The Covariation Between TpA Deficiency, CpG Deficiency, and G+C Content of Human Isochores Is Due to a Mathematical Artifact

Laurent Duret and Nicolas Galtier†*

*Laboratoire de Biométrie, Génétique et Biologie des Populations, Université Claude Bernard, Villeurbanne, France; and †Laboratoire Génome, Populations, Interactions, Université Montpellier 2, Montpellier, France

CpG and TpA dinucleotides are underrepresented in the human genome. The CpG deficiency is due to the high mutation rate from C to T in methylated CpG's. The TpA suppression was thought to reflect a counterselection against TpA's destabilizing effect in RNA. Unexpectedly, the TpA and CpG deficiencies vary according to the G+C contents of sequences. It has been proposed that the variation in CpG suppression was correlated with a particular chromatin organization in G+C-rich isochores. Here, we present an improved model of dinucleotide evolution accounting for the overlap between successive dinucleotides. We show that an increased mutation rate from CpG to TpG or CpA induces both an apparent TpA deficiency and a correlation between CpG and TpA deficiencies and G+C content. Moreover, this model shows that the ratio of observed over expected CpG frequency underestimates the real CpG deficiency in G+C-rich sequences. The predictions of our model fit well with observed frequencies in human genomic data. This study suggests that previously published selectionist interpretations of patterns of dinucleotide frequencies should be taken with caution. Moreover, we propose new criteria to identify unmethylated CpG islands taking into account this bias in the measure of CpG depletion.

Introduction

The frequency of occurrence of dinucleotides within genomes has received much attention because some of them significantly depart their expectations with respect to base composition. Classically, these deviations are measured by the ratio of observed over expected dinucleotide frequency (XpYo/e = $d_{XY}/n_X \cdot n_Y$, where n_X denotes the frequency of nucleotide X, and d_{XY} is the frequency of the dinucleotide XpY). It is well known that in vertebrate species the frequency of dinucleotide CpG is much (up to fivefold) lower than the product of C and G frequencies (Bird 1980). Although the deviation is less pronounced, the TpA dinucleotide is also significantly underrepresented in vertebrate genomes (Hanai and Wada 1988; Beutler et al. 1989; Karlin and Mràzek 1997). In mammals, the CpG deficiency is the consequence of a mutational bias. CpG is the target of DNA-methyltransferase activity, resulting in the methylation of cytosine. Methylated CpG's have a high mutation rate toward TpG (or CpA on the complementary strand), decreasing the CpG frequency. In the regions that escape methylation in the germ line (e.g., the promoter regions of housekeeping genes), CpG dinucleotides are less suppressed. As a consequence, these regions (the so-called ''CpG islands'') appear relatively CpG-rich compared with the rest of the genome (Antequera and Bird 1999).

The reason for the TpA scarcity is not clearly understood. However, UpA appears to be a preferential target for ribonucleases (Beutler et al. 1989). Moreover, Beutler et al. (1989) noticed that TpA is more stringently excluded in DNA destined to be expressed in the cytosol (exons of protein-coding genes and tRNA and rRNA genes) than in nontranscribed Y-chromosomal DNA, DNA that is expressed only in mitochondria, and DNA that is degraded within the nucleus (intron DNA). This led the authors to propose that, by reason of their instability, there was a selective pressure against UpA dinucleotides in mRNA, tRNA, or rRNA sequences.

Unexpectedly, the deficiencies in CpG and TpA dinucleotides, measured by the ratio of observed to expected dinucleotide frequency (CpGo/e, TpAo/e), varies according to the G+C contents of human genes: CpG depletion is lower and TpA depletion higher in G+C-rich than in G+C-poor genes (Hanai and Wada 1988). The same trend has been observed within genes: both the G+C content and the CpGo/e ratio are higher in 5′ untranslated regions (UTRs) than in 3′ UTRs (Pesole et al. 1997). On a larger scale, it has been shown that CpGo/e is higher in G+C-rich parts of the genome (G+C-rich isochores) than in G+C-poor regions (Bernardi et al. 1985; Aissani and Bernardi 1991; Jabbari and Bernardi 1998). Interestingly, these correlations (positive and negative, respectively) between sequence CpGo/e or TpAo/e and G+C content have also been found in RNA viruses (Rima and McFerran 1997). However, the reason for these correlations was not established.

We propose that the observed TpA deficiency (on one hand) and the observed correlations between G+C content, CpG deficiency, and TpA deficiency (on the other hand) are essentially indirect consequences of the mutational CpG depletion. We first present an intuitive argument explaining the reasons for these effects, and we then quantify them through an improved model of dinucleotide evolution that accounts for overlaps between successive dinucleotides.

Intuitive Argument

Consider a random sequence with independent dinucleotide frequencies (i.e., no neighboring effect). Now suppose that many CpGs change to either TpG or CpA.

Key words: dinucleotides, CpG, TpA, CpG islands, methylation, isochores.

Address for correspondence and reprints: Laurent Duret, Laboratoire de Biométrie, Génétique et Biologie des Populations, UMR CNRS 5558, Université Claude Bernard, 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne cedex, France. E-mail: duret@biomserv.univ-lyon1.fr.

Mol. Biol. Evol. 17(11):1620–1625. 2000

Table 1
Impact of CpG Mutations on Base Composition and Dinucleotide Deficiency (observed/expected frequency) in a G+C-Poor and a G+C-Rich Sequence

	G+C-Poor Sequence		G+C-Rich Sequence	
	Before Mutation[a]	After Mutation[a]	Before Mutation[a]	After Mutation[a]
Base frequency				
A	0.30	0.31	0.20	0.23
T	0.30	0.31	0.20	0.23
C	0.20	0.19	0.30	0.27
G	0.20	0.19	0.30	0.27
G+C	0.40	0.37	0.60	0.54
Fraction of mutated CpG	0.67		0.67	
Number of mutated CpG's	27		60	
CpG				
Observed	40	13	90	30
Expected	40.00	34.82	90.00	72.82
CpGo/e	1.00	0.38	1.00	0.41
TpA				
Observed	90	90	40	40
Expected	90.00	98.22	40.00	52.97
TpAo/e	1.00	0.92	1.00	0.76

[a] A random sequence (1,000 nt) is mutated by changing 67% of its CpG's to TpG or CpA. The base composition and the observed and expected numbers of dinucleotides are indicated before and after mutation of CpG's.

This would increase the number of A's and T's. However, this would not change the number of TpAs and would not generate new CpGs, since new dinucleotides arising from such changes can only be NpT, TpG, CpA, and ApN (where N is A, C, G, or T). Thus, this CpG depletion has three noteworthy consequences. First, since the number of T's and A's, but not the number of TpAs, increases, the ratio TpAo/e ($d_{TA}/n_T \cdot n_A$) tends to decrease. Second, since the number of C's and G's has decreased, the expected number of CpGs in the new sequence is smaller than the number of CpGs in the original sequence. Therefore, the CpGo/e ratio underestimates the real CpG depletion. Finally, both effects are enhanced when the G+C content is high: the relative increase in A and T frequencies induced by CpG depletion is higher, resulting in an increased TpA deficiency and a decrease in the expected number of CpGs.

To give an idea of the impact of CpG mutations, we calculated the G+C content and the observed over expected dinucleotide frequencies in a random sequence where 67% of CpG's would have been changed to TpG or CpA. We calculated these values for G+C contents (before CpG mutations) of 40% and 60%. As shown in table 1, in both cases, CpG depletion induces an apparent TpA deficiency. The CpGo/e ratio is higher than the ratio of final/initial CpG frequencies (0.33). Thus, the CpGo/e ratio underestimates the real mutation pressure on CpG dinucleotides. Finally, the CpGo/e and TpAo/e ratios are, respectively, higher and lower in the G+C-rich than in the G+C-poor sequence, in agreement with observations from the human genome.

A Model of Dinucleotide Evolution

The above approach is obviously simplistic because it does not take into account the dynamics of the mutation process. Therefore, we developed a model to simulate the evolution of dinucleotide content in a sequence where CpG dinucleotides are under mutation pressure toward TpG or CpA. Nucleotide evolution is usually modeled as a four-state Markov chain, where each change from one state to another is assigned a rate in continuous time (e.g., see Yang 1995). Sved and Bird (1990) applied this approach to dinucleotides using a 16-state Markov chain. Their model, however, did not account for the overlap between successive dinucleotides. The overlap induces dependencies among dinucleotide frequencies that have important effects. A consequence of neglecting the overlap is that the sum of the equilibrium frequencies of, say, ApN dinucleotides is different from the sum of the equilibrium frequencies of NpAs in Sved and Bird's (1990) study (their table 2), while both should be equal to the frequency of A.

Table 2
Base Composition and Dinucleotide Deficiency (observed/expected frequency) in Coding and Noncoding Regions of Human Genes

	G+C%	TpAo/e	CpGo/e
Coding regions	55 ± 8	0.51 ± 0.1	0.46 ± 0.1
Introns	49 ± 9	0.68 ± 0.1	0.27 ± 0.1
3′ UTR	48 ± 11	0.64 ± 0.2	0.24 ± 0.1

NOTE.—Average values ± SD. $N = 545$ human genes for which the complete genomic sequence (introns and exons, including the entire 3′ UTR) was available in GenBank.

The simple model of dinucleotide evolution that we present here accounts for the overlap between successive dinucleotides. This model involves three parameters. It is assumed that any nucleotide not currently involved in a CpG doublet evolves according to Tamura's (1992) model. This model has two parameters, namely, the transition/transversion ratio κ and the equilibrium G+C content θ, i.e., the G+C content that would be reached if a sequence evolved according to this process during a very long time. Under Tamura's (1992) model, a change from anything to G or C occurs at rate θ if it is a transversion and at rate $\kappa\theta$ if it is a transition, while changes to A or T have rates of $1 - \theta$ or $\kappa(1 - \theta)$. Nucleotides involved in a CpG doublet also evolve according to Tamura's (1992) model but with an increased transition rate κ_1. No other assumption is made about neighboring effects. When κ_1 is equal to κ, the model reduces to Tamura's (1992) model, and the equilibrium G+C content is θ. For $\kappa_1 > \kappa$, however, the equilibrium G+C content is lower than θ, since C's and G's involved in CpG doublets change to T or A at a higher rate than the reverse. Let $\mathbf{D}(u) = (d_{ij}(u))$ be the 4×4 matrix of dinucleotide frequencies at time u. Nucleotide frequencies $\mathbf{N}(u) = (n_i(u))$ derive from $\mathbf{D}$: for any j, $n_j = \Sigma_j d_{ij} = \Sigma_j d_{ji}$. To determine how the above-described evolutionary forces apply on $\mathbf{D}$, note that the instantaneous rates of change of any nucleotide depend only on its state and the states of its two neighbors. If $\mathbf{T}(u) = (t_{ijk}(u))$ is the matrix of trinucleotide frequencies, one can write:

$$d_{xy}(u + \mathrm{d}u) = d_{xy}(u) + \sum_i \sum_j \sum_k t_{ijk}(u) \times \sum_m r(i, j \rightarrow m, k) \times b((x, y), (i, j \rightarrow m, k))\mathrm{d}u. \quad (1)$$

In equation (1), $r(i, j \rightarrow m, k)$ is the rate of change from trinucleotide (*ijk*) to trinucleotide (*imk*), deductible from the model. For instance, r(A, A→C, T) equals θ, and r(A, C→T, G) equals $\kappa_1.(1 - \theta)$. Factor b((*x*, *y*), (*i*, *j*→*m*, *k*)) in equation (1) is the balance for dinucleotide (*xy*) when a (*ijk*) to (*imk*) change occurs, i.e., the difference between the number of (*xy*) dinucleotides included in trinucleotide (*imk*) and the number of (*xy*) dinucleotides included in (*ijk*). For instance, b[(A, C), (A, C→T, G)] is −1 (one AC is lost by changing ACG to ATG), b[(A, A), (A, G→A, A)] is 2, and b[(A, A), (C, C→T, C)] is 0. In words, equation (1) states that the overall change for dinucleotide (*xy*) is the sum over all trinucleotides (*ijk*) and all possible changes for *j* of the frequency of that trinucleotide times the probability of that change times the effect of that change on (*xy*) occurrence. Trinucleotide frequencies can be deduced from dinucleotide ones:

$$t_{ijk}(u) = \frac{d_{ij}(u)d_{jk}(u)}{n_j(u)}, \quad (2)$$

where $n_j = \Sigma_j\, d_{ij} = \Sigma_j\, d_{ji}$ is the frequency of nucleotide *j*. Equation (2) assumes that dependencies do not extend farther than two bases, i.e., that the probability of the state of one nucleotide depends only on its neighbors. We checked this approximation from simulations and found it to be very good.

Equation (1) written for all possible (*x*, *y*) forms a system of 16 differential equations that describe the instantaneous dynamics of dinucleotide frequencies under our model. This system can hardly be solved analytically—in contrast to the analogous system in models describing nucleotide evolution—essentially because the expression of t_{ijk} includes products between d_{ij}'s, making it nonlinear. However, equation (1) allows one to quickly simulate the evolution of dinucleotides and to deduce equilibrium frequencies given θ, κ, and κ_1. The simulation process is the following:

1. Start from dinucleotide frequencies **D**(0).
2. Deduce nucleotide **N**(0) and trinucleotide **T**(0) frequencies.
3. Compute **D**(d*u*) for some arbitrary small d*u* using equation (1).
4. Deduce **N**(d*u*) and **T**(d*u*) and iterate until equilibrium is achieved.

Results

We made use of this process to compare the predictions of our model with human DNA sequence data. G+C content is variable across the human genome according to the so-called ''isochore structure'' (Bernardi et al. 1985). In contrast, there is no evidence that transition/transversion ratios within or outside CpG doublets vary in distinct parts of the genome. Therefore, we used fixed values for parameters κ and κ_1 but allowed θ to vary. Estimates for κ (2.1) and κ_1 (27.6) were obtained from a study of hemophilia B polymorphism in the United Kingdom (Giannelli, Anagnostopoulos, and Green 1999) and were largely concordant with the data of Cargill et al. (1999) and Halushka et al. (1999).

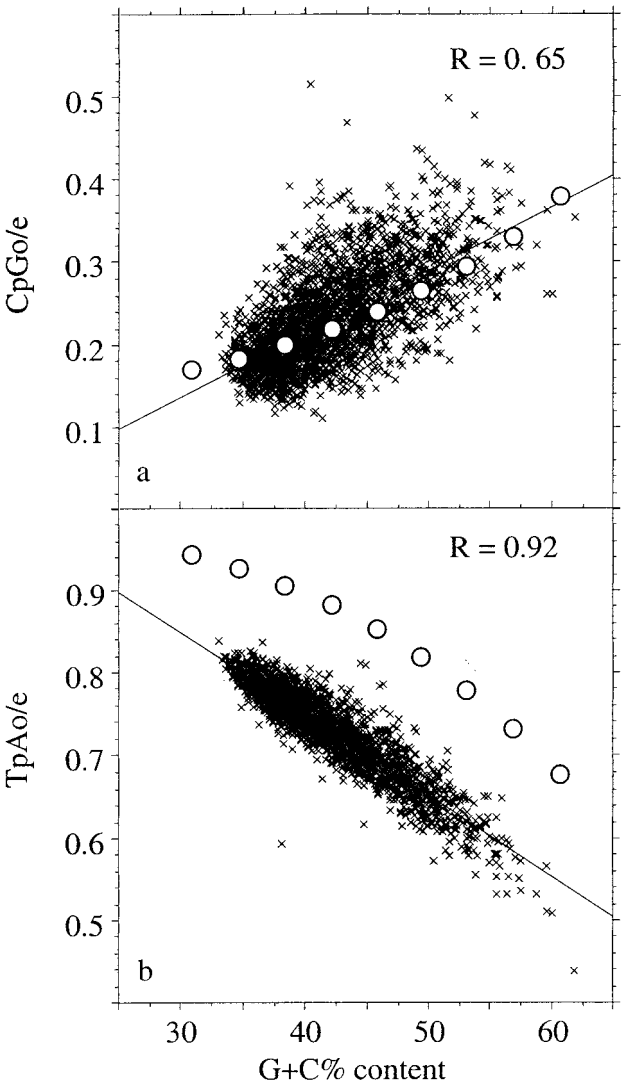

FIG. 1.—Correlation between G+C content and the ratio of observed/expected TpA and CpG dinucleotide frequencies in the human genome. Observed frequencies are in 3,073 genomic sequences longer than 100 kb. Solid lines represent linear regression, and open circles represent frequencies predicted according to the model.

Figure 1 displays the relationship between CpG and TpA depletion (observed/expected frequencies) and G+C content in 3,073 human DNA sequences longer than 100 kb. Sequences were retrieved from GenBank release 115 (December 15, 1999) using the ACNUC database (Gouy et al. 1985). As previously reported with smaller data sets, a significant correlation between CpGo/e and G+C content was found: CpG deficiency was lower in G+C-rich regions. A moderate TpA deficiency also appeared, correlated with G+C content as well (but negatively). The regression lines are shown.

We simulated the evolution of dinucleotide frequencies until equilibrium was achieved using the empirical estimates for κ and κ_1 indicated above and various θ values. The relationship between the predicted equilibrium G+C content and equilibrium CpG and TpA deficiencies under the model assumptions are shown in figure 1 (open circles). Figure 1*a* shows a very good fit between CpG frequencies in observed and simulated sequences. Note that we did not even try to fit the parameters to the data: we used empirical estimates of mutational rates. This suggests that this model correctly represents CpG evolution. In particular, the correlation between CpG depletion and G+C content is predicted by the model.

Interestingly, our model also predicts some TpA deficiency at equilibrium, although no specific mutational mechanism has been assumed with respect to TpA dinucleotides. Moreover, in agreement with the observation on real sequences, our model predicts a negative correlation between TpAo/e and G+C content (fig. 1*b*). Note that the slope of the correlation is the same in real and in simulated sequences. Thus, variations of the TpAo/e ratio according to the G+C content are probably simply a direct consequence of CpG depletion. Using a related approach but a different model, Bulmer (1986) did not predict any TpA deficiency. One should note, however, that the TpAo/e ratio is lower in real sequences than expected according to our model. Thus, other factors contribute to the deficiency of TpA in human sequences.

In summary, these simulations confirm the qualitative predictions of the simplistic example presented in table 1. In agreement with real data, our model predicts that: (1) an increased mutation rate from CpG to TpG and CpA induces an apparent depletion in TpA, (2) this apparent TpA depletion increases with G+C content, and (3) the CpGo/e ratio underestimates the real mutation pressure on CpG dinucleotides, all the more as the sequence is G+C-rich; as a consequence, (4) CpGo/e and TpAo/e are correlated (positively and negatively, respectively) to G+C content.

Discussion

Our model has several implications regarding the biological significance of the variations of the CpGo/e and TpAo/e ratios along the genome. First, since the CpG deficiency is related to the level of methylation, and since the level of methylation is correlated with chromatin organization and gene expression, it has been proposed that the positive correlation between CpGo/e and G+C might reflect a particular chromatin organization associated with transcriptionally active DNA or some differences in the regulation of gene expression in G+C-rich compared with G+C-poor isochores (Bernardi et al. 1985; Aissani and Bernardi 1991; Jabbari and Bernardi 1998). However, the fit between our simulations and the data observed on real sequences suggests that this correlation is simply due to the fact that the ratio CpGo/e underestimates the deficiency in CpG dinucleotides in G+C-rich regions. Therefore, there is no evidence for variation in the rate of CpG hypermutation with respect to the isochore G+C content, and the apparent reduced CpG deficiency in G+C-rich isochores is simply a mathematical artifact.

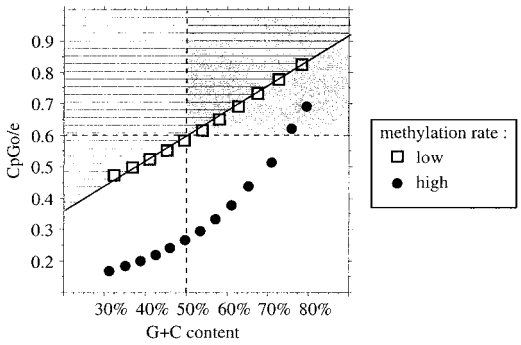

FIG. 2.—Correlation between G+C content and the observed/expected CpG dinucleotide ratio predicted by our model for different rates of CpG methylation: a high rate (κ_1 = 27.6, circles) and a low rate (κ_1 = 8.0, squares). The regression line for the low rate of methylation is indicated (R^2 = 0.997). The area above the regression line (hatched) corresponds to undermethylated sequences. The gray rectangle corresponds to sequences that would be predicted to be unmethylated CpG islands according to the criteria classically used in the literature (Gardiner-Garden and Frommer 1987).

Second, this result raises the question of the definition of CpG islands. The essential feature of CpG islands is the absence of methylation (at least in the germ line). Originally, CpG islands were identified as genomic regions that were very rich in cleavable sites for mCpG-sensitive restriction enzymes (Bird 1986). The sequencing of these unmethylated regions revealed relatively high G+C contents and CpGo/e ratios. These features have been used to detect CpG islands by computer analysis of genomic sequences. Classically, CpG islands are identified as DNA regions (≥200 bp) with a G+C content higher than 50% and a CpGo/e ratio higher than 0.6 (Gardiner-Garden and Frommer 1987). Are these criteria relevant to identify all nonmethylated islands? The relatively high G+C content in CpG islands can be explained in part by the fact that CpG depletion tends to decrease the G+C content in the rest of the genome. However, it is also possible that this latter property reflects a bias in the original method for the detection of unmethylated DNA regions: the recognition sites for mCpG-sensitive restriction enzymes contain at least 50% G+C and hence are more frequent in G+C-rich than in G+C-poor DNA. It is therefore not clear whether this latter criterion is necessary to identify unmethylated DNA regions. Indeed, it has been shown that nonmethylated islands in fish genomes are G+C-poor (Cross et al. 1991). The CpGo/e ratio reflects the mutability of CpG's and thus is an indicator of the level of methylation in the germ line. However, as we have shown, in G+C-rich regions, this ratio underestimates the real CpG depletion. According to our model, a CpGo/e ratio of 0.6 with a G+C content of 50% corresponds to a rate of transition at the CpG doublet about 3.5 times as low as that in the rest of the genome (i.e., κ_1 = 8.0). Figure 2 displays the CpGo/e ratio predicted by our model for different G+C contents and for two

values of κ_1: high methylation rate ($\kappa_1 = 27.6$, genomic average rate) and low methylation rate ($\kappa_1 = 8.0$, CpG islands rate). This figure shows that according to the classical criteria (CpGo/e $\geq$ 0.6, G+C $\geq$ 50%), a highly methylated G+C-rich region (>70%) would be erroneously considered as a CpG island. Conversely, an undermethylated G+C-poor region (<50%) would not be identified as a CpG island. We therefore suggest that the criteria to identify CpG islands should be set according to the G+C-content of sequences. According to our simulations (fig. 2), the threshold of CpGo/e as a function of G+C frequency to assess the presence of unmethylated islands can be calculated with the following formula:

$$\text{CpGo/e} \geq 0.206 + 0.80\text{GC}. \quad (3)$$

This formula is consistent with the standard definition of CpG islands when the G+C content equals 50%, and it accounts for the CpG/GC relationship when the G+C content is different from 50%.

Third, Beutler et al. (1989) have noted that the TpAo/e ratio is lower in exons of protein-coding genes and tRNA and rRNA genes than in Y-chromosomal DNA, mitochondrial DNA, and introns. They also have shown that UpA has a destabilizing effect on RNAs. This led them to propose that a selective pressure against TpA is acting in DNA sequences destined to be expressed in the cytosol. However, it should be noted that protein-coding regions and tRNA or rRNA genes are characterized by a relatively high G+C content (on average, 55%–60%) compared with the other sequences they analyzed (less than 49% in introns, 44% in mitochondria, and 39% in chromosome Y genomic sequences). According to our model, these differences in G+C content could explain the differences in TpAo/e. To directly test the selectionist hypothesis proposed by Beutler et al. (1989), we compared the TpAo/e ratios in coding regions, introns, and 3′ UTRs of human genes. Coding regions and 3′ UTRs are part of the mRNA (and hence are destined to be expressed in the cytosol), whereas introns are not. Therefore, according to the selectionist hypothesis, TpAo/e should be lower in coding regions and 3′ UTRs than in introns. On the other hand, coding regions are relatively G+C-rich compared with introns and 3′ UTRs. Therefore, according to our model, TpAo/e should be lower in coding regions than in 3′ UTRs and introns. As shown in table 2, the data fit with our model and not with the selectionist hypothesis.

Since human genomic DNA is essentially nontranscribed, another prediction of the selectionist hypothesis is that TpAo/e should be lower in 3′ UTRs and introns than in genomic sequences. On the contrary, we found that the average TpAo/e ratios in 3′ UTRs (0.64 ± 0.20) and introns (0.68 ± 0.13) were very close to those of large (>100 kb) genomic sequences of similar base composition (0.67 and 0.66, respectively, calculated according to the regression slope presented in fig. 1).

Therefore, contrary to what has been proposed (Beutler et al. 1989) there is no evidence that TpA dinucleotides are more counterselected in exons than in introns or in transcribed than in nontranscribed DNA. As shown with our model, CpG depletion induces an apparent TpA depletion that depends on the G+C content of sequences. The differences in the TpAo/e ratios between the different sequences analyzed by Beutler et al. (1989) are merely a consequence of their differences in G+C content. However, as mentioned previously, the CpG depletion does not totally explain the observed TpA deficiency in the human genome. Karlin and Mràzek (1997) proposed that the deficiency in TpA might be due to its low thermodynamic stacking energy in DNA. They also suggested that because of the presence of TpA as part of many regulatory signals (e.g., TATA box, polyadenylation signal), TpA suppression might help to avoid inappropriate binding of regulatory factors. Although these are useful working hypotheses, they are still speculative, and the reason for the TpA depletion remains to be determined.

Acknowledgments

We thank two anonymous referees for their helpful comments. This work was supported by the Centre National de la Recherche Scientifique.

LITERATURE CITED

AISSANI, B., and G. BERNARDI. 1991. CpG islands: features and distribution in the genome of vertebrates. Gene **106**: 173–183.

ANTEQUERA, F., and A. BIRD. 1999. CpG islands as genomic footprints of promoters that are associated with replication origins. Curr. Biol. **9**:R661–R667.

BERNARDI, G., B. OLOFSSON, J. FILIPSKI, M. ZERIAL, J. SALINAS, G. CUNY, M. MEUNIER-ROTIVAL, and F. RODIER. 1985. The mosaic genome of warm-blooded vertebrates. Science **228**:953–958.

BEUTLER, E., T. GELBART, J. H. HAN, J. A. KOZIOL, and B. BEUTLER. 1989. Evolution of the genome and the genetic code: selection at the dinucleotide level by methylation and polyribonucleotide cleavage. Proc. Natl. Acad. Sci. USA **86**:192–196.

BIRD, A. P. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res. **8**:1499–1504.

———. 1986. CpG-rich islands and the function of DNA methylation. Nature **321**:209–213.

BULMER, M. 1986. Neighboring base effects on substitution rates in pseudogenes. Mol. Biol. Evol. **3**:322–329.

CARGILL, M., D. ATSHULER, J. IRELAND et al. (17 co-authors). 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat. Genet. **22**:231–238.

CROSS, S., P. KOVARIK, J. SCHMIDTKE, and A. BIRD. 1991. Non-methylated islands in fish genomes are GC-poor. Nucleic Acids Res. **19**:1469–1474.

GARDINER-GARDEN, M., and M. FROMMER. 1987. CpG islands in vertebrate genomes. J. Mol. Biol. **196**:261–282.

GIANNELLI, F., T. ANAGNOSTOPOULOS, and P. M. GREEN. 1999. Mutation rates in human. II. Sporadic mutation-specific rates and rate of detrimental human mutations inferred from Hemophilia B. Am. J. Hum. Genet. **65**:1580–1587.

GOUY, M., C. GAUTIER, M. ATTIMONELLI, C. LANAVE, and G. DI PAOLA. 1985. ACNUC, a portable retrieval system for nucleic acid sequences databases: logical and physical designs and usage. Comp. Appl. Biosci. **1**:167–172.

HALUSHKA, M. K., J.-B. FAN, K. BENTLEY, L. HSIE, N. SHEN, A. WEDER, R. COOPER, R. LIPSHUTZ, and A. CHAKRAVARTI.

1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat. Genet. **22**:239–247.

HANAI, R., and A. WADA. 1988. The effects of guanine and cytosine variation on dinucleotide frequency and amino acid composition in the human genome. J. Mol. Evol. **27**:321–325.

JABBARI, K., and G. BERNARDI. 1998. CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. Gene **224**:123–128.

KARLIN, S., and J. MRÀZEK. 1997. Compositional differences within and between eukaryotic genomes. Proc. Natl. Acad. Sci. USA **94**:10227–10232.

PESOLE, G., S. LUINI, G. GRILLO, and C. SACCONE. 1997. Structural and compositional features of untranslated regions of eukaryotic mRNAs. Gene **205**:95–102.

RIMA, B. K., and N. V. MCFERRAN. 1997. Dinucleotide and stop codon frequencies in single-stranded RNA viruses. J. Gen. Virol. **78**:2859–2870.

SVED, J., and A. BIRD. 1990. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc. Natl. Acad. Sci. USA **87**:4692–4696.

TAMURA, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol. Biol. Evol. **9**:678–687.

YANG, Z. 1995. On the general reversible Markov process model of nucleotide substitution: a reply to Saccone et al. J. Mol. Evol. **41**:254–255.

EDWARD HOLMES, reviewing editor

Accepted June 27, 2000