Selection against CpG dinucleotides in lentiviral genes: a possible role of methylation in regulation of viral expression

Eugene G.Shpaer and James I.Mullins
Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA 94305-5402, USA

Received May 24, 1990; Revised and Accepted August 23, 1990

ABSTRACT

Extremely low frequencies of CpG dinucleotides are found in the genomes of the lentivirus subfamily of retroviruses, including the human, simian and feline immunodeficiency viruses (HIV1, HIV2, SIV, and FIV, respectively), equine infectious anemia virus (EIAV), and the ovine lentivirus, Visna. The occurrence of CpG dinucleotides is greater in the 2-3 (NCG) than in the 1-2 (CGN) codon-defined frame, as well as in the *gag* and *env* genes, compared to the more conserved *pol* gene. These differences suggest that CpG depletion in lentiviruses occurs as a result of selection against CpG rather than due to mutational bias, the latter is responsible for low CpG frequencies in vertebrate genomes. CpG levels in the onco-retrovirus subfamily are reduced to a lesser extent, principally due to mutational bias. The difference between the retrovirus subfamilies appears to reflect their evolutionary origin, that is, lentiviruses have no known endogenous counterparts whereas most oncoviruses have endogenous cellular counterparts with which they can undergo recombination. Furthermore, we suggest that the number of CpG dinucleotides in a lentiviral genome determines the maximum potential DNA methylation level of the provirus, which in turn affects viral transcription in host cells.

INTRODUCTION

In vertebrate genomes, 60–90% of all CpG dinucleotides are methylated at position five on cytosine and about 90% of all methylated cytosines occur in CpG dinucleotides [1, 2]. CpG is the rarest dinucleotide in vertebrate DNA, being found at 20–25% of its expected frequency [1, 2, 5]. This low frequency is due to a mutational bias stemming from high frequency of spontaneous cytosine deamination (C → U) and failure to repair this mutation in the case of methylated cytosines (5-mC → T) [3, 4]. Methylation of C-residues often negatively correlates with gene expression whereas demethylation of cytosines usually leads to an increase of transcription [2, 6, 7]. 'Islands' (clusters) with high levels of unmethylated CpG dinucleotides are found 5′ to housekeeping genes and sometimes 3′ to both housekeeping and tissue-specific genes [8–10], and these clusters appear to be involved in the control of gene expression [11].

Human immunodeficiency virus type 1 (HIV1) long terminal repeat (LTR) directed gene expression in murine and Vero cells is susceptible to transcriptional inactivation by methylation [12]. Furthermore, methylation of several murine, feline and avian retroviral proviruses correlates negatively with their expression (reviewed by Cooper [2,13–17]). These retroviruses could often be activated by 5-azacytidine which decreases the level of methylation [2, 15, 16].

To examine the role CpG dinucleotides in lentiviral gene regulation, we determined the frequencies of CpG dinucleotides in 11 lentiviruses derived from 8 host species and compared them with several other retroviruses and a set of human genes. The CpG level of the HIV1 genome had previously been determined to be very low [18, 19]. We report here that the decrease in CpG levels varies both among the various lentiviral sequences and in genes within a single virus. We suggest that selection against CpG dinucleotides and not mutational bias is responsible for CpG depletion in lentiviral genes and propose a hypothesis for the role of cytosine methylation in the regulation of lentiviral transcription.

METHODS

Lentiviral sequences were obtained from the compilation of nucleic acid sequences in Human Retroviruses and AIDS [20]; human and other retroviral sequences were taken from GenBank release 61. Full-length lentiviral genomes were analyzed, including: Human immunodeficiency virus type 1, isolates BRU ($HIV1_{BRU}$), SF2 ($HIV1_{SF2}$) and ELI ($HIV1_{ELI}$); Simian immunodeficiency viruses from: African green monkey (SIV_{agm}); macaque, strain BK28 (SIV_{mac}); sooty mangabey, strain PBJ14 (SIV_{smm}) and mandrill (SIV_{mnd}); Human immunodeficiency virus type 2, strain FG (HIV2); Equine infectious anemia virus (EIAV); Visna lentivirus, Icelandic strain (Visna); Feline immunodeficiency virus (FIV). Several retroviruses of different types were studied for comparison: Simian SRV1 type D retrovirus (SRV1, GenBank locus: SIVRV1CG); AKV murine leukemia virus (MuLV, locus: MLOCG); Human T-cell leukemia virus type 1 (HTLV1, locus: HL1PRCAR); Rous sarcoma virus (RSV, locus: ALRCG).

We used the IG-suite programs (IntelliGenetics) and additional programs in 'C' (Sun 4). These additional programs (1) measure dinucleotide frequencies in 3 reading frames and compare them with expected frequencies; (2) scan several aligned sequences and find frequencies of a given dinucleotide within the 'windows' along these sequences. We used the method of Feng and Doolittle [21] to align nucleotide sequences and build phylogenetic trees.

RESULTS

Phylogenetic trees were built separately for the *gag*, *pol* and *env* proteins of the lentiviruses (data not shown, see [22, 23] for retroviral and [24] lentiviral phylogeny). These 'trees' have similar branching orders, suggesting that no major recombinational events had occurred during evolution of lentiviruses, i.e., all 3 genes evolved together in the same host cells and under control of the same replication enzymes. Therefore, any nucleotide features (e.g., frequencies of dinucleotides) that differ among the genes of a given lentivirus can be attributed to constraints on gene or protein function and structure rather than on a different evolutionary environment.

Very low CpG levels in lentiviruses

The frequency of CpG dinucleotides in retroviral *gag*, *pol* and *env* genes are diagrammed at the top of Figure 1, with observed and expected CpG and GpC frequencies in the three genes taken together shown at the bottom. As can be seen, the level of CpG dinucleotides is lower in *pol* genes than in *gag* and *env* genes in lentiviruses and SRV1. The percentage of CpG in all 11 lentiviral *gag* genes taken together is 1.05%, 1.08% in *env*, and 2.5 fold lower, 0.43%, in *pol*. Conversely, CpG frequencies are lower in *env* (3.26%) than in *pol* (3.68%) genes in the three onco-retroviruses, with *gag* having the highest CpG levels (4.62%).

The total number of CpG's is the lowest in SIV_{mnd} genes, it is approximately two fold higher in HIV1 genes and ~5 fold higher in HIV2, SIV_{agm} and Visna. CpG dinucleotides occur much more frequently in the three onco-retroviruses in comparison to lentiviruses (Fig.1). There are only 2 CpG's in the 2770 nucleotide non-overlapping part of SIV_{mnd} *pol* gene (Figs. 1, 3), which is the lowest level observed to date for any gene.

The ratios of the actual versus the expected number of CpG dinucleotides in each codon-defined position are presented in Fig. 2, a set of human genes [compiled in 25] is shown for comparison. CpG frequencies in lentiviral genes and SRV1 are lower in the 1–2 frame (CGN) than in the 2–3 frame (NCG), whereas this tendency is reversed for MuLV, HTLV1 and human genes.

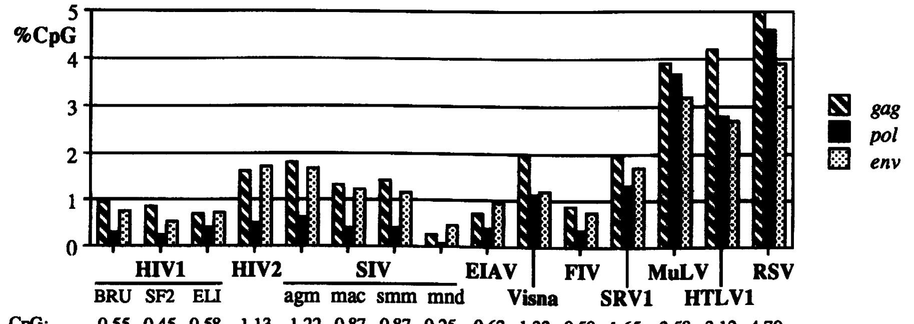
The frequencies of CpG dinucleotides as they occur along human and simian lentiviral genomes are shown in Fig. 3. GpC levels are presented as a control: these values are very close to the expected and higher than CpG frequencies in all 75 'windows' shown in Fig. 3. CpG levels increase in overlapping gene coding regions (*gag* and *pol*; *env* and *rev*; *env* and *nef*). For each virus under study, the number of CpG dinucleotides in the LTR, the region between the 5′ LTR and *gag*, and in the whole virus are shown in Table 1. The CpG level is the highest in the region between the 5′ LTR and *gag* gene and it is higher than 5.5% for all studied viruses, except SIV_{mnd}, FIV, EIAV and HIV2.

Lowered CpG frequencies: mutational bias or selection

Low levels of CpG dinucleotides might arise due to high mutation rate of CpG's and/or selection against them. If the mutation rate is high then CpG's that are already present in the genome will *disappear quickly*. Selection against CpG means that mutations leading to CpG are rarely fixed (become substitutions) because they negatively affect viral fitness, i.e., CpG dinucleotides *rarely emerge*. The analysis of the mechanism of CpG depletion in viral genes, presented below, is based on two assumptions: (1) synonymous substitutions occur more frequently in evolution than non-synonymous substitutions, this is in agreement with the observation that the rate of synonymous changes in lentiviral *pol* genes are about 10 times higher than non-synonymous changes [26]. (2) CpG levels in viral genes are low, which we have shown to be true for the viruses discussed above.

There are more synonymous mutation options for NCG codons ('G' can mutate to any nucleotide without a coding change) than for CGN codons (only the transversion CGR to AGR is synonymous). Therefore, a high mutation rate of the CpG dinucleotide would decrease the number of NCG codons more than CGN. This is true for human genes where mutational bias is known to be responsible for CpG depletion ([27], Fig. 2) as well as for MuLV and HTLV1. However, we observe the opposite for SRV1 and three lentiviral genes: Frequencies of NCG codons are about twice as high as CGN codons (Fig. 2).

The retroviral reverse transcriptase and endonuclease encoded



	BRU	SF2	ELI	HIV2	agm	mac	smm	mnd	EIAV	Visna	FIV	SRV1	MuLV	HTLV1	RSV
CpG:	0.55	0.45	0.58	1.13	1.22	0.87	0.87	0.25	0.62	1.32	0.58	1.65	3.58	3.12	4.70
GpC:	4.02	3.99	4.06	4.69	4.25	3.85	4.28	3.23	3.06	3.82	3.05	4.44	5.25	6.56	6.54
Expected:	3.82	3.77	3.75	4.38	4.20	4.09	3.99	3.25	3.27	3.75	3.05	4.17	7.03	6.23	6.94

Figure 1. Frequencies of the CpG dinucleotide in retroviral genes. Values were calculated in the following way: the number of CpG dinucleotides was divided by gene length and multiplied by 100%. Only non-overlapping parts of genes were used for analysis. The combined percentages of CpG and GpC awell as expected frequencies for these two dinucleotides in *gag*, *pol* and *env* genes taken together are shown below the figure. The expected frequencies were obtained by multiplying the mononucleotide frequencies of 'C' by those of 'G' in the three genes.

by the *pol* gene are more conserved than the envelope protein (*env*). That is why the rate of evolution is approximately 3 times lower for *pol* than for *env* [22], i.e., non-synonymous mutations in *pol* are rejected 3 times more often than in the *env* gene. A high mutation rate for CpG dinucleotides (about two-thirds of all mutations are non-synonymous) should lead to greater CpG depletion in less conserved proteins, i.e., it would decrease CpG level in *env* more than *pol*. This is true for MuLV and HTLV1, but again the tendency is opposite for lentiviruses and SRV1 (Fig. 1).

Suppose a negative selection exists against CpG dinucleotides, then the CpG levels in genes would reflect the balance two forces—an evolutionary pressure to eliminate CpG dinucleotides, and random mutations that sometimes lead to CpG. Mutations in CpG dinucleotides do not contribute significantly to this balance due to low CpG levels. There are 4236 NCH (H=A,T,C) codons in all 33 lentiviral genes studied that can change to synonymous NCG codons with one point mutation, while only AGR codons (Arg) can mutate synonymously (at 971 sites) to CGN codons. That is why random synonymous mutations would lead to NCG codons ~4 times more often than to CGN, resulting in NCG codons occurring more frequently than CGN—this is true for lentiviral genes (Fig. 2). Furthermore, non-synonymous mutations leading to CpG have a higher probability of being rejected in *pol* than in *env* genes (see above), which explains why CpG dinucleotides occur less frequently in *pol* genes as compared to *env* in lentiviruses and SRV1 (Fig. 1).

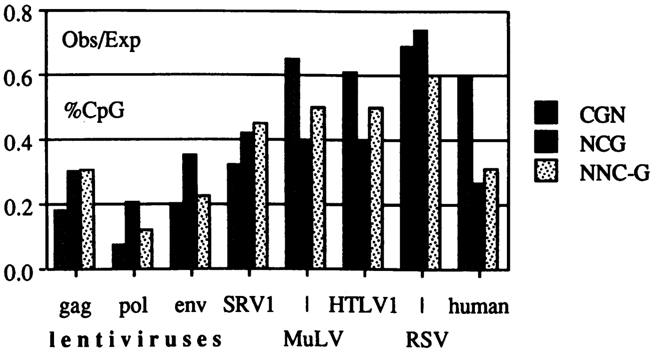


Figure 2. Ratios of the observed versus expected number of CpG dinucleotides in 3 codon-defined positions in retroviral and human genes. Data are presented for 11 pooled lentiviral genes, 4 oncoviruses (*gag*, *pol* & *env* summarized) and 14 pooled human genes (4885 codons, [25]). For example; there are a total of 4284 codons in 11 lentiviral *gag* genes; 19% of 4284 codons start with 'C' (CNN); 18.8% – have 'G' in the 2-nd position (NGN); so, the expected frequency of CGN is: 0.19×0.188×4284 = 153. There are, however, only 27 CGN codons in all *gag* genes, so the height of the bar is 27/153 = 0.18. For NCG codons in 11 lentiviral *gag* genes the number of NCG is 61, the expected number is 204 and the height of the bar is 61/204 = 0.3.

Two independent criteria, that the number of NCG codons is higher than that of CGN and the frequency of CpG in *env* is higher than in *pol*, argue that CpG depletion in lentiviral genes and SRV1 is the result of selection against CpG dinucleotides. The same criteria suggest that the CpG level is decreased in MuLV and HTLV1 mainly due to a mutational bias. The CpG frequencies are only slightly decreased in RSV (Fig. 1, 2) and thus it is unclear whether this is the result of selectional or mutational bias. (We shall discuss the relatively high CpG levels found in *gag* genes below.)

DISCUSSION

Lowered CpG levels may occur by chance, due to a high mutation rate in CpG's, or as a result of selection. The decrease in CpG levels in all viruses in this study is statistically significant ($P < 10^{-4}$), which argues against the first explanation. The foregoing analysis suggests that a biased mutational rate is responsible for low CpG frequencies in onco-retroviruses, while selection against CpG's is affecting the evolution of lentiviruses.

Evolution of retroviruses and CpG dinucleotides

The pattern of CpG frequencies in viral genes correlates with the phylogeny of retroviruses. Lentiviruses comprise one class of retroviruses—they have the lowest CpG frequencies (Fig. 1) and CpG depletion in all of them has a 'selectional pattern' (NCG higher than CGN and *env* higher than *pol*, Figs. 1, 2). Type D retroviruses (SRV1, SRV2 and Human endogenous retrovirus-like element HERV-K) are evolutionarily close [22, 23] and they all have a 'selectional pattern' of CpG depletion similar to lentiviruses. Type C retroviruses (MuLV, Feline leukemia virus [FeLV] and Baboon endogenous virus) are on one branch of the evolutionary tree and analysis of their CpG frequencies suggests that mutational bias, rather than selection, is responsible for decreased CpG levels in all three viruses.

Table 1. Frequencies of CpG dinucleotides in different retroviral regions

Virus	LTR	between 5′ LTR & *gag*	Total
HIV1$_{BRU}$	11/634(1.7/6.7%)	13/155(8.4/8.1%)	92/9766 (0.94/4.5%)
HIV1$_{SF2}$	10/634(1.6/6.4%)	15/155(9.8/8.6%)	88/9737 (0.90/4.4%)
HIV1$_{ELI}$	9/634(1.4/6.1%)	13/155(8.4/8.1%)	89/9712 (0.92/4.3%)
HIV2	13/628(2.1/6.4%)	12/245(4.9/7.3%)	156/9885 (1.58/5.1%)
SIV$_{agm}$	17/726(2.3/6.0%)	19/212(9.0/8.2%)	167/9779 (1.71/4.7%)
SIV$_{mac}$	14/806(1.7/5.9%)	13/234(5.6/6.8%)	130/10277(1.26/4.7%)
SIV$_{smm}$	14/594(2.4/6.5%)	13/234(5.6/7.4%)	129/9996 (1.29/4.7%)
SIV$_{mnd}$	11/794(1.4/5.5%)	4/174(2.3/6.7%)	61/9835 (0.62/3.9%)
EIAV	9/321(2.8/4.2%)	5/143(3.5/8.4%)	84/8228 (1.02/3.6%)
VISNA	12/414(2.9/5.3%)	20/328(6.1/7.5%)	152/9519 (1.60/4.1%)
FIV	10/355(2.8/4.3%)	9/272(3.3/5.6%)	91/9474 (0.96/3.2%)
SRV1	17/346(4.9/6.7%)	19/154(12.4/7.8%)	171/8171 (2.1 /4.4%)
MuLV	20/626(3.2/6.9%)	40/493(8.1/7.8%)	314/8890 (3.5 /7.0%)
HTLV1	45/756(6.0/7.8%)	4/ 46(8.9/9.9%)	344/9055 (3.8 /6.5%)
RSV	16/334(4.8/5.1%)	22/278(7.9/9.3%)	477/9625 (5.0 /7.2%)

Frequencies are shown in the following format:
Number of CpG / Length of the region (Percentage of CpG observed/expected)

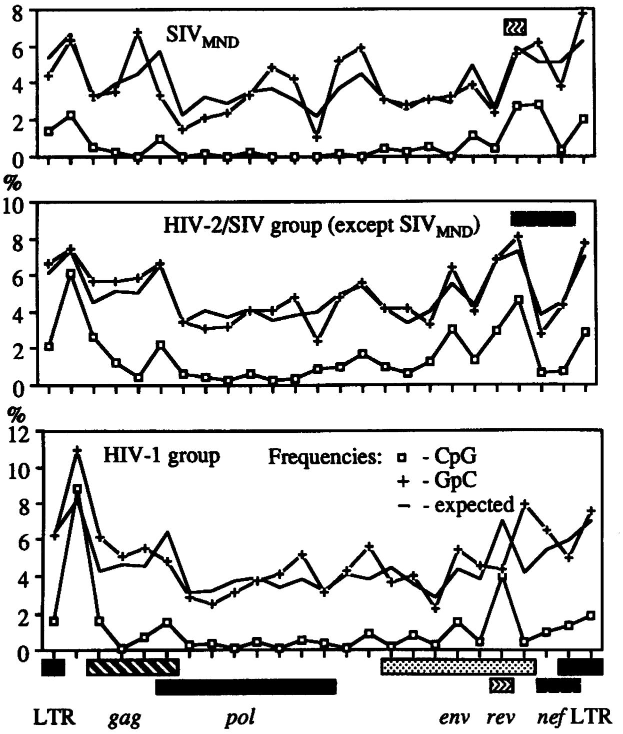


Figure 3. Occurrence of CpG (□) and GpC (+) dinucleotides along human and simian lentiviral genomes. Expected frequencies are shown by a plain line. The aligned viral genomes were 'cut' into 25 'windows' for analysis, so that all genes and LTRs start and end exactly at the border of a window. Averaged frequencies are presented for the HIV1 and HIV2/SIV groups (viruses within each group have very similar profiles). Positions of *rev* and *nef* genes and their overlap with *env* differ among viruses. The short open reading frames, *tat, vif, vpx*, are not shown.

Doolittle [23] suggested that exogenous retroviruses exist as 'short-lived bursts' upon the germline-encoded endogenous retroviruses. During a long endogenous state the evolution is controlled by the host, it is slow and CpG depletion can arise due to mutational bias (similar to host genes). During short periods of exogenous reproduction the evolution rate increases dramatically due to the high error level of reverse transcriptase, absence of proof-reading and a short reproduction cycle [23]. The question then arises: what fraction of present-day viral sequences emerged from exogenous versus endogenous evolution?

We suggest that CpG frequencies can be used as 'markers' of retroviral evolution. According to these markers, a significant part of the present-day MuLV (as well as those of FeLV and avian viruses) sequence is mainly a result of endogenous evolution, since the pattern of CpG depletion is mutational, similar to host genes. This finding correlates well with the presence of multiple copies of endogenous virus-like elements in these hosts and the demonstration of recombination between exogenous and endogenous viral sequences during exogenous virus infections [28–30]. On the other hand, lentiviral and SRV1 sequences have no 'traces' of endogenous evolution in their CpG pattern, therefore, it seems that their sequences emerged during exogenous evolution. Furthermore, no endogenous counterparts are known for lentiviruses. Interestingly, this correlation, does not hold true for HTLV1—whereas the pattern of CpG frequencies suggests mainly mutational origin, no endogenous counterparts have yet been identified.

Frequencies of CpG dinucleotides and functional constraints

The degeneracy of the genetic code permits any protein to be encoded without the inclusion of CpG dinucleotides. CGN codons can be replaced by the synonymous AGR codons (Arg), NCG codons are always synonymous to NCH (H=A,T,C), and NNC-G context is always synonymous to NNT-G ('−' denotes separation of adjacent codons). However, the usage of CpG is indispensable in overlapping coding regions for certain pairs of amino acids in different frames, as well as in some functional sites, e.g., the binding sites for transcription factor Spl in the LTR, the tRNA primer binding sequence between the 5′ LTR and gag. In agreement with these predictions, the results presented in Fig. 3 clearly show that the highest levels of CpG occur in overlapping parts of genes, LTRs, and in the region between the 5′ LTR and *gag*.

Regions 5′ to mammalian genes usually contain CG-rich 'islands' (clusters) with high CpG levels [7–11] and it seems that many retroviruses mimic this feature (Fig. 3, Table 1)—both CpG and [C+G] levels are high in the region between the 5′ LTR and *gag*. The actual number of CpG is even higher than the one expected in HIV1, SIV$_{agm}$, SRV1 and MuLV (Fig. 3, Table 1). The CG-rich islands 5′ to mammalian genes have been suggested to play a role as enhancer-like elements [7] although no similar function for the region between the 5′ LTR and *gag* has been described. In many viruses, especially Visna, MuLV and HTLV1, the CG-rich island partly overlaps with the 5′ part of the *gag* gene, which explains why *gag* often has the highest CpG levels compared to *env* and *pol* genes (Figs. 1, 3).

CpG levels in non-overlapping parts of lentiviral genes which have no known functional constraints on CpG usage are the result of a balance between neutral mutations (all synonymous ones and non-synonymous mutations that preserve protein functions) and selection against CpG dinucleotides.

Hypothesis for the effect of methylation on lentiviral transcription

Selection against CpG's means that low CpG frequencies increase fitness of lentiviruses, i.e., CpG levels play a certain biological role. We suggest that the number of CpG dinucleotides in lentiviruses determines the maximum possible methylation level of the provirus, which in turn may play a role in the regulation of gene transcription during the viral life cycle. DNA genomes of lentiviruses are initially synthesized by the viral reverse transcriptase and are not methylated prior to integration into the host DNA. It is not known to what extent lentiviral proviruses are methylated during periods of transcriptional activity or latency. It was shown, however, that another retrovirus, Moloney murine leukemia virus (Mo-MuLV), is methylated between the eighth and sixteenth days after infection [13, 14]. If methylation of HIV1 occurs at the same rate as it does in Mo-MuLV infection, one can suggest that an initial period of viral replication is correlated with the absence of methylation after which latency can result as a consequence of methylation. A similar mechanism for HIV latency was proposed by Bednarik *et al.* [12]. Kypr *et al.* [18] also suggested that low CpG levels in lentiviruses might be important to avoid cytosine methylation.

The number of CpG dinucleotides in the viral genome determines the maximum possible level of methylation; if the 'latent period' is correlated with methylation it might continue longer or is more likely to occur in viruses with higher CpG levels, which may in turn have an impact on viral pathogenicity. In that vein it is interesting that SIV$_{agm}$ and HIV2 have the highest CpG levels among human/simian lentiviruses, and are of apparently lower pathogenicity in their hosts than HIV1. SIV_{mnd} has both the lowest CpG level and lacks a third Sp1 transcription factor binding site usually present in lentiviral enhancers [31]. Although the pathogenicity of this virus is unknown, the expected reduction in gene expression due to a weaker enhancer could conceivably be compensated for by an overall low methylation level of the provirus, thus rendering the virus less susceptible to transcriptional inactivation.

Since no evidence of selection against CpG is found in oncoretroviruses, it is unclear if CpG levels are of biological importance in these viruses.

Selection appears to decrease the frequencies of CpG dinucleotides in lentiviral genes, suggesting that low CpG levels are of biological importance in these viruses. The hypothesis on the role of cytosine methylation in the regulation of lentiviral transcription might contribute to the understanding of the factors that control lentiviral expression in the host cells.

ACKNOWLEDGEMENTS

This work supported by U.S. Public Health Service Awards AI85007 and CA01058. We thank Prof. R.Doolittle and Dr. R.Alberro for providing computer programs for phylogenetic analysis. We thank Prof. S.Karlin, P.Edmonson, Drs. M.Bachmann, P.Bucher, C.Contag, and J.Jurka for valuable comments on the manuscript.

REFERENCES

1. Bird, A. (1986) *Nature*, **321**, 209–213.
2. Cooper, D.N. (1983) *Hum. Genet.*, **64**, 315–333.
3. Coulondre, C., Miller, J.H., Farabough, P.J. and Gilbert, W. (1978) *Nature*, 274, 775–780.
4. Bird, A.P. (1980) *Nucleic Acid Res.*, **8**, 1499–1504.
5. Josse, J., Kaiser, A.D. and Kornberg, A. (1961) *J. Biol. Chem.*, **236**, 864–875.
6. Cedar, H. (1988) *Cell*, **53**, 3–4.
7. Jones, P.A. (1986) *Cancer Res.*, **46**, 461–466.
8. Gardiner-Garden, M. and Frommer, M. (1987) *J. Mol. Biol.*, **196**, 261–282.
9. Cooper, D.N. and Gerber-Huber, S. (1985) *Cell Differ.*, **17**, 199–205.
10. McClelland, M. and Ivarie, R. (1982) *Nucleic Acid Res.*, **10**, 7865–7877.
11. Wolf, S.F. and Migeon, B.R. (1985) *Nature*, **314**, 467–469.
12. Bednarik, D.P., Mosca, J.D. and Raj, N.B.K. (1987) *J. Virol.*, **61**, 1253–1257.
13. Niwa, O., Yokota, Y., Ishida, H. and Sugahara,T. (1983) *Cell*, **32**, 1105–1113.
14. Gautsch, J.W. and Wilson, M.C. (1983) *Nature*, **301**, 32–37.
15. Hoffmann, J.W., Steffen, D., Gusella, J., Tabin, C., Bird, S., Cowing, D. and Weinberg, R.A. (1982) *J. Virol.*, **44**, 144–157.
16. Groffen, J., Heisterkamp, N., Blennerhassett, G. and Stephenson, J.R. (1983) *Virology*, **126**, 213–227.
17. Ponta, H., Günzburg W.H., Salmons B., Groner, B., Herrlich, P. (1985) *J. Gen. Virol.*, **66**, 931–943.
18. Kypr, J., Mrazek, J. and Reich, J. (1989) *Biochim. Biophys. Acta*, **1009**, 280–282.
19. Ohno, S. and Yomo, T. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 1218–1222.
20. Myers, G., Rabson, A.B., Josephs, S.F., Smith, T.F., Berzofsky, J.A. and Wong-Staal, F. (eds.) (1989) Human Retroviruses and AIDS 1989. A Compilation and Analysis of Nucleic Acid and Amino Acid Sequences. Los Alamos National Laboratory, Los Alamos.
21. Feng, D.-F. and Doolittle, R.F. (1987) *J. Mol. Evol.*, **25**, 351–360.
22. McClure, M.A., Johnson, M.S., Feng, D.-F. and Doolittle, R.F. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2469–2473.
23. Doolittle, R.F., Feng, D.-F., Johnson, M.S and McClure, M.A. (1989) *The Quart. Review Biol.*, **64**, 1–30.
24. Olmsted, R.A., Hirsch, V.M., Purcell, R.H., and Johnson, P.R. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 8088–8092.
25. Hanai, R. and Wada, A. (1990) *J. Mol. Evol.*, **30**, 109–115.
26. Yokoyama, S., Chung, L. and Gojobori, T. (1988) *Mol. Biol. Evol.*, **5**, 237–251.
27. Hanai, R. and Wada, A. (1988) *J. Mol. Evol.*, **27**, 321–325.
28. Weiss, R.A., Mason, W.S. and Vogt, P.K. (1973) *Virology*, **52**, 535–552.
29. Coffin, J. (1982) In Weiss, R., Teich, N., Varmus, H. and Coffin, J. (eds.), RNA Tumor Viruses. Cold Spring Harbor Laboratory, New York, pp. 1109–1204.
30. Overbaugh, J., Riedel, N., Hoover, E.A. and Mullins, J.I. (1988) *Nature*, **332**, 731–734.
31. Dewhurst, S., Embretson, J.I., Anderson, D.C., Mullins, J.I. and Fultz, P.N. (1990) *Nature*, **345**, 636–640.