CG doublet difficulties in vertebrate DNA

SIR — Contrary to the idea proposed by Adams and Eason [1], the dinucleotide CG does not appear to be less avoided at higher G+C contents in vertebrate DNA sequences. The most meaningful relation in terms of the methylcytosine deamination hypothesis for the suppression of CpG [1-3] is the difference between the normalized frequencies of doublets CG and TG as a function of % G+C. These normal statistics refer to the doublet frequencies expected from base composition and are calculated from permutations of the bases in the sequence [4]. I have therefore made plots of CpG suppression (CG – TG) versus % G+C for all human, mouse and chicken sequences of suitable length in our information system ACNUC, which includes the latest GenBank release [4]. No correlation between the two parameters is found. The curves of CG – TG against % G+C are flat with high dispersion for each species. A few sequences do show less avoidance of CpG over 60% G + C, but this is not a general trend. Indeed, although the CG – TG dispersion is large at all G+C contents with the 84 human sequences studied, it is greatest above 60% G+C.

Another test of the idea was made by sectioning the Epstein-Barr virus genome sequence (in latent form it can be highly methylated). The 172 kilobase sequence was divided into 17 segments of 10,000 bases each and the normal statistic CG – TG plotted against % G+C (since the statistic is cumulative the best comparison is with sequences of exactly the same length). This likewise gave a flat, dispersed curve. Furthermore, *Drosophila* sequences, which are not methylated [3,5], also produced a curve of like appearance. Finally, plots for *Escherichia coli* and phages T7 and lambda are rather similar to those above (the T7 and lambda complete genome sequences were cut into 2,000-base segments for the analysis).

To add to Max's findings [2], in my human workfile seven sequences do show somewhat less CpG avoidance than the others; c-*myc* 5′ end; enkephalin 5′ flank and start of intron C; alpha-actin mRNA; c-*myc* exon 2; enkephalin in exons 1 and 2; tRNA Glu gene and flanks; 45S rRNA transcription origin region (which at 75.3% G+C shows no evidence of CpG suppression). The last five sequences have more than 60% G+C, but nine others of over this amount reveal strong CpG avoidance. In the mouse sample (100 sequences of 1-2 kilobases each, 27 from segmenting longer sequences), the seven of lowest CpG avoidance were: hypoxanthine phosphoribosyl transferase exon 1; 5′ end 45S rRNA precursor; first 1,240 bases of MHC class II H2-IA-beta exons 2-6; RSRGO repeat element 5′ to rRNA transcription origin; first 1,180 bases of MUSCFOS; second half of MUSCMYCONC; 5′ flank, exons 1 and 2 and first introns in MHC class I gene Q10. The first two sequences have over 60% G+C. Thus instead of a general increase in CpG normalized frequency at high G+C contents, we see particular sequence families exhibiting low CpG avoidance, while taken altogether sequences reveal the avoidance at least as strongly above 60% G+C as below. I wish I had a new hypothesis for the rarity of CpG in DNA.

RICHARD GRANTHAM
Institut d'Evolution Moléculaire,
Université Lyon I,
69622 Villeurbanne, France

1. Adams, R.L.P. & Eason, R. *Nature* **312**, 407-408 (1984).
2. Max, E.E. *Nature* **310**, 100 (1984).
3. Bird, A.P. *Nucleic Acids Res.* **8**, 1499-1504 (1980).
4. Grantham, R. *et al. Bull. Inst. Pasteur* (in the press).
5. Bird, A.P. *Nature* **307**, 503-504 (1984).

SIR — In a recent *News and Views* article [1], I discussed our published observations on "CpG-rich" segments of mammalian genes, and described a model we proposed [2] that might explain local persistence of this usually rare dinucleotide in 5′ regions of certain genes. We suggested that these regions might have been protected from the usual mechanism of CpG loss (deamination of 5-methylCpG to TpG) if these CpG-rich regions were undermethylated in germ-line DNA. The article provoked several responses, including some published in *Nature* [3,4].

Erickson [3] addresses the methylation status of class I MHC genes, a category of genes in which we noticed that CpG-rich regions occur consistently in the region 5′ of intron 3. Erickson cites his experimental evidence that class I MHC genes of the mouse are highly methylated in sperm DNA, which may seem to contradict our undermethylation model; but I believe that his data do not provide a good test of this model. The particular experiment referred to (Fig.3 in ref.5) shows a Southern blot of mouse sperm DNA disgested with *Msp*I and *Hpa*II and probed with a mouse class I MHC probe. If the CpG-rich 5′ regions of the MHC genes were completely demethylated while the 3′ regions were highly methylated, the 5′ regions might be so thoroughly digested at closely spaced *Hpa*II sites that the fragments generated from 5′ regions of the genes would be too small to be detected in routine Southern blots. (In the H-2L^d sequence, for example, the largest *Hpa*II fragment contained in the CpG-rich region 5′ of intron 3 is 197 bp; most fragments are less than 100 bp.) Therefore, one might reasonably predict that the only fragments detectable on routine Southern blots of the *Hpa*II digest would be large fragments generated from the highly methylated 3′ end of the gene, yielding bands similar to what is observed in Erickson's Fig.3. Since the undermethylation model is consistent with the Southern blot reported in Erickson's paper, while this blot is also consistent with the opposite prediction (methylation of the 5′ regions), it is apparent that this experiment does not test the model. (This does not imply a defect in Erickson's data; his experiments were simply designed for other reasons than to explore this particular model.) An additional caveat should be mentioned. Mature sperm represents a state of male germ-line DNA that is terminal and relatively transient compared with that of the long-lived precursor stem cells, cells which may have a different pattern of DNA methylation than mature sperm. Therefore the methylation status of mature sperm DNA may not reflect the methylation pattern most significant in influencing the accumulation of deamination mutations over successive generations. This caveat obviously applies equally well to the experiments that seem to support our undermethylation model by demonstrating regional undermethylation in CpG-rich regions of several genes in mature sperm (as cited in the *News and Views* piece).

Adams and Eason [4] suggest that G+C-rich regions of mammalian DNA may be held so tightly as a double helix that deamination of 5-methyl-cytosine residues, a reaction thought to occur only on single-stranded DNA, may be prevented in these regions; as a consequence, deamination-caused 'CpG suppression' would be absent in these G+C-rich areas. This provides an alternative model to the one we proposed explaining CpG-rich regions as a consequence of undermethylation of these regions in germ-line DNA.

One theoretical reservation that I have about the 'tight helix' model is that the notion that deamination occurs only on single-stranded DNA is based on *in vitro* experiments on deamination events catalysed by bisulphite [6,7]. Since the requirement of single-stranded DNA for the bisulphite-catalysed reaction may result from poor steric access of the catalyst in double-stranded DNA, it is not necessarily true that the *in vivo* deamination, not bisulphite-catalysed, would show the same preference for single-stranded DNA.

If one accepts the extrapolation and assumes that *in vivo* deamination also requires single-stranded DNA, then the tight helix model seems to apply well to the genes we considered in our earlier paper [2], since in these genes the regions that are relatively CpG-rich by and large are C+C-rich also. However, this correlation is not always found. Mouse satellite DNA is composed of a highly repeated DNA segment which demonstrates no CpG suppression. In the 234-bp unit sequence [8], the observed CpG frequency (3.4%) is almost exactly that predicted from G+C content, yet this sequence does not fall into the category of G+C rich sequences described in the tight helix model since the G+C content is only 36.8%. It does, however, fit the undermethylation model since available data [9,10] suggest that this satellite DNA is undermethylated in germ-line cells (mature sperm and spermatogonia) although it is highly methylated in other tissues. Furthermore, even when CpG