J Mol Evol (1983) 19:286–288

Journal of Molecular Evolution

Letter to the Editor

CpG Frequency in Large DNA Segments

Gregory G. Lennon and Nigel W. Fraser

The Wistar Institute, Philadelphia, Pennsylvania 19104, USA

Summary. The relationship between the deficiency of CpG dinucleotides and the coding-noncoding segments of DNA has been examined. Analysis of five human α-like globin DNA sequences and five human β-like globin DNA sequences reveal that there is no apparent difference between protein coding and noncoding portions of DNA. Rather CpG deficiency appears to be a property of long contiguous segments of DNA consisting of several genes and their intergenic regions. Thus we propose that CpG deficiency is not involved with translation or transcription but rather is related to chromosomal constraints.

Key words: Eukaryotic DNA – CpG deficiency – Globin genes

The largest deviation from random DNA nucleotide sequence in eukaryotic DNA is that of the deficit of CpG dinucleotides [1]. Amino acid sequence data supported the hypothesis that the CpG shortage in the DNA of vertebrates might be due to a shortage of codons containing this doublet in the polypeptide coding regions of DNA [2]. This was further supported by the finding that tRNA, 5S RNA and ribosomal RNA did not show a deficit of CpG. When a comparison of the nucleotide sequences in HeLa cell mRNA and HnRNA by RNA fingerprinting techniques was performed, it was found that they were both indeed CpG deficient [3]. While HnRNA appeared to be very similar to DNA, mRNA appeared to be slightly less deficient in the nearest neighbor CpG. Recently, Nussinov [4] claimed that the CpG dinucleotide deficiency seen in many eukaryote DNA sequences is *not* due to coding requirements based on a comparison of the mostly non-coding Ig light J cluster sequence with other gene coding sequences. A simple way to support Nussinov's conclusion is to compare coding and non-coding segments within specific genes. This type of analysis has been successfully used for SV40 [5].

In this paper we have compared the CpG frequencies of the 5′ untranslated, exon, intron and 3′ untranslated segments of the known human globin genes (Table 1). We have calculated the randomly expected number for each segment from its base composition and compared it with the actual number of CpGs counted in that segment. It can be seen that, in general, for any given gene these segments have similar CpG frequencies. Given that the introns are larger than the exons in Table 1 and are just as depressed (if not more so) in the ratio of actual to expected number of CpG dinucleotides, we can infer that globin HnRNA will be lower in CpG than globin mRNA, in agreement with the experimental data on whole HeLa cell mRNA and HnRNA [3].

These data do not support explanations of CpG paucity due to selection at the level of translation [2, 6]. Neither do the earlier data of Fraser et al. [3] nor the analysis of vertebrate protein sequences, which has shown that amino acids with a high proportion of codons ending in C occur with significantly reduced frequency when preceding amino acids whose codons start with G [7]. Rather, the data in Table 1 would be consistent with structural explanations of CpG deficiency.

Salser [8] suggested that CpG sequences are poorly represented in eukaryotic DNA because they are hot spots for mutations. That CpG can be methylated and mutated to TpG is known [9]. In the case of the α-globin structural gene which has the expected frequency of CpG, Salser suggests that there are structural constraints important for mRNA function which select against DNA mutations reducing CpG frequency. Clearly,

Offprint requests to: N.W. Fraser

Table 1. Ratio of Actual/Expected CpG

	α Cluster					β Cluster				
Sequence[1]	Ψζ1	ζ1	Ψα1	α2	α1	ε	G_γ	A_γ	δ	β
Expressed in	-	Embryo	-	Fetus/ Adult	Fetus/ Adult	Embryo	Fetus	Fetus	Adult	Adult
Nucleotides in sequence analysed	2640	2520	813	1138	900	3919	1648	1628	1076	2052
5′ non-coding to CAP	0.11*	0.22*	0	1.06	1.17	0.23*	0.50	0.50	0.17*	0*
CAP to Exon 1	0.67	0.33	0	0	0	0.33	1.00	1.00	0	0
Exon 1	0.50	0.50	0.33*	0.90	0.90	0.17*	0*	0*	0.14*	0.29*
Intron A	0.35*	0.40*	0.11*	0.67	0.67	0*	0.12*	0.12*	0*	0*
Exon 2	1.09	1.09	0.29*	0.95	0.95	0.21*	0*	0*	0.07*	0.12*
Intron B	1.03	1.07	0.17*	0.94	0.84	0.19*	0.23*	0.16*	0.12*	0.15*
Exon 3	0.93	1.00	0.14*	0.58	0.58	0.20*	0.20*	0.20*	0.11*	0.10*
Exon 3 to Poly A	0.45*	0.82	0.20*	0.55*	0.36*	0*	0	0	0*	0.20
Poly A to End of DNA Sequenced	0.75	0.63*	-	0.06*	-	0.20*	-	-	0.12*	0*

*Actual number of CpG in segment significantly different by χ^2 test from random expectation at $P < 0.05$

[1] DNA sequence data was obtained from the nucleic acid sequence database, National Biomedical Research Foundation and from Dr. T. Maniatis [14]

it is difficult to make this argument for the ζ-pseudogene, a non-functional gene within the α-globin cluster, which is relatively high in CpG content (see Table 1).

It has also been suggested that the selective pressure against CpG may be to diminish the extent to which DNA may be methylated [10, 11]. It is known that the extent of methylation of DNA is related to regulation of its transcriptional activity [12, 13]. Given the large difference between the coordinately regulated α-globin and the β-globin genes (Table 1), selection against methylation is unlikely to be the reason for low CpG frequency.

DNA viruses of eukaryotic cells seem to fall into two classes with regard to CpG deficiency: 1) those that are low in CpG, mainly the small viruses, as typified by SV40, and 2) those that are not, mainly the large viruses, as typified by HSV-1. In line with the fact that whole viral genomes may or may not be CpG deficient, we would like to suggest that CpG distribution, whether high or low, will be relatively uniform over large blocks of DNA such as large gene clusters. From Table 1 the α-globin cluster, a contiguous region spanning at least 20 kb (from which 8 kb was analyzed) is relatively high, while the β-gene cluster (at least 40 kb) is relatively low. Rather than supporting a transcriptional or translational constraint on CpG frequency, we propose that these data point to a structural constraint at the DNA level, such as a constraint of chromosomal organization or replication.

Acknowledgements. We wish to thank Dr. T. Maniatis for making available sequence data prior to publication (Proudfoot et al. (1982) [14]), and Dr. M.O. Dayhoff for use of the nucleic acid sequence database. This work is supported by Grant No. AI 16815. G.L. supported by NIH predoctoral trainee grant GM 07071.

References

1. Swartz MN, Trantner TA, Kornberg A (1962) Enzymatic synthesis of deoxyribonucleic acid. XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. J Biol Chem 273:1961–1967
2. Subak-Sharpe JH, Burk RR, Crawford LV, Morrison JM, Hay J, Keir HM (1966) An approach to evolutionary relationships of mammalian DNA viruses through analysis of the pattern of nearest neighbor base sequences. Cold Spring Harbor Symp Quant Biol 31:737–748
3. Fraser NW, Burdon RH, Elton RA (1975) Comparison of nucleotide sequences in HeLa cell mRNA and HnRNA. Nucleic Acid Res 2:2131–2146
4. Nussinov R (1981) The universal dinucleotide asymmetry rules in DNA and the amino acid codon choice. J Mol Evol 17:237–244
5. Grantham R (1978) Viral, prokaryote and eukaryote genes contrasted by mRNA sequence indexes. FEBS Lett 95: 1–11
6. Russell GH, Walker PMB, Elton RA, Subak-Sharpe JH (1976) Doublet frequency analysis of fractionated vertebrate nuclear DNA. J Mol Biol 108:1–23
7. Bullock E, Elton RA (1972) Dipeptide frequencies in proteins and the CpG deficiency in vertebrate DNA. J Mol Evol 1:315–325
8. Salser W (1975) Globin mRNA sequences: Analysis of base pairing and evolutionary implications. Cold Spring Harbor Symp Quant Biol 40:985–1002
9. Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acid Res 8:1499–1504
10. Jukes TH (1978) Codons and nearest-neighbor nucleotide pairs in mammalian messenger RNA. J Mol Evol 11: 121–127
11. Holliday R, Pugh JE (1975) DNA modification mechanisms and gene activity during development. Science 187:226–232
12. Van Der Ploeg LHJ, Flavel RA (1980) DNA methylation in human α δ β-globin locus in erythroid and non-erythroid tissues. Cell 19:947–958
13. Doerfler W (1981) DNA methylation of regulatory signal in eukaryotic gene expression. J Gen Virol 57:1–20
14. Proudfoot NJ, Gil A, Maniatis T (1982) The structure of the human zetaglobin and a closely linked nearby identical pseudogene. Cell 31:553–563

Received August 27, 1982/Revised March 1, 1983