J Mol Evol (1994) 38:36–40

Journal of Molecular Evolution

Possible Implications of CpG Avoidance in the Flatworm *Schistosoma mansoni*

Héctor Musto,[1] Helena Rodríguez-Maseda,[2] Fernando Alvarez,[2,3] José Tort[1]

[1] Departamento de Bioquímica, Facultad de Ciencias, Tristán Narvaja 1674, Montevideo 11200, Uruguay
[2] Cátedra de Genética, Facultad de Medicina, Montevideo, Uruguay
[3] Departamento de Genética, Facultad de Ciencias, Montevideo, Uruguay

Received: 1 April 1993/Accepted 16 April 1993

Abstract. We report the analysis of the biases of CpG, TpG, and CpA of all the DNA sequences data from the Trematode *Schistosoma mansoni*. Our results show CpG avoidance whereas TpG and CpA frequencies are over the expected values. These characteristics are similar to the biases displayed by methylated genomes, but in platyhelminths 5mC has not been detected by biochemical methods. The possible implications of this CpG shortage are discussed.

Key words: CpG shortage — Methylation — Genome organization — *Schistosoma mansoni* — Platyhelminths

Introduction

It has been recognized for a long time that the dinucleotide CpG is present in vertebrate DNA less frequently than would be expected from base composition (Josse et al. 1961; Swartz et al. 1962). Also, TpG and CpA are significantly elevated with respect to expected values calculated as random base associations. These findings have been explained by the spontaneous deamination of 5mC generating the transition mutation products TpG and CpA (Salser 1977). On the other hand, animals with no CpG shortage do not display TpG and CpA excess (Bird 1980, 1983).

In vertebrates the analysis of DNA sequence data has shown that CpG is disfavored while TpG and CpA are both over the expected values (Nussinov 1984; McClelland and Ivarie 1982; Hanai and Wada 1988, 1990), not only in untranslated regions (5′, 3′ and introns), but even in the three codon positions of translated DNA. Again, these data have been interpreted in terms of the presence of 5mC (McClelland and Ivarie 1982). Furthermore, it was recently postulated that the CpG mutations to TpG in actin genes in lower eukaryotes were caused by methylation (Drouin 1991) as in vertebrate genes (Savatier et al. 1985; Green et al. 1990).

On the other hand, nonmethylated genomes like arthropods or mitochondria display non-CpG shortage (*Drosophila*) and TpG and CpA around the expected values (both genomes) (Nussinov 1984; Hanai and Wada 1990), so it might be concluded that the simultaneous presence of CpG avoidance and TpG + CpA excess is a characteristic of methylated genomes. Using biochemical methods like restriction enzymes and HPLC, 5mC was not detected in the platyhelminths *Schistosoma mansoni* and *Spirometra mansonoides* (Simpson et al. 1982; Cox et al. 1990).

Correspondence to: H. Musto, Departamento de Bioquímica, Facultad de Ciencias, Tristán Narvaja 1674, Montevideo 11200, Uruguay

In this paper we report the CpG, TpG, and CpA biases of all the DNA sequence data from *S. mansoni*. Contrary to the expectations, our results show CpG shortage and TpG and CpA frequencies over the expected values in almost all regions considered. These characteristics are very similar to the biases found in methylated genomes, so taking the possibility of methylation into account, two hypothesis can be considered: (1) there are indeed very little amounts of 5mC in *S. mansoni*, and (2) these genomes were methylated in the past. A further hypothesis has nothing to do with 5mC but takes into account the different C+G levels of the DNA regions that contain the sequences analyzed, since we found that the observed frequency of CpG increases linearly with the C+G content of the surrounding sequences and introns. This is in line with the observed increased frequency of CpG going from C+G-poor to C+G-rich isochores both in mammals and plants (Bernardi et al. 1985; Montero et al. 1990), so we postulate that the biases found in this organism are due to the constraints determined by the genome organization, as might happen even in methylated genomes.

Materials and Methods

Sequences were classified as protein coding, repeated sequences, and ribosomal DNA according to the definitions of the authors. In the case of protein coding, these were further divided in 5′ untranslated (5′UT), 3′ untranslated (3′ UT), translated, and introns. Only different sequences were analyzed. In total we obtained 37 coding sequences comprising 40,060 base pairs (bp), 28 5′UT (4,187 bp), 32 3′ UT (7,329 bp), and introns of four genes (5,447 bp). Furthermore, we analyzed three different repeated sequences (938 bp) and three ribosomal sequences (2,560 bp).

The expected dinucleotide values were calculated by dividing the product of the mononucleotide occurrences by the number of dinucleotides. The significance of the deviations of observed (O) from expected (E) values was measured using the χ^2 test calculated as $(O - E)^2/E$. As there is one degree of freedom, $\chi^2 > 4$ is significant at $P < 0.05$; $\chi^2 > 7$ is significant at $P < 0.01$; $\chi^2 > 11$ is significant at $P < 0.001$; and $\chi^2 > 15$ is significant at $P < 0.0001$ (Johnson 1990).

All the sequences were extracted from GenBank (Release 74.0) using the ACNUC retrieval system (Gouy et al. 1984).

Results and Discussion

CpG, TpG, and CpA Biases in Translated DNA

Table 1 displays all the translated sequences analyzed together with the O/E ratios for the considered dinucleotides and the C+G content of each sequence. It can be seen that CpG is diminished in relation to the expected values in 28/37 genes, and, on the other hand, TpG O/E value is always over 1, and CpA is increased in 33/37 sequences. Furthermore, in 27/28 of the sequences where there is a CpG shortage, TpG and CpA are simultaneously over the expected values. When all the sequences are analyzed together (see Table 2) it can be seen that all deviations are statistically significant.

This is the situation expected in methylated genomes, since in animals 5mC is present only in the dinucleotide CpG, and when 5mC mutates to T it gives rise, after a round of duplication without repair, to TpG in one chain of the DNA molecule and CpA on the other (Salser 1977; Bird 1980); this has been taken as an explanation for the CpG avoidance found in mammals translated DNA (McClelland and Ivarie 1982).

Table 1. List and dinucleotides ratios of translated genes investigated

Genes	CpG	TpG	CpA	C + G
SCMCTSB	0,80	1,44	1,13	40,67
SCMEGFRA	0,74	1,39	1,20	38,57
SCMFABP14	1,01	1,45	1,11	41,04
SCMGLUPER	0,83	1,15	1,06	41,37
SCMGPROTEI	1,03	1,17	0,99	40,96
SCMGSTM	0,78	1,39	1,05	42,77
SCMHGBA	0,78	1,33	1,23	34,11
SCMHGPRT	0,98	1,46	0,92	39,25
SCMHMGCOB	0,83	1,24	1,24	36,31
SCMSAT1A	0,78	1,38	1,30	43,14
SCMANT70	0,95	1,36	1,19	47,07
SCMANTIH	0,66	1,57	1,07	35,86
SCMANTTEG	0,88	1,27	1,36	30,53
SCMCALPAIN	0,77	1,31	1,19	35,66
SCMVACAG	0,67	1,22	1,15	27,67
SCMHSP86	0,84	1,23	1,10	43,23
SCMCBPB	0,56	1,42	1,33	37,14
SCMSMA	1,05	1,31	0,94	33,70
SCMCPROT	0,73	1,46	1,46	48,55
SCMPRSM	0,87	1,38	1,19	35,03
SCMCHRA	0,67	1,34	1,51	53,37
SCMP48EGG	1,13	1,36	1,21	34,94
SCMMEGA	0,98	1,31	1,39	47,93
SCMFERA	0,79	1,25	1,18	39,85
SCMFERB	0,98	1,23	1,32	36,80
SCMFS800	1,01	1,17	1,32	27,87
SCMFSPA	0,58	1,32	1,63	48,70
SCMMYHC	1,01	1,51	1,01	28,86
SCMPMYA1	1,16	1,30	1,12	31,14
SCMTPM	1,10	1,41	0,80	38,97
SCMNPK	0,86	1,26	1,15	35,36
SCMSMC74	0,80	1,38	1,11	39,25
SCMSMOXI	1,05	1,36	1,34	28,21
SCMSMOXIII	0,70	1,26	1,22	29,94
SCMSODM	0,81	1,13	1,43	40,36
SCMIMP23A	0,67	1,46	1,20	42,51
SCMSM24A	0,67	1,54	1,28	31,42

Translated sequences analyzed: Obs/Exp values are displayed for each dinucleotide and each gene (C + G is the %C + G in the corresponding sequence)

Table 2. Dinucleotides ratios in translated genes

Din	Obs	Exp	Obs/Exp	χ^2
CpG	1216	1434,89	0,85	33,39
TpG	3216	2434,52	1,32	250,86
CpA	2634	2246,26	1,17	66,93

Total values for all the sequences considered in Table 1 (for significance of χ^2, see Materials and Methods)

Table 3. Dinucleotides ratios in flanking regions and introns

	Din	Obs	Exp	Obs/Exp	χ^2
a) 5′ UT	CpG	102	108,86	0,94	0,43
	TpG	219	203,76	1,07	1,14
	CpA	310	253,19	1,22	12,75
b) 3′ UT	CpG	99	139,91	0,71	11,96
	TpG	431	336,86	1,28	26,31
	CpA	449	389,12	1,15	9,21
c) INTRONS	CpG	123	172,79	0,71	14,35
	TpG	404	350,01	1,15	8,33
	CpA	339	275,53	1,23	14,62

The same as Table 2 but for (a) 5′ UT, (b) 3′ UT, and (c) introns

Table 4. Dinucleotides ratios in repeated sequences and rDNA

	Din	Obs	Exp	Obs/Exp	χ^2
a) Repeated sequences	CpG	34	41,76	0,81	1,44
	TpG	95	59,30	1,60	21,50
	CpA	84	54,74	1,53	15,64
b) rDNA	CpG	166	167,49	0,99	0,01
	TpG	209	205,24	1,02	0,07
	CpA	129	144,27	0,89	1,62

The same as Table 2 but for (a) repeat sequences and (b) rDNA

Dinucleotide Biases in 5′ UT, 3′ UT, and Introns

Table 3a shows the CpG, TpG, and CpA biases found in the 5′ UT region. Although the figures must be taken into account with care, because most analyzed sequences are derived from cDNA clones and the number of nucleotides is not big enough, it can be seen that there are nonsignificant CpG shortages and TpG excess, and CpA is the only significant deviated doublet. These data are rather coincident with the non-CpG shortage found in the 5′ region of mammals and plant housekeeping genes, the so-called CpG islands, which constitute regions of nonmethylated DNA associated with genes (Bird 1986; Aïsani and Bernardi 1991; Antequera and Bird 1988).

On the other hand, (Table 3b), it can been seen that there are significant deviations of the three dinucleotides in the 3′ UT region, which displays the same biases that are found in an "average" mammalian gene (McClelland and Ivarie 1982), as happens with introns. (See Table 3c.)

Biases in Repeated Sequences and rDNA

Table 4a shows that in repeated sequences CpG is around 20% diminished in relation to expected values; this deviation is nonsignificant, very probably because of the rather small number of observed and expected CpGs; and again, TpG and CpA are significantly overrepresented. This could be explained assuming that, as happens in mammals, repeated sequences are highly methylated (Ehrlich et al. 1982).

On the other hand, there is no CpG shortage in rDNA (Table 4b), which coincides with the behavior of this doublet in rDNA from humans and toads (our results, not published).

CpG Frequency and C+G Content of UT Regions and Introns

The analysis of CpG, TpG, and CpA biases in the different regions is coincident with the expectations (and behaviors found) in methylated genomes. But in mammals it has been found that CpG shortage usually is strong in genes embedded in C+G-poor isochores, while it becomes weaker and almost disappears in genes localized in the C+G-richest isochores (Bernardi et al. 1985); in other words, CpG shortage depends on the C+G levels of the isochores which contain the genes.

Although, as noticed, the 5′ and 3′ data of *S. mansoni* are not large enough, we decided to analyze the correlation between the frequency of CpG and the C+G content of the available UT regions and introns. Figure 1 shows unequivocally that as long as the C+G content of UT regions and introns becomes greater, the CpG shortage becomes weaker.

Conclusions

As we have shown, the CpG, TpG, and CpA biases in all the regions analyzed in *S. mansoni* are compatible with the biases displayed by methylated genomes like that of mammals, since according to one current opinion (Salser 1977; Bird 1980), CpG avoidance is mainly due to the spontaneous deamination of 5mC to T giving rise to TpG and CpA.

On the other hand, Ohno has suggested that CpG/TpA avoidance and TpG/CpT/CpA excess are not related to CpG methylation and decay but are due to a universal rule that applies both to coding and noncoding sequences (Ohno 1988; Yomo and Ohno 1989). Our results do not contradict his opinion, but we should remark: (1) In a pool of nonver-

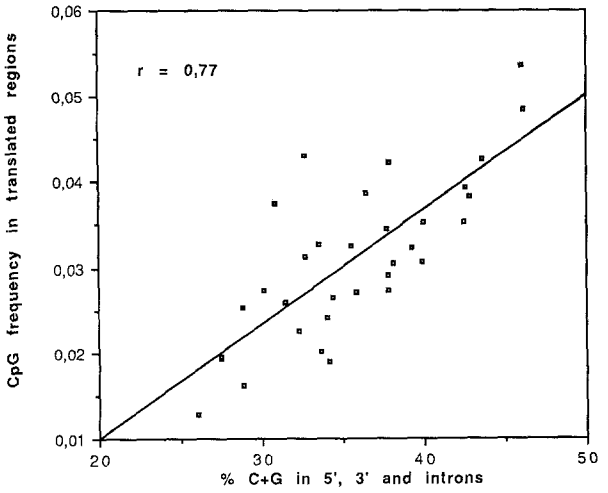

Fig. 1. Plot of CpG frequency in each translated sequence vs the C + G content of the corresponding 5′ + 3′ + introns, in each case, when available. r is coefficient of correlation.

tebrate sequences (poorly or not methylated at all), comprising 50 different sequences and 56,819 bp, CpG is only slightly diminished and TpG and CpA are around the expected values (Nussinov 1984), and (2) mitochondria genome (again nonmethylated) displays CpG avoidance but TpG and CpA are around the expected values (Nussinov 1984; Hanai and Wada 1990; and our results, not published). So we think that although the universal rule does apply to methylated genomes, more data on unmethylated organisms are needed to discard the role of 5mC in generating the CpG, TpG, and CpA biases. We think that finding, at the same time, CpG shortage and TpG + CpA excess is the fundamental argument for the role of deamination of 5mCpG in generating the biases.

The situation in *S. mansoni* is rather different, since it displays the biases but is not methylated (Simpson et al. 1982; Cox et al. 1990). Explaining our findings in accordance with the most accepted hypothesis (Salser 1977; Bird 1980) requires one to assume either (1) past methylation (and the biases are the "footprints") or (2) tissue- (germline) or (3) stage-specific methylation. These explanations seem unlikely because (1) one should explain why 5mC disappeared, (2) HPLC is powerful enough to detect low levels of 5mC, and (3) the digestion of nuclear DNA of *Fasciola hepatica,* a species closely related to *S. mansoni,* shows identical patterns with *Hpa*II and *Msp*I in two different stages (our results, not published). So the biases we found very probably have nothing to do with the presence of 5mC. Indeed, the fact that CpG frequency in exons increases linearly with the C + G content of the DNA harboring the sequence (Fig. 1) indicates that the CpG shortage in coding sequences is neither random nor constant but depends on the C + G content of the DNA harboring the sequence. An identical situation is found in mammals and plants (Bernardi et al. 1985; Montero et al. 1990), where methylation and the organization of the genome in isochores are both present, so we postulate that the observed CpG biases in *S. mansoni* are the consequence of the existence of different C + G levels in different DNA regions. This implies an isochore genome organization in this parasite.

Another consequence of our findings is that clearly it remains to be established how much of the biases found in methylated genomes are due to the deamination of CpG and how much to other factors like compositional constraints. What is clear, however, is that finding the biases is not enough of an argument to assume the presence of 5mC.

Acknowledgments. We wish to thank Drs. Ricardo Ehrlich and Ekaterina Scvortzoff for discussions, and specially Prof. Giorgio Bernardi for helpful suggestions and critically reading the manuscript. Part of this work was supported by Proyecto de Desarrollo de Ciencias Básicas (PEDECIBA), Universidad de la República, Uruguay.

References

Aïsani B, Bernardi G (1991) CpG islands: features and distribution in the genomes of vertebrates. Gene 106:173–183

Antequera F, Bird AP (1988) Unmethylated CpG islands associated with genes in higher plant DNA. EMBO J 7:2295–2299

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm blooded vertebrates. Science 228:953–958

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8:1499–1504

Bird AP (1983) DNA modification. In: Maclean N, Gregory SP, Flavell RA (eds) Eukaryotic genes, their structure, activity and regulation. Butterworth & Co, London, pp 53–67

Bird AP (1986) CpG-rich islands and the function of DNA methylation. Nature 321:209–213

Cox GS, Phares CK, Schmidt RA (1990) Molecular characterization of the *Spirometra mansonoides* genome: renaturation kinetics, methylation, and hybridization to human cDNA probes. Biochem Biophys Acta 1049:134–144

Drouin G (1991) Non random CpG mutations affect the synonymous codon usage of moderately GC-rich single copy actin genes. J Mol Evol 33:237–240

Ehrilch M, Gama-Sosa MA, Huang L-H, Midgett RM, Kuo KC, McCune RA, Gehrke C (1982) Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic Acids Res 10:2709–2721

Gouy M, Milleret F, Mugnier C, Jacobzone M, Gautier C (1984) ACNUC: a nucleic acid sequence data base and analysis system. Nucleic Acids Res 12:121–127

Green PM, Montandon AJ, Bentley DR, Ljung R, Nilsson IM, Gianelli F (1990) The incidence and distribution of CpG → TpG transitions in the coagulation factor IX gene. A fresh look at CpG mutational hotspots. Nucleic Acids Res 18:3227–3231

Hanai R, Wada A (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

Hanai R, Wada A (1990) Doublet preference and gene evolution. J Mol Evol 30:109–115

Johnson AM (1990) Comparison of dinucleotide frequency and

codon usage in Toxoplasma and Plasmodium: evolutionary implications. J Mol Evol 30:383–387

Josse J, Kaiser AA, Kornberg A (1961) Enzymatic synthesis of deoxyribonucleic acid VIII. Frequencies of nearest neighbour base sequences in deoxyribonucleic acid. J Biol Chem 236:864–875

McClelland M, Ivarie R (1982) Asymmetrical distribution of CpG in an 'average' mammalian gene. Nucleic Acids Res 10:7855–7877

Montero LM, Salinas J, Matassi G, Bernardi G (1990) Gene distribution and isochore organization in the nuclear genome of plants. Nucleic Acids Res 18:1859–1867

Nussinov R (1984) Doublet frequencies in evolutionary distinct groups. Nucleic Acids Res 12:1749–1763

Ohno S (1988) Universal rule for coding sequence construction: TA/CG deficiency-TG/CT excess. Proc Natl Acad Sci USA 85:9630–9634

Salser W (1977) Globin mRNA sequences: analysis of base pairing and evolutionary implications. Cold Spring Harbor Symp Quant Biol 40:985–1002

Savatier P, Trabuchet G, Faure C, Cheboune Y, Gouy M, Verdier G, Nigon VM (1985) Evolution of the primate beta-globin gene region. High rate of variation in CpG dinucleotides and in short repeated sequences between man and chimpanzee. J Mol Biol 182:21–29

Simpson A, Sher A, McCutchan T (1982) The genome of *Schistosoma mansoni:* isolation of DNA, its size, bases and repetitive sequences. Mol Biochem Parasitol 6:125–137

Swartz MN, Trautner TA, Kornberg A (1962) Enzymatic syntheses of deoxyribonucleic acid XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. J Biol Chem 237:1961–1967

Yomo T, Ohno S (1989) Concordant evolution of coding and noncoding regions of DNA made possible by the universal rule of TA/CG deficiency-TG/CT excess. Proc Natl Acad Sci USA 86:8452–8456