Hypothesis

Analysis of CpG dinucleotide frequency in relationship to translational reading frame suggests a class of genes in which mutation of this dinucleotide is asymmetric with respect to DNA strand

David P. Leader*, Beate Peter**, Birgit Ehmer***

Department of Biochemistry, University of Glasgow, Glasgow G12 8QQ, UK

Received 20 October 1995

Abstract Results are described from application of a computer program that compares the expected and actual incidence of CpG dinucleotides in relation to the codon reading frame of genes, assuming a conserved amino acid sequence and normalizing for the third-position incidences of C and G in the remainder of the sequence. Sequences encoding certain proteins showed a pronounced bias in favour of CpG in the {3,1} compared with the {2,3} codon position; whereas sequences encoding related proteins expressed to a similar extent or in the same tissue did not. We propose that the cases exhibiting this bias reflect a difference between the two strands of the DNA duplex in their susceptibility to loss of CpG dinucleotides by mutation. Although in vertebrates this loss of CpG dinucleotides from the sense strand might reflect strand-asymmetry in deamination of 5-methylcytosine residues, the fact that a similar CpG codon bias is found in some invertebrates indicates that other factor(s) must also be involved.

Key words: CpG; Dinucleotide; Actin; Methylation; Computer program

1. Introduction

It is well known that vertebrate genomes have an overall deficiency in the dinucleotide, CpG. One explanation for this is that CpG dinucleotides are the target for methylation by eukaryotic methyltransferases, and the 5-methylcytosine resulting from the action of this enzyme is prone to spontaneous deamination to thymine. Certain regions of vertebrate genomes that do not reflect the overall deficiency in CpG contain unmethylated CpG dinucleotides which may be important for transcriptional activity (reviewed in [1]). Methylation cannot, however, be the only explanation for the CpG deficit, as it is also observed in invertebrates, albeit to a lesser extent [2].

There are other features of the CpG deficit which are difficult to explain. One of these, which we had noted previously in mouse actin cDNA sequences [3], is an overall greater deficit in the {2,3} codon position (defined in Fig. 1) compared with the {3,1} codon position [2,4]. We have been interested in analysing this difference in a more rigorous fashion, taking into account both the constraints of the amino acid sequence and the third-position nucleotide composition of the gene. A computer program, 'Dinucleotides', was written to perform such an analysis, and its application to some highly-conserved vertebrate and invertebrate sequences identified several that exhibited an extreme bias in favour of CpG in the {3,1} codon position. We suggest that in these cases the differences in mutation rates of CpG from the different codon positions, rather than being related to translational requirements, may reflect a difference in the mutation rates from the two strands of the DNA duplex.

2. Computation of 'random expected occurrence of CpG' in different codon positions

To simplify the program to calculate the number of CpG dinucleotides one could expect to occur in the {2,3} and {3,1} codon positions of a 'cDNA' (here used as shorthand for the amino acid coding regions of a gene), it is expedient to assume that the amino acid sequence it encodes is totally conserved, as is, indeed, the case for the actin isoforms that provoked this study. It was then necessary to determine:

(i) how many codons there are with C in position 2 (to allow CpG at {2,3}) and how many with G in position 1 (to allow CpG at {3,1}), and

(ii) for these respective cases, how the degeneracy of the

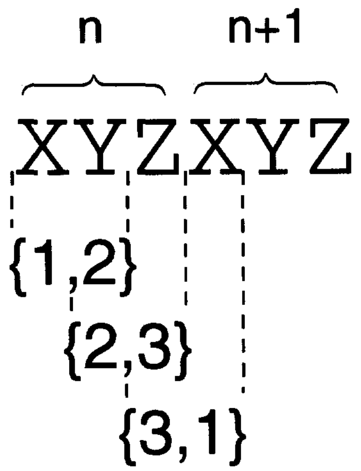

Fig. 1. Definition of {1,2}, {2,3} and {3,1} codon positions for considering CpG occurrence. Two consecutive codons (n and n + 1) are shown.

*Corresponding author. Fax: (44) (141) 330 4620.
E-mail: d.leader@udcf.gla.ac.uk

***Present address:* Pfizer Central Research, Sandwich, Kent, UK

****Present address:* Department of Zoology, University of Würzburg, Germany

SSDI 0014-5793(95)01268-0

genetic code determines the theoretical likelihood of a C in position 2, or a G in position 1.

This is referred to as 'random expected occurrence of CpG' and was computed as follows:

2.1. CpG in the {2,3} codon position

Codons of the type NCM (where N and M are any nucleotides) include NCG, in which the CpG is in the {2,3} codon position. There are four amino acids encoded by codons of the this type (Ser, Pro, Thr and Ala), each being four-fold degenerate. Thus, the 'random expected occurrence of CpG' in the {2,3} position is simply a factor of 0.25 of the number of amino acids specified by codons of the NCM type.

2.2. CpG in the {3,1} codon position

CpG can occur in the {3,1} codon position where an amino acid, the codons of which begin with G (Arg, Cys, Gly, Trp, and certain Ser codons), is preceded by an amino acid, the degenerate codons for which include one with C in the third position. To compute the random expected occurrence of such CpG dinucleotides each codon preceding a G was assessed individually as follows:

(a) Where C occurs as a member of a four-codon (Val, Pro, Thr, Ala, Gly), three-codon (Ile) or two-codon (Asn, Asp, Cys, His, Phe, Tyr) family allowing C in the third position, the expected occurrences were taken as factors of 0.25, 0.33 and 0.5, respectively, of the number of amino acids of each type preceding a codon beginning with G.

(b) Where C occurs as a member of a six-codon family, the expected occurrences were taken as a factor of 0.16 of the occurrence of Arg and Leu (in which only 1 of the 6 codons has C in the third position), and 0.33 for Ser (in which 2 in 6 codons has C in the third position). (The six-codon families do, of course, comprise 4-codon and 2-codon subsets. Although the likelihood of mutation from one subset to another is less than that of a mutation within a subset, it was necessary to consider the subsets together because of the nature of the computer program, which first translates the nucleotide sequence and then performs calculations on the basis of this.)

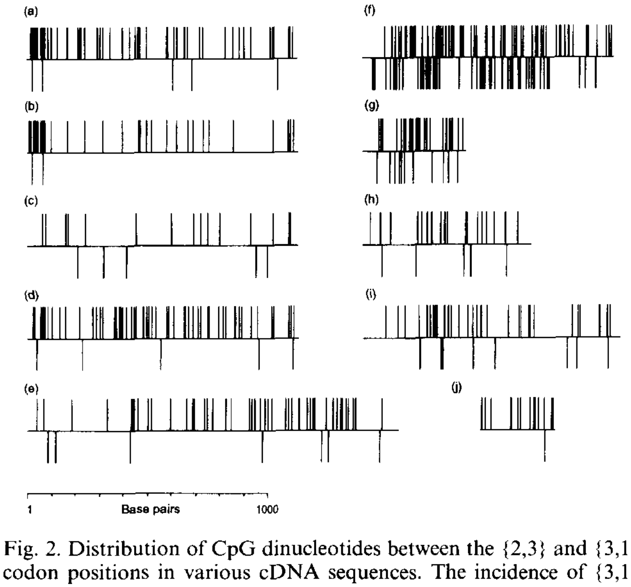

Fig. 2. Distribution of CpG dinucleotides between the {2,3} and {3,1} codon positions in various cDNA sequences. The incidence of {3,1} CpGs is shown above the horizontal line, and that of {2,3} CpGs below the horizontal line, in each case. The sequences are: (a) human cytoplasmic β-actin, (b) mouse β-cytoplasmic actin, (c) human smooth muscle α-actin, (d) *A. nidulans* γ-like actin, (e) human glucose 6-phosphate dehydrogenase, (f) human transcription factor IL6, (g) human thymidine kinase, (h) human ζ-globin, (i) human Gx-α, and (j) human histone H4. Most of the original sequences are taken from version 3 of the 'Human CpG-island database', which may be downloaded from many Internet sites, including that at EMBL, Heidelberg.

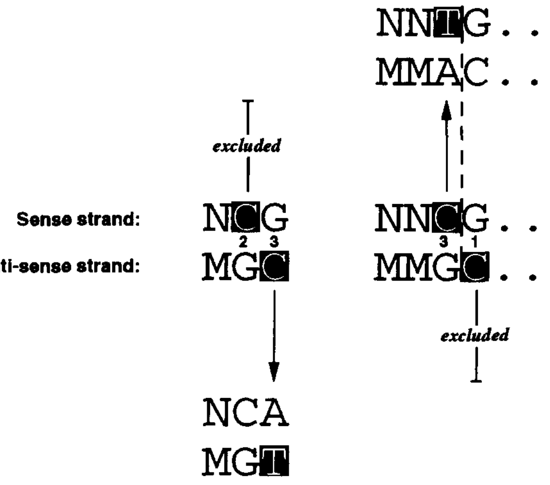

Fig. 3. Consequences of the mutation of cytosine in CpG dinucleotides occupying different codon positions. Bases undergoing mutation and their products are indicated by colour reversal. Where a mutations would involve an amino acid change (selected against in conserved amino acid sequences) the mutation is marked 'excluded' and the product is not indicated. N and M represent any complementary base-pair.

Codons not allowing C in the third position were excluded from consideration. The overall 'random expected occurrence of CpG' was computed as the simple arithmetical sum of the expected occurrences at all the positions in the region under consideration.

3. Computational allowance for overall base-composition of DNA

The computations described above assume that the chances of each of the four different nucleotides occupying a completely unrestrained position are the same. Clearly, this assumption is incorrect if the overall base-composition is other than 50% (G + C) — as is the case for many of the mammalian actin isoforms. However a simple correction for overall GC-content [4] is inadequate because of three further complexities. One is that different regions of vertebrate genomes differ in overall base composition [5], so that the base composition of each individual coding sequence needs to be considered; second, that base frequencies in the third codon position need to be considered as these generally differ from those in the positions constrained by amino acid sequence (and differ for G and C); and third, that the expected occurrence in the third position must be computed taking into account the constraints of a conserved amino acid sequence. We therefore considered the actual overall occurrence of C and G at the third codon position only, and related this to the expected random occurrence given the specific codon degeneracy in each case in a similar manner to that for CpG in the {2,3} position, described above. (Situations in which a third position C or G could be part of a CpG dinucleotide were, of course, excluded from this analysis.) The results of this 'correction' are presented separately in the computer program, for clarity.

The computer program, 'Dinucleotides', uses subroutines of version 8 of the GCG package [6] and runs under VMS or Unix. It also provides a simple graphical representation of the position of {2,3} and {3,1} dinucleotides along a nucleotide sequence, which is a useful visual aid and can identify local regions of interest. It is available either from DPL or from Internet sites to which it will be submitted.

4. Certain cDNAs are much less deficient in CpG in the {3,1} than in the {2,3} codon position

Mouse cytoplasmic γ-actin was the starting point for our analysis as the conservation of certain CpG dinucleotides had become apparent in comparing it to several corresponding processed pseudogene sequences in which many transitions from CpG were evident [7]. Like other mammalian actins, its amino acid sequence is totally conserved within vertebrates (reviewed in [8]), and a high overall content of individual G and C nucleotides suggested that there would be a sufficient number of CpG residues to make the results of analysis meaningful. Having written the computer program we were able to analyse many more cDNA sequences with a high (G + C) content, especially useful being a compilation of human genes containing CpG islands [9]. For a more uniform comparison our discussion will, therefore, focus on the human sequences.

Results from some of our analyses are given in Table 1. What had struck us originally with mouse cytoplasmic γ-actin was the much lower frequency of occurrence of CpG in the {2,3} codon position than in the {3,1} codon position, and it in can be seen that in the human actins this is most marked for the cytoplasmic β-actin. It should be stressed that the correction for third position G and C de-emphasizes rather than exaggerates this effect: the uncorrected ratios of actual/expected (not included in Table 1 for compactness) were 22% and 149% for the {2,3} and {3,1} positions, respectively (see also the simple incidence graph, Fig. 2a). The high incidence of C and, even greater, G in the third codon positions in the rest of this and other sequences is the

Table 1
Occurrence of CpG with respect to codon position in some cDNAs

	Number CpG		CpG actual/expected (corr.)		Ratio
	{2,3}	{3,1}	{2,3}	{3,1}	{3,1}/{2,3}
(a) Human actins					
cytoplasmic-β	5	50	14%	60%	4.3
cytoplasmic-γ	13	47	36%	60%	1.7
skeletal muscle-α	17	66	49%	80%	1.6
cardiac muscle-α	5	19	17%	24%	1.4
smooth muscle-α	5	15	18%	21%	1.2
smooth muscle-γ	3	10	10%	12%	1.3
(b) Other actins					
mouse cytoplasmic-β	2	30	6%	41%	7.1
chick cytoplasmic-β	2	17	7%	24%	3.6
Aspergillus nidulans	5	56	21%	80%	3.8
silkworm	11	67	41%	110%	2.7
(c) Human high CpG					
histone H4	1	24	18%	68%	3.8
glc 6-P dehydrogenase	7	47	18%	46%	2.6
thymidine kinase	6	20	24%	58%	2.4
Gx-α	8	35	33%	59%	1.8
ζ-globin	11	27	57%	101%	1.8
histone H3	7	15	60%	102%	1.7
histone H1	13	14	43%	63%	1.5
hsp 70	36	92	52%	71%	1.4
serotonin receptor	25	43	58%	65%	1.1
c-myc	25	43	62%	65%	1.0
tubulin-β	24	56	62%	64%	1.0
thrombmodulin	48	106	81%	83%	1.0
Transcription factor IL6	56	65	104%	103%	1.0
c-jun	33	41	88%	83%	0.9
α2-adrenergic receptor	51	82	111%	104%	0.9
oxytocin-neurophysin	10	23	93%	81%	0.9
somatostatin receptor	35	48	78%	67%	0.9
neurotrophin 3	15	15	79%	66%	0.8
histone H2B	7	7	170%	58%	0.3
(d) Human low CpG					
ubiquitin	3	12	19%	37%	1.9
triose phosphate isomerase	6	19	26%	36%	1.4
p53	9	13	23%	30%	1.3
elongation factor 1α	5	12	18%	16%	0.9
ribosomal protein S8	4	8	36%	31%	0.9
enolase	7	12	19%	12%	0.6

The source of the original sequences is as in Fig. 2.

reason for the difference between uncorrected and corrected figures.

There are five other isoforms of mammalian actin, differing by less than 7% in amino acid sequence, and each was subjected to analysis (Table 1). Cytoplasmic γ-actin showed a quite similar bias to CpG in the {3,1} position as β-actin (although involving different actual residues), as did α-skeletal actin. However, α-cardiac actin and α-smooth muscle actin (Fig. 2c), although still maintaining some bias, were considerably more deficient in CpG in the {3,1} position, and γ-smooth muscle actin very much more so. The murine actins showed a pattern very similar to the human actins, although again involving distinct CpG dinucleotides (e.g. Fig. 2b for mouse cytoplasmic β-actin). When more distant species were considered different trends were seen. Chicken cytoplasmic β-actin cDNA has an apparently high {3,1} : {2,3} ratio, but in the context of a much lower {3 : 1} CpG incidence than for the mammalian actins (Table 1). A cytoplasmic γ-type actin of *Aspergillus nidulans*, however, shows a very pronounced bias towards {3,1} CpG (Fig. 2d), as does a silkworm cytoplasmic actin (Table 1).

It became apparent in examining mammalian cDNA sequences for other proteins that normal or near normal incidence of CpG was uncommon, so we have focussed our attention on compilations of sequences with high GC-contents or, especially, those reported to contain CpG islands. Even among the latter, many sequences were deficient in CpG (e.g. the CpG islands were restricted to the promoter region), and we have only listed some of these in Table 1 by way of illustration. Among those that we have listed as having a high CpG-content (at least in the {3,1} position), the majority do not show a marked bias of {3,1} CpG over {2,3} CpG dinucleotides (e.g. Fig. 2f). However, some examples were identified (Fig. 2e and Fig. 2g–j). One with a highly conserved amino acid sequence (and hence most suitable to our analysis) is histone H4. This is an quite an interesting example, as histones H1 and H2B do not show a similar bias, and that in histone H3 is less extreme (Table 1). Other cDNAs identified with a {3,1} CpG bias are those for human glucose 6-phosphate dehydrogenase, thymidine kinase, ζ-globin and the G-protein, Gx-α.

Finally it should be emphasized that the results in Table 1 do not purport to be representative of human coding sequences. The purpose of the table is to present that small minority of coding sequences that we have found to show a CpG bias similar to the cytoskeletal actin genes, with only a few examples of the predominating other types of sequence included as comparators.

5. Hypothesis for strand-specific mutation

There have been previous suggestions to explain a global higher occurrence of CpG in the {3,1} position compared with the {2,3} position [4]. These have been in terms of translational requirements: either optimization for a complement of tRNAs or the avoidance of 'sticky' codons (those with too great hydrogen-bonding to the anticodon). Our analysis of individual sequences has revealed a special group of extreme cases, for which such explanations do not provide an answer. This is well illustrated by the differences in the various similarly expressed histones of a single species.

We suggest a different explanation, starting from the premise that there is a tendency for CpG to mutate. Our key point is that the self-complementary nature of the CpG dinucleotide means that one has to consider mutation of CpG from both strands of DNA, and that the position occupied by the cytosine residues on the two different strands of any given CpG dinucleotide will differ with respect to the reading frame. The observed different distribution of CpG dinucleotides in relation to codon position can therefore be accounted for by a different mutation rate from the two DNA strands. This is explained diagrammatically in Fig. 3. If one considers both strands of the DNA duplex there are six possible positions for a cytosine reside to be mutated in a CpG dinucleotide in relation to the reading frame. Four of these involve transition at codon positions 1 or 2, which are assumed to be selected against as they would alter the amino acid sequence. To simplify the diagram we have not illustrated the two excluded mutations from {1,2} CpGs, which we have not considered in this paper. The two from the {2,3} and {3,1} position have been marked 'excluded', and — for greater clarity — the products are not shown. When the two remaining possibilities are considered, it can be seen that to maintain the amino acid sequence, mutation of a CpG cytosine in the {3,1} codon position is only allowed from the sense strand, whereas mutation from the {2,3} codon position is only allowed from the anti-sense strand. Thus we suggest that in those sequences observed to have a marked {3.1} bias there is a lower rate of CpG mutation from the sense strand than from the anti-sense strand.

It should be mentioned that this is not the first time an inequality of mutation rate between the two strands of DNA has been suggested. We have become aware that such a suggestion has been made previously for primate globins [10]. However, in that case — which was not concerned with the reading frame or restricted to a consideration of CpG — the suggestion was the opposite of that made here: that there was a lower general rate of mutation from the anti-sense strand.

6. Discussion

A recent paper by Karlin and Burge [2] has also analysed CpG frequency in relation to codon position in a much larger number of human genes than considered here, and found that, overall, the relative abundance of CpG dinucleotides in the {2,3} codon position was lower than that in the {3,1} codon position, although both values were less than the 'expected' value of 1 (0.39 and 0.47, respectively). It is therefore pertinent to explain the difference in objectives and approach of that study and the work presented here. Karlin and Burge were interested in obtaining statistically significant global averages for dinucleotide frequencies, making it necessary for them to sample a large number of sequences, and appropriate for them to use average observed frequencies of C and G in the different codon positions as a basis of their calculations. In contrast, the work presented here addresses a minority of individual cases in which there is extreme CpG bias with respect to the {2,3} and {3,1} codon positions. After our initial observation of this in the mouse cytoskeletal actins, our interest was to see whether this phenomenon extended to any other genes, and were therefore obliged to focus on databases of GC-rich and CpG-rich genes. The vast majority of individual coding sequences are poor in CpG dinucleotides, and although it is valid to include them in a global analysis of the type conducted by Karlin and Burge [2], it would be statistically invalid to try to make conclusions about CpG bias with respect to codon position in such individual cases. It was this focus on *individual* genes that accounts for the different basis of our calculations. We adopted the standpoint that the occurrence of nucleotides in position 2 was determined by the individual amino acid sequence, and, as explained in section 3, used the observed overall frequency of each nucleotide in position 3 as the basis of calculation of expected frequency at that position.

Our premise that there is a tendency for CpG to mutate is obviously influenced by the well-known spontaneous deamination of vertebrate 5-methyl cytosine (found at CpG dinucleotides) to thymidine. Thus, for vertebrates it might be possible to extrapolate our suggestion to one of strand-specific hemimethylation, although there are no experimental data bearing on this. Caution is required in this respect, however, because of the finding of a similar CpG bias towards the {3,1} position in certain invertebrate actins, the DNA of which is not methylated [11]. Nevertheless, the rapid mutation of CpGs in the mouse γ-actin pseudogenes, and the increase in incidence of CpA and TpG where there is a CpG deficiency supports our premise of a mutational phenomenon. Although there are a number of enzymes that handle the two strands of the DNA duplex differently, the restricted nature of the phenomenon we have identified implies there are special circumstances in these cases. What these special circumstance are remains an open question.

References

[1] Cross, S.H. and Bird, A.P. (1995) Curr. Opin. Genet. Dev. 5, 309–314.

[2] Karlin, S. and Burge, C. (1995) Trends Genet. 11, 283–290.

[3] Peter, B. (1988) Ph.D. Thesis, Glasgow University.

[4] Schorderet, D.F. and Gartler, S.M. (1992) Proc. Natl. Acad. Sci. USA 89, 957–61.

[5] Bernadi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985) Science 228, 953–958.

[6] Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucl. Acids Res. 12, 387–395.

[7] Peter, B., Man, Y.M., Begg, C.E., Gall, I.G. and Leader, D.P. (1988) J. Mol. Biol. 203, 665–675.

[8] Herman, I.M. (1993) Curr. Opin. Cell Biol. 5, 48–55.

[9] Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) Genomics 13, 1095–107.

[10] Wu, C.-I. and Maeda, N. (1987) Nature 327, 169–170.

[11] Fidel, S., Doonan, J.H. and Morris, N.R. (1988) Gene 70, 283–293.