mathews and Ji 1992).

For the above reasons, the following condition should hold:

$$K_m \ll C \ll K'_m \tag{5}$$

where $C = \sum_{k=1}^{4} C_k$ is the total concentration of nucleotide precursors under the normal physiological conditions of mammalian germ cells. The relation in Eq. (5) allows us to simplify Eq. (2) and show that the rate of misincorporation v'_k, as well as $v'_k(i)$, is approximately linearly proportional to the concentration of the incorrect nucleotides in vivo—that is,

$$v'_k \approx \frac{V'_{max}}{K'_m} \sum_{i \neq k} C_i$$

$$v'_k(i) \approx \frac{V'_{max}}{K'_m} C_i \tag{6}$$

Let N_k be the relative concentration of nucleotide k $(\Sigma_{k=1}^{4} N_k = 1)$ so that $C_k = N_k C$ and $C'_k = \Sigma_{i \neq k} C_i = (1 - N_k)C$. Given that the correct nucleotide at a given site is *k*, Eq. (4) can be approximated by

$$\sigma_m(k) = \frac{\alpha(1 - N_k)}{\alpha(1 - N_k) + \dfrac{N_k}{1 + \beta N_k}} \tag{7}$$

$$\sigma_m(k \rightarrow i) = \frac{\alpha N_i}{\alpha(1 - N_k) + \dfrac{N_k}{1 + \beta N_k}} \tag{8}$$

where the two parameters α and β are given by

$$\alpha = \frac{V'_{max}}{V_{max}} \cdot \frac{K_m}{K'_m} \tag{9}$$

$$\beta = \frac{C}{K_m} \tag{10}$$

For mathematical simplicity, we assume that both α and β are the same for all kinds of misincorporations. The parameter α, which is called the discrimination coefficient against incorrect nucleotides, determines the magnitude of the probability of misincorporation (Echols and Goodman 1991), ranging from 10^{-5} to 10^{-6} (Wong et al. 1991). The value of β, as shown later, is a key parameter to describe the variation of mutation rate with $G + C$ content. When $\beta = 0$, the current model is essentially equivalent to the models of Wolfe (1991) and Eyre-Walker (1992b). However, the condition $\beta = 0$ is not compatible with the normal physiological condition. In fact, in the typical concentration of a precursor pool, *C* is about 100 μM, and the estimate of β, based on Dresler et al.'s (1988) data in human cells, is from $\beta = 20$ (100 μM/5 μM) to $\beta = 100$ (100 μM/1 μM). Although there are no data available to estimate directly the value of β in mammalian germ cells, it is reasonable to assume that the value is, like that in human somatic cells, larger than 20. Besides, we note that the biochemical interpretation of $\beta \rightarrow \infty$ is that the rate of correct incorporation v_k approaches its maximum rate V_{max} in vivo. At any rate, the assumption of $\beta = 0$ is biologically unrealistic and perhaps underlies the failure of the models of Wolfe (1991) and Eyre-Walker (1992b) to explain the correlation of mutation rate with $G + C$ content.

The expected rate of misincorporation, denoted by $\bar{\sigma}_m$, is therefore obtained by averaging all occurrences of misincorporation over the entire sequence. That is,

$$\bar{\sigma}_m = \sum_{k=1}^{4} f_k \sigma_m(k) \tag{11}$$

where f_k is the frequency of nucleotide *k* in the sequence, and $\sigma_m(k)$ is given by Eq. (7).

Correction of Misincorporations and Next-Nucleotide Effect. Now let us consider the efficiency of correction for misincorporation during DNA replication and also the next-nucleotide effect. As seen from Fig. 1, there are two mechanisms for correction if a misincorporation occurs: the first is during the formation of the phosphodiester bond (the E_1 pathway) and the second is before the entry into the next incorporation step (the E_2 pathway) (Kunkel 1992a). The next-nucleotide effect means that the higher the concentration of the next correct nucleotide following a misincorporation is, the less efficient the correction will be, for less time is available for correction to occur (Kunz and Kohalmi 1991). For the Klenow polymerase in *E. coli,* the E_1 pathway appears to be less important than the E_2 pathway; it is not known whether the E_1 pathway exists in eukaryotes. Anyway, the E_2 pathway is subject to the next-nucleotide effect whereas the E_1 pathway is not.

Let $\sigma_n(k)$ be the probability of noncorrection for a given site where the next correct nucleotide is *k* (the subscript *n* stands for *noncorrection*). Furthermore, let $P_c(1)$ be the probability that a misincorporation is corrected through the E_1 pathway and $P_c(2)$ be the probability that a misincorporation is corrected through the E_2 pathway.

Then by definition the following relation should hold:

$$\sigma_n(k) = (1 - P_c(1))(1 - P_c(2))$$

Since the correction efficiency through the E_1 pathway is not affected by the next-nucleotide effect, the probability of noncorrection through this pathway, which is denoted by

$$h_1 = 1 - P_c(1)$$

can be assumed to be constant, independent of the relative concentrations of nucleotide precursors. By contrast, the probability of noncorrection of a misincorporation through the E_2 pathway, $1 - P_c(2)$, depends on the extension rate from the upstream mismatched nucleotide relative to that from the correct upstream nucleotide—that is,

$$1 - P_c(2) = \frac{h_2 + bN_k}{1 + bN_k}$$

(Mendelman et al. 1990), where b is a constant to describe the strength of the next-nucleotide effect and $0 < h_2 < 1$ is the minimum probability of noncorrection through the E_2 pathway when $N_k = 0$. From the above equations, we obtain

$$\sigma_n(k) = h_1 \times \left(\frac{h_2 + bN_k}{1 + bN_k}\right) \tag{12}$$

It has been suggested that the correction for misincorporations increases on average the replication fidelity by about 100-fold, which indicates that $\sigma_n(k) \sim 10^{-2}$. (For review, see Kunkel 1988, 1992.) Therefore, under equal concentrations of the four types of nucleotide precursors—that is, i.e., $N_k = 0.25$—value of b (and also h_2) should be less than 0.04 if all corrections of misincorporations are via the E_2 pathway and should be less than 0.4 if only half of them are via the E_2 pathway. For this reason, we assume that the real value of b, though unknown, is less than 1.

In Wolfe's (1991) and Eyre-Walker's (1992b) models, the probability of noncorrection is proportional to

$$\frac{N_k}{E + N_k}$$

where E is a constant (Ninio 1987). Therefore, a misincorporation is definitely corrected (i.e., $\sigma_n(k) = 0$), if the concentration of the next correct nucleotide k in the precursor pool is zero, i.e., $N_k = 0$. This might be biologically unrealistic and probably overestimates the next-nucleotide effect on the mutation rate.

The average rate of noncorrection of misincorporation over the entire sequence, denoted by $\bar{\sigma}_n$, is given by

$$\bar{\sigma}_n = \sum_{k=1}^{4} f_k \sigma_n(k) \tag{13}$$

where f_k is the frequency of nucleotide k in the sequence and $\sigma_n(k)$ is given by Eq. (12).

Mutation Rate. Combining misincorporation and correction, we obtain the mutation rate u as

$$u = \bar{\sigma}_m \times \bar{\sigma}_n = \left(\sum_{k=1}^{4} f_k \sigma_m(k)\right)\left(\sum_{k=1}^{4} f_k \sigma_n(k)\right) \tag{14}$$

The important role of postreplication repair for DNA replication fidelity has been recognized both in prokaryotes and eukaryotes. (See Modrich 1991 for a review.) However, this factor is not taken into consideration in the present study, because it is not yet clear to us how to incorporate this factor into the model.

If the strand-symmetric nature of DNA replication is assumed, we have equal G content and C content in a sequence—that is, $f_G = f_C = 0.5f$ and $f_A = f_T = 0.5(1 - f)$, where f is the $G + C$ content in the sequence; since both parental DNA strands are templates, mutation can be generated by misincorporation on either the leading or the lagging strand in replication. The direct mutation pressure model (Sueoka 1988, 1992), driven by imbalanced DNA precursors, gives the equilibrium $G + C$ content of the sequence as

$$f = \frac{u_{(AT \to GC)}}{u_{(AT \to GC)} + u_{(GC \to AT)}} \tag{15}$$

where $u_{(AT \to GC)}$ is the total mutation rate from A or T to G or C, and $u_{(GC \to AT)}$ is vice versa—that is,

$$u_{(AT \to GC)} = u_{A \to G} + u_{T \to C} + u_{T \to G} + u_{A \to C}$$
$$u_{GC \to AT)} = u_{G \to A} + u_{C \to T} + u_{G \to T} + u_{C \to A}$$

Here the mutation rate from the correct nucleotide k to the incorrect nucleotide i, denoted by $u_{k \to i}$, can be written as

$$u_{k \to i} = \sigma_m(k \to i) \times \bar{\sigma}_n$$

($i = A, T, C, G, k \neq i$), where $\bar{\sigma}_n$ is the correction rate given by Eq. (13) and $\sigma_m(k \to i)$, the probability of misincorporation from the correct nucleotide k to the incorrect nucleotide i is given by Eq. (8).

Results

Equilibrium G + C *Content of Sequences*

Let $N = N_G + N_C$ be the proportions of *dGTP* and *dCTP* in the pool of nucleotide precursors. Furthermore, let us make the simplifying assumption that the concentrations of *dCTP* and *dGTP* are the same, and so are the concentrations of *dATP* and *dTTP*—that is, $N_G = N_C$ and $N_A = N_T$. Noting that the magnitude of α is less than 10^{-5} and substituting Eq. (8) into Eq. (15), we can simplify the expression for the equilibrium $G + C$ content as follows:

$$f = \frac{N^2[2+\beta(1-N)]}{2+(\beta-4)N(1-N)} \tag{16}$$

As shown in Fig. 2, the relationship between the equilibrium $G + C$ content (f) and the proportion of $G + C$ in nucleotide precursors (N) depends on only the parameter β. The curve is sigmoidal when $\beta = 0$, similar to the result given by Eyre-Walker (1992b). However, it approaches the line $f = N$ when $\beta \to \infty$. Actually, for $\beta = 20$, the curve is quite close to $f = N$. Since the value of β is fairly large, i.e., $\beta > 20$, the equilibrium $G + C$ content (f) is approximately equal to the proportion of $G + C$ in nucleotide precursors (N) in mammalian germ cells under the normal physiological conditions. This result confirms the intuitive statement on the equilibrium $G + C$ content by Wolfe et al. (1989).

Mutation Rate and G + C *Content*

First, let us define a new variable, H, as

$$H = 2(N_G + N_C)(N_A + N_T) = 2N(1 - N) \quad (17)$$

which is a measure for the distribution of DNA precursor concentrations: $0 \le H \le 0.5$. Substituting Eq. (15) into Eq. (11) and Eq. (13), respectively, we obtain the expressions for the rate of misincorporation $\bar{\sigma}_m$ and the rate of noncorrection $\bar{\sigma}_n$ when the $G + C$ content of the sequence is at equilibrium. After some algebraic simplifications, we arrive at

$$\bar{\sigma}_m = \frac{\alpha}{2}(1 + H)\left(\frac{8 + 4\beta + \beta^2 H}{4 + (\beta - 4)H}\right) \quad (18)$$

$$\bar{\sigma}_n = h_1 h_2 + \left(\frac{h_1(1 - h_2)b}{8 + 4b + b^2 H}\right)$$
$$\left(\frac{16 + 4(\beta + b - 6)H - (4\beta + 4b - b\beta)H^2}{4 + (\beta - 4)H}\right) \quad (19)$$

Since the mutation rate is $u = \bar{\sigma}_m \times \bar{\sigma}_n$, Eqs. (16) to (19) can give the expected relationship between the mutation rate u and the equilibrium $G + C$ content f. Let us consider the simplest case where the next-nucleotide effect is so weak that $\bar{\sigma}_n \approx h_1 h_2$ in Eq. (19). It occurs when the E_2 pathway is little affected by the relative concentration of the next correct precursors—that is, $b \to 0$. The mutation rate is then given by

$$u = \frac{\alpha}{2}h_1 h_2(1 + H)\left(\frac{8 + 4\beta + \beta^2 H}{4 + (\beta - 4)H}\right) \quad (20)$$

In combination with Eq. (16) and Eq. (17), Eq. (20) shows that the mutation rate u reaches its maximum with $f = N = H = 0.5$, and reduces to its minimum when $f = 0$ or $f = 1$. Let R be the relative mutation rate, which is defined by

$$R = \frac{u - u_0}{u_{0.5} - u_0} \quad (21)$$

where u_0 is the mutation rate when $f = N = H = 0$ and $u_{0.5}$ is the mutation rate when $f = N = H = 0.5$—that is,

$$u_0 = \frac{1}{2}\alpha h_1 h_2(2 + \beta)$$

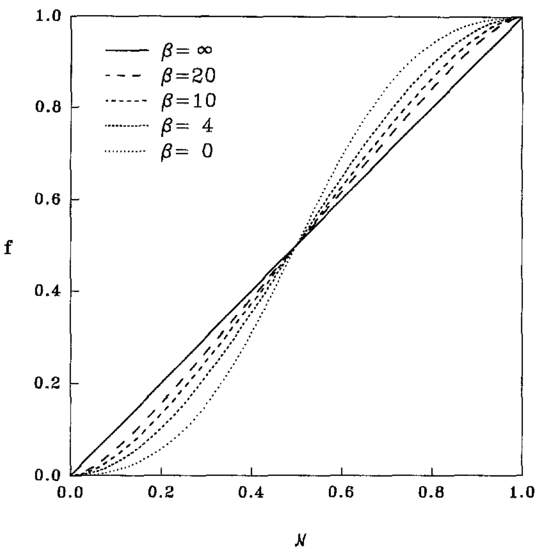

Fig. 2. The relation between the equilibrium $G + C$ content in a sequence (f) and the $G + C$ proportion in the nucleotide precursor pool (N), given $\beta = 0, 4, 10, 20$, and ∞.

$$u_{0.5} = \frac{3}{4}\alpha h_1 h_2(4 + \beta)$$

The relation between the relative mutation rate R and the equilibrium $G + C$ content of the sequence is illustrated in Fig. 3, for $\beta = 0, 4, 10, 20$, and ∞. As β increases, the curve becomes more inverted-V-shaped and eventually

$$R \to 4f(1 - f)$$

when $\beta \to \infty$. Since the real value of β is fairly large (i.e., $\beta > 20$), changes in the relative concentrations of nucleotide precursors during the germ-cell cycle lead to an inverted-V-shaped correlation of mutation rate with $G + C$ content if the next-nucleotide effect is weak.

The Next-Nucleotide Effect

Now we consider the next-nucleotide effect on mutation rate in some detail. As discussed above, since the value of b is less than 1 and β is larger than 20, the condition that $b \ll \beta$ holds. Therefore Eq. (19) is simplified to

$$\bar{\sigma}_n \approx h_1 h_2 + h_1 h_2 \epsilon\left(\frac{4 + (\beta - 6)H - (1 + \beta)H^2}{4 + (\beta - 4)H}\right) \quad (22)$$

where the coefficient of the next-nucleotide effect ϵ is defined by

$$\epsilon = \frac{b(1 - h_2)}{(2 + b)h_2} \quad (23)$$

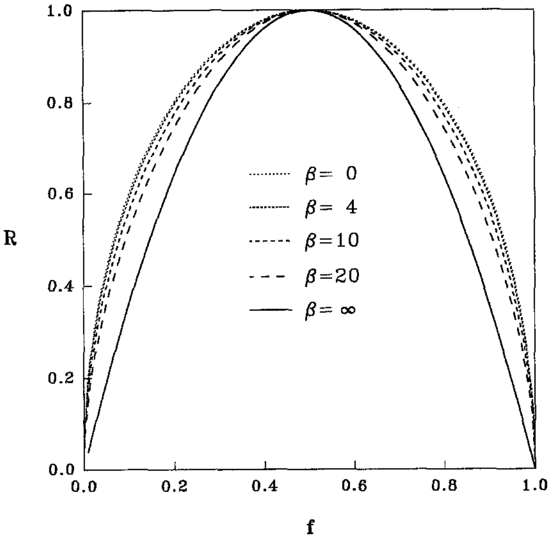

Fig. 3. Distribution of the relative mutation rate (R) with respect to the $G + C$ content (f) when the next-nucleotide effect is weak (i.e., $\epsilon = 0$ or $MC = 1$), given $\beta = 0, 4, 10, 20$, and ∞.

The mutation rate u then can be expressed as

$$u = \frac{\alpha}{2} h_1 h_2 (1 + H)\left(\frac{8 + 4\beta + \beta^2 H}{4 + (\beta - 4)H}\right)\left(1 + \epsilon\frac{4 + (\beta - 6)H - \beta H^2}{4 + (\beta - 4)H}\right) \quad (24)$$

In order to obtain a biological interpretation for the strength of the next-nucleotide effect, let MC be the maximum capacity of the next-nucleotide effect, which is defined as the ratio of the maximum to the minimum of $\sigma_n(k)$, with respect to the relative concentrations of precursors. From Eq. (12), it can be shown that

$$MC = \frac{max(\sigma_n(k))}{min(\sigma_n(k))} = \frac{h_2 + b}{h_2(1 + b)} = \frac{1 + (2 + b)\epsilon}{1 + b} \quad (25)$$

given $h_2 \ll 1$. If $b \ll 1$, we have

$$MC \approx 1 + 2\epsilon \quad (26)$$

Therefore, the strength of the next-nucleotide effect is measured by the maximum number of folds that the noncorrection probability ($\sigma_n(k)$) can be increased by the next-nucleotide effect. In the case of $\epsilon \to 0$, or $MC \to 1$, which indicates a weak next-nucleotide effect, Eq. (24) reduces to Eq. (20). As also indicated by Wolfe (1991) and Eyre-Walker (1992b), the general conclusion is that the curve of the mutation rate vs the $G + C$ content of the sequence is flattened by an increase of the next-nucleotide effect, because the effect elevates the mutation rate of the sequence at extreme $G + C$ contents compared to that at an intermediate $G + C$ content.

Figure 4 shows the relationship between the relative mutation rate (R) and the $G + C$ content (f) for $\beta = 20$. When $\epsilon = 0$ ($MC = 1$), a strong correlation of mutation rate with $G + C$ content exists and leads to an inverted-V-shaped distribution. When $\epsilon = 0.5$ ($MC = 2$), a weaker correlation of the mutation rate with $G + C$ content leads to an inverted-U-shaped distribution. Numerical analysis indicates that for $\beta = 20$, the peak of mutation rate at $f = 0.5$ is readily observed when $\epsilon < 0.5$ ($MC < 2$).

In general, it can be shown that the relationship between the mutation rate (u) and the $G + C$ content of the sequence (f) is inverted-V(or U)-shaped, with the maximum value at $f = 0.5$, if the coefficient of the next-nucleotide effect ϵ is not larger than the critical value ϵ_0, i.e.,

$$\epsilon \leq \epsilon_0$$

where ϵ_0 is given by

$$\epsilon_0 = \frac{64 + 8\beta + \beta^2}{16 + 14\beta + \beta^2} \quad (27)$$

If $\epsilon > \epsilon_0$, the relationship between u and f is bimodal-like, with the local minimum value at $f = 0.5$ (the valley point; see Fig. 4). For $\beta = 20$, the critical value ϵ_0 is about 0.897.

The value of ϵ_0 depends on the parameter β. As can be seen from Eq. (27), ϵ_0 begins from $\epsilon_0 = 4$ when $\beta = 0$, decreases gradually until its minimum value $\epsilon_0 \approx 0.896$ when $\beta \approx 21.86$, and then increases to $\epsilon_0 = 1$ when $\beta \to \infty$. Therefore, we may conclude that the correlation of mutation rate with $G + C$ content is predicted to be inverted-U-shaped by the current model if

$$\epsilon \leq 0.896 \text{ or } MC \leq 2.8$$

under the normal physiological conditions in the mammalian germ cells.

Discussion

Origin of GC-*Rich Isochores*

We have shown above that under the normal physiological condition in germ cells, the $G + C$ content in a sequence (f) and the $G + C$ proportion in the nucleotide precursors (N) are approximately equal at equilibrium. This implies that any shift in the $G + C$ proportion of

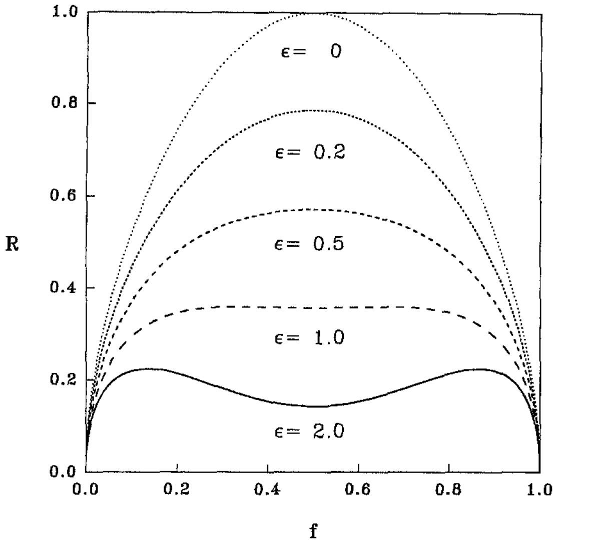

Fig. 4. The effect of the next-nucleotide concentration (ϵ) on the distribution of the relative mutation rate (R) with respect to the $G + C$ content (f), for $\beta = 20$.

nucleotide precursors can gradually lead to a corresponding shift in the $G + C$ content in the sequence. Furthermore, the conclusion that $f \approx N$ at equilibrium is independent of the strength of the next-nucleotide effect. Therefore, a DNA segment (i.e., a replicon) will become *GC* rich if it is replicated at a time when the nucleotide precursor pool is rich in *G* and *C*. Thus, *GC*-rich isochores can arise under the present model.

Mathews and Ji (1992) compiled the relative concentrations of nucleotide precursors in different mammalian cell lines. The data indicated that the $G + C$ proportions of nucleotide precursors during a cell cycle are roughly from 44% to 55%, and for most cases, they are below 50%. The study of Bernardi et al. (1985) indicated that the compositional distribution of isochores in the human genome is characterized by (1) two classes of *GC*-poor isochores (39% and 41% *GC*, respectively), representing about two-thirds of the genome and (2) three classes of *GC*-rich isochores (45%, 49%, and 53% *GC*, respectively), representing one-third of the genome. These observations can be explained by our model, given the data of Mathews and Ji (1992). Therefore, the mutationist hypothesis may be able to explain the origin of *GC*-rich isochores in the genome of mammals.

Inverted V-Shaped Distribution of Mutation Rates

It is interesting that our model predicts that an inverted-V distribution of mutation rates in relation to $G + C$ content occurs if the next-nucleotide effect is weak, i.e., $MC \rightarrow 1$. When the strength of the next-nucleotide effect is intermediate, i.e., $1 < MC < 2$, the curve becomes flatter so that the distribution is inverted-U-shaped, but, as illustrated in Fig. 4 for $\beta \geq 20$, the peak of mutation rate at 50% *GC* can be identified. This peak still exists theoretically when $2 < MC < 2.8$, but it eventually disappears when $MC > 2.8$. There are few data available for estimating the value of *MC* in mammalian germ cells. In spite of this, the fact that an inverted-U-shaped distribution of mutation rate with respect to $G + C$ content has been observed (Bulmer et al. 1991; Wolfe and Sharp 1993) implies that the value of *MC* in mammalian germ cells may be less than 2.

There is also biochemical evidence to suggest that *MC* is not very large. In *E. coli*, the probability of misincorporation is about 10^{-5}; the probability of no correction during replication (proofreading) is about 10^{-2}; and the probability of no postreplication repair is about 10^{-3} (Kunkel 1988, 1992a; Modrich 1991). Therefore, proofreading is actually less important than misincorporation and postreplication repair in determining the mutation rate in prokaryotes. In mammals, three DNA polymerases—α, δ, and ϵ—are shown to be involved in nuclear DNA replication. Only polymerases δ and ϵ contain an associated 3′-5′ exonuclease which is essential for the correction of misincorporations, but the role of these polymerases in eukarytic DNA replication is still not clear. However, the energy cost of correction for misincorporations by exonucleolytic activity might be unacceptably high due to too much excision of correct incorporations if more-than-100-fold fidelity is contributed by proofreading (Fersht 1985). Therefore, in mammals, as in prokaryotes, the mutation rate may not be greatly affected by the next-nucleotide effect.

However, in addition to the next-nucleotide effect, there are some other factors that can blur the inverted-V-shaped distribution. First, synonymous substitutions are only nearly selectively neutral. Bias of codon usage may occur in mammalian genomes, although it is difficult to detect. Consequently, synonymous rates are only rough estimates of mutation rates. Second, the substitution rate estimated from comparison between sequences is actually the average value over lineages. Therefore, when the estimated synonymous rate is from two relatively distantly related taxa, e.g., mammalian orders, the condition of equilibrium may not hold. Finally, estimates of synonymous substitution rates are subject to large stochastic errors, especially when the sequence length is short. Therefore, a failure to observe an inverted-V-(or U-)shaped distribution of substitution rates with respect to $G + C$ contents may not be disproof of the mutationist hypothesis.

Acknowledgments. We thank Dr. T.A. Kunkel, Dr. D. Hewett-Emmett, Dr. A. Eyre-Walker, and Dr. K.H. Wolfe for suggestions. This research is supported by NIH grants.

References

Bernardi G (1989) The isochore organization of the human genome. Annu Rev Genet 23:637–661

Bernardi G, Mouchiroud D, Gautier C (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol 28:7–28

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Bulmer M, Wolfe KH, Sharp PM (1991) Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. Proc Natl Acad Sci USA 88:5974–5978

Dresler SL, Frattini MG, Robinson-Hill RM (1988) In situ enzymology of DNA replication and ultraviolet-induced DNA repair synthesis in permeable human cells. Biochemistry 27:7247–7254

Echols H, Goodman MF (1991) Fidelity mechanisms in DNA replication. Annu Rev Biochem 60:477–511

Eyre-Walker A (1992a) Evidence that both $G + C$ rich and $G + C$ poor isochores are replicated early and late in the cell cycle. Nucleic Acids Res 20:1497–1501.

Eyre-Walker A (1992b) The role of DNA replication and isochores in generating mutations and silent substitution rate variance in mammals. Genet Res 60:61–67

Fersht (1985) Enzyme structure and mechanism, 2nd ed. WH Freeman, New York, p 308, 363–367

Goldman MA (1988) The chromatin domain as a unit of gene regulation. Bioessays 9:50–55

Holmquist GP (1987) Role of replication time in the control of tissue of specific gene expression. Am J Hum Genet 40:151–173

Holmquist GP (1988) DNA sequences in G-bands and R-bands. In: Adolph KW (ed) Chromosomes and chromatin. CRC Press, Boca Raton, p 76

Holmquist GP (1992) Chromosome bands, their chromatin flavors, and their functional features. Am J Hum Genet 51:17–37

Ikemura T, Aota S (1988) Global variation in G + C content along vertebrate genome DNA: possible correlation with chromosome band structures. J Mol Biol 203:1–13

Kunkel TA (1988) Exonucleolytic proofreading. Cell 53:837–840

Kunkel TA (1992a) DNA replication fidelity. J Biol Chem 267: 18251–18254

Kunkel TA (1992b) Biological asymmetries and the fidelity of eukaryotic DNA replication. Bioessays 14:303–308

Kunz BA, Kohalmi SE (1991) Modulation of mutagenesis by deoxyribonucleotide levels. Annu Rev Genet 25:339–359

Leeds JM, Slabaugh MB, Mathews CK (1985) DNA precursor pools and ribonucleotide reductase activity: distribution between the nucleus and the cytoplasm of mammalian cells. Mol Cell Biol 5:3443–3450

Mathews CK, Ji J (1992) DNA precursor asymmetries, replication fidelity, and variable genome evolution. Bioessays 14:295–301

Mendelman LV, Petruska J, Goodman MF (1990) Base mispair extension kinetics. J Biol Chem 265:2338–2346

Modrich P (1991) Mechanisms and biological effects of mismatch repair. Annu Rev Genet 25:229–253

Ninio J (1987) Kinetics devices in protein synthesis, DNA replication, and mismatch repair. Cold Spring Harbor Symp Quant Biol 52: 639–646

Sueoka N (1988) Directional mutation pressure and neutral molecular evolution. Proc Natl Acad Sci USA 85:2653–2657

Sueoka N (1992) Directional mutation pressure, selective constraints, and genetic equilibria. J Mol Evol 34:95–114

Wolfe KH (1991) Mammalian DNA replication: mutation biases and the mutation rate. J Theor Biol 149:441–451

Wolfe KH, Sharp PM (1993) Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. J Mol Evol 37:441–456

Wolfe KH, Sharp PM, Li WH (1989) Mutation rate differ among regions of the mammalian genome. Nature 337:283–285

Wong I, Patel SS, Johnson KA (1991) An induced-fit kinetic mechanism for DNA replication fidelity: direct measurement by single-turnover kinetics. Biochemistry 30:526–537