J Mol Evol (1995) 40:343–346

JOURNAL OF MOLECULAR EVOLUTION

Letter to the Editor

Compositional Correlations in the Nuclear Genes of the Flatworm *Schistosoma mansoni*

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

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

Received: 4 October 1994

Abstract. We have investigated the genome organization in the flatworm *Schistosoma mansoni.* First, we analyzed the compositional distributions of the three codon positions. Second, we investigated the correlations that exist between (1) the GC levels of exons against flanking regions, (2) the GC levels of third codon positions against flanking regions, (3) the dinucleotide frequencies of exons against flanking regions, and (4) the GC levels of 5′ against 3′ regions. The modality of the distribution of third codon positions, together with the significant correlations found, leads us to propose that the nuclear genome of this species is compositionally compartmentalized.

Key words: Genome organization — Isochores — *Schistosoma mansoni* — Platyhelminths

In a recent paper (Musto et al. 1994) we have postulated that the CpG shortage in the translated regions of the flatworm *Schistosoma mansoni* (class Trematoda) is neither the result of the spontaneous deamination of the modified base 5mC (Salser 1977), as generally accepted (Bird 1980, 1983; McClelland and Ivarie 1982; Schorderet and Gartler 1992), nor due to a "universal rule" that applies to both coding and noncoding sequences (Ohno 1988; Yomo and Ohno 1989). On the contrary, we have suggested that the CpG avoidance is due to compositional constraints (Bernardi and Bernardi 1986) determined by different GC levels of the regions of DNA that contain the genes. Indeed, we have found that the CpG frequency in exons is significantly correlated with the GC level of the DNA harboring the sequence, as is the case in mammals and plants (Bernardi et al. 1985; Montero et al. 1990).

Our suggestion implies that the genome of *S. mansoni* is made up of isochores, or isochore-like structures, which are long, compositionally homogeneous DNA segments which can be subdivided into a small number of families characterized by different GC levels. Isochores have been studied in detail in vertebrates and plants (for reviews see Bernardi 1989, 1993) and have been demonstrated in the unicellular parasites *Plasmodium cynomolgi* (McCutchan et al. 1988) and *Trypanosoma brucei* and *T. equiperdum* (Isacchi et al. 1993). Furthermore, regional variations in GC levels have been detected in the chromosome III of the yeast *Saccharomyces cerevisiae* (Karlin et al. 1993; Sharp and Lloyd 1993). So, compositional compartmentalization appears to be a phylogenetically very widespread situation.

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

In a series of papers, Bernardi and his colleagues have demonstrated that in vertebrates compositional correlations exist between the GC level of exons (and especially third codon position) and the GC level of the isochores in which they are embedded, as well as between exons and the introns from the same genes. (See for instance Bernardi et al. 1985, 1988; Bernardi and Bernardi 1985, 1986; Aïssani et al. 1991; D'Onofrio et al. 1991; Mouchiroud et al. 1991.) Furthermore, the compositional distribution of third codon positions from compartmentalized nuclear genomes (like those of vertebrates) tends to be multimodal, while noncompartmentalized genomes (like those of most bacteria) generally display unimodal distributions (Bernardi 1993; D'Onofrio and Bernardi 1992).

All these features are due to the fact that exons (and their codon positions) and introns display the same GC bias of the isochore in which they are embedded. As a consequence, finding these compositional correlations and a nonunimodal compositional distribution (in particular of third codon positions) may be taken as indicative of a compartmentalized genome. However, it should be remarked that a nonunimodal distribution may have different causes which are independent of the level of compartmentalization of the genome. For instance, different GC levels in third codon positions might be due to different biases associated with the levels of gene expression and the actual populations of isoaccepting t-RNAs (Ikemura 1981a,b, 1982).

In this paper we have analyzed the compositional patterns and the compositional correlations of the nuclear genes of *S. mansoni*. For a description of the sequences analyzed, see Musto et al. (1994).

The compositional patterns of the three codon positions are shown in Fig. 1. The distribution of first codon positions (Fig. 1a) is multimodal with three major peaks at 35, 42.5, and 50% GC, respectively. Figure 1b shows that the distribution of second codon positions is unimodal with the highest concentration of genes at 32.5%, and trails both toward high and low values, ranging from 12.5 to 62.5% GC. Third codon positions (Fig. 1c) show two important features: first, 25% of the genes (9/37) have GC levels equal to or lower than 20%; and the highest value is only 50%; second, the distribution seems to be bimodal since two peaks are evident, one at 25% and the other at 40% GC.

The low GC levels of third codon positions may be a characteristic of flatworms, since the nuclear encoded genes of Turbellarians (free-living platyhelminths) tend to display this feature (J. García-Fernandez, J. Baguñà, and E. Salo, personal communication). On the other hand, the bimodality of the distribution is, as mentioned above, suggestive of the existence of compositional compartments. To investigate this possibility further, the compositional correlations of exons and third codon positions against flanking sequences (defined in Musto et al. 1994) were analyzed.

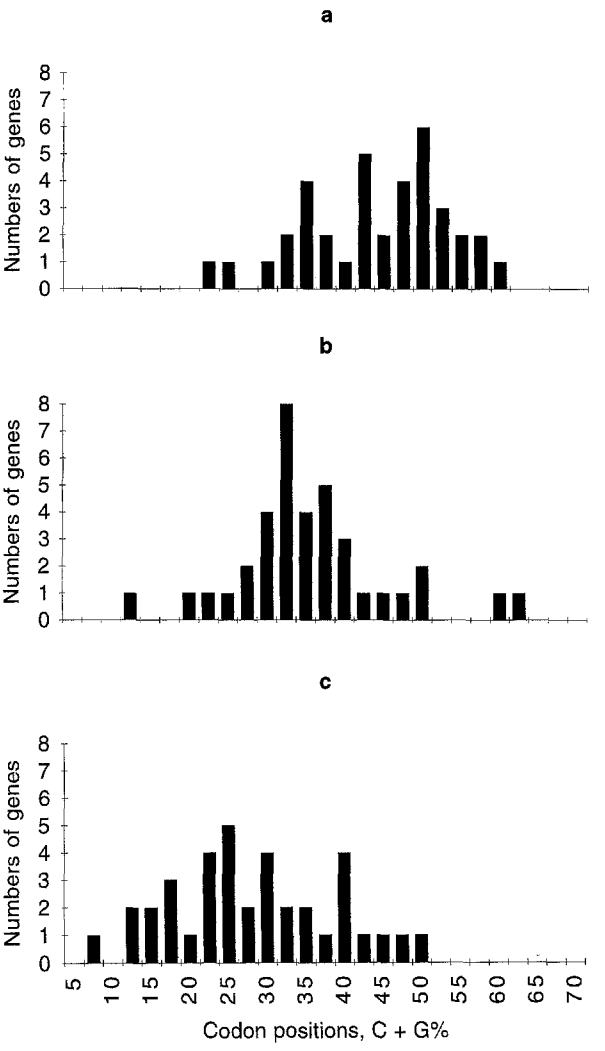

Fig. 1. Compositional pattern of the genome of *S. mansoni*. Histograms show the distributions of GC levels of first **(a)**, second **(b)**, and third **(c)** codon positions.

Figure 2a and b shows the plots of GC levels of exons and third codon positions, respectively, against the GC levels of the flanking regions, which comprise 16,963 bp. (For details, see Musto et al. 1994.) In Fig. 2a, the least-square line through the points exhibits a slope of 0.58, and the correlation coefficient of 0.50 is significant ($P < 0.01$). In Fig. 2b the slope is 1.21 and the correlation coefficient is 0.64 ($P < 0.001$). Although in the last plot a point (marked with an arrow) was not taken into consideration, it is evident that, as happens in compositionally compartmentalized genomes (see for instance Aïssani et al. 1991), a linear relationship exists between the GC levels of exons (and third codon positions) and the GC levels of the DNA regions harboring the sequences analyzed.

Further, the distribution of points and the slope greater than unity in the plot of Fig. 2b suggests that when coding sequences are embedded in GC-poor regions, their third codon positions are even poorer in GC, while when located in the GC-richest regions they attain

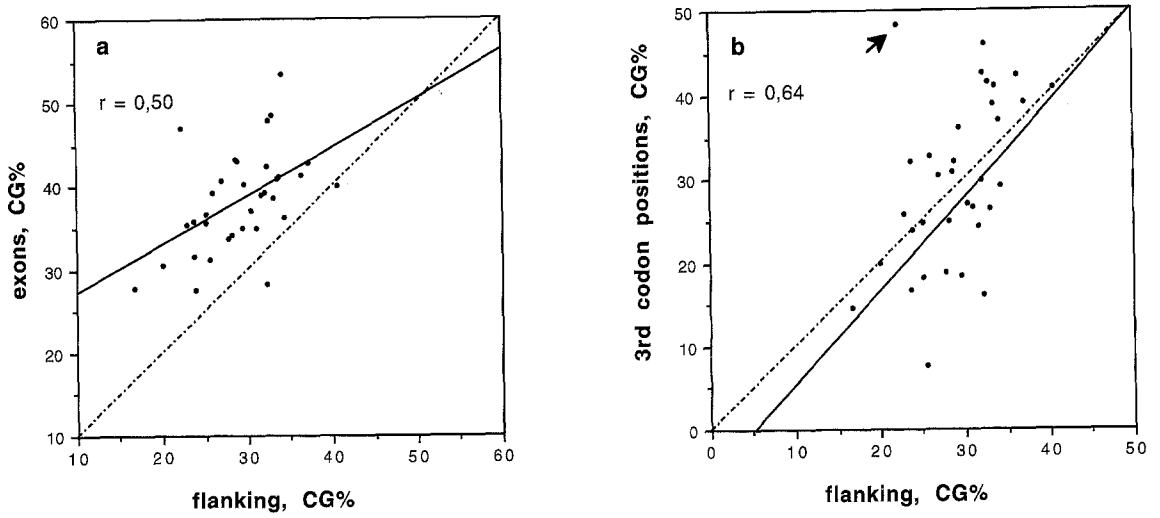

Fig. 2. Plots of GC levels of exons (**a**) and third codon positions (**b**) against the corresponding GC levels of the flanking regions. The correlation coefficients (r) are given. The point marked by an *arrow* in **b** was not considered in calculating r. The *diagonal line* (slope = 1) is indicated by a *point–dash line.*

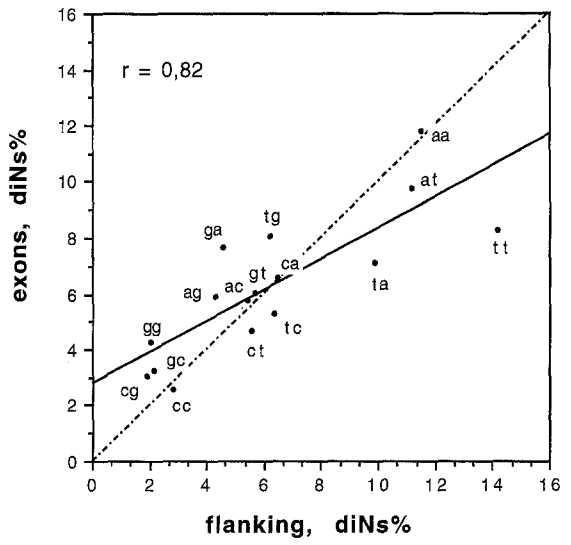

Fig. 3. Plot of dinucleotide (diNs) frequencies of exons against flanking regions. Each diN is indicated; r and *diagonal line* as in Fig. 2.

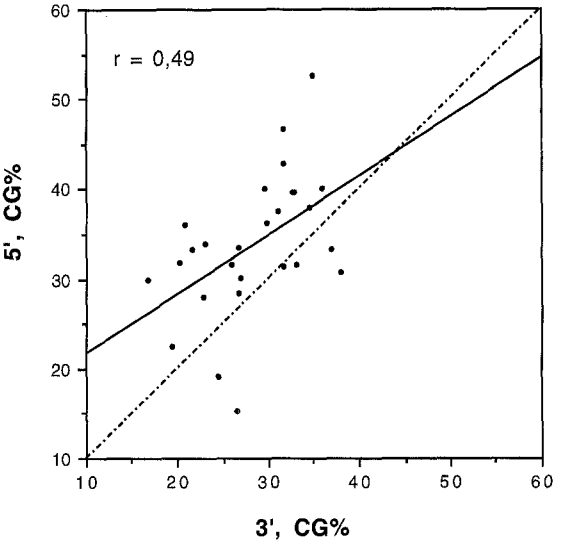

Fig. 4. Plot of GC levels of 5′ regions against 3′ regions. r and *diagonal line* are as in Fig. 2.

not only the highest GC values but are even higher than the surrounding sequences. A nearly identical situation is found in mammals (Aïssani et al. 1991).

More evidence of the homogeneous composition of the exons and their flanking regions is shown in Fig. 3, where the dinucleotide frequencies of the exons and the translated regions are plotted against each other. It can be seen that there is a highly significant correlation (r = 0.82, $P < 0.0001$), which suggests that the compositional homogeneity extends to other levels of orders such as dinucleotide usage and might not be limited to the base composition.

Figure 4 shows the plot of GC levels of 5′ against 3′ regions. The correlation coefficient is 0.49 ($P < 0.01$), which means that the compositional homogeneity includes the surrounding regions of the exons. This is an important point since it suggests that each independently considered region is compositionally correlated with the others. This point is confirmed by plots of GC levels of exons against the GC levels of 5′ and 3′ regions, since the correlation coefficients are 0.38 ($P < 0.05$) and 0.70 ($P < 0.0001$), respectively (data not shown).

In conclusion, we have found (1) that the compositional pattern of third codon positions is bimodal and (2) that there are linear and significant correlations among the GC levels of exons and third codon positions against the GC levels of flanking regions, between the dinucleotide frequencies of exons against flanking regions, and between the GC levels of 5′ against 3′. Further, in a previous paper (Musto et al. 1994), we found that the level of CpG in exons is linearly correlated with the GC levels of flanking regions.

All these results are compatible with our previous suggestion that the nuclear genome of *S. mansoni* is compositionally compartmentalized, although experimental work is needed to confirm this hypothesis. If it turns to be true, this proposal could be important for understanding the genomic organization of metazoans, since platyhelminths are in a special position from a phylogenetic point of view, being at the base of all triblastic animals displaying bilateral symmetry (Barnes 1977; Strickberger 1990).

Acknowledgments. We wish to thank Dr. Giorgio Bernardi for helpful discussions, encouragement, and critical reading of the manuscript. Part of this work was supported by Proyecto de Desarrollo de Ciencias Básicas (PEDECIBA) and Comisión Sectorial de Investigación Científica (CSIC), Universidad de la República, Uruguay. H.M. and H.R.-M. also thank the UNESCO/Third World Academy of Sciences for short-term fellowships.

References

Aïssani B, D'Onofrio G, Mouchiroud G, Gardiner K, Gautier C, Bernardi G (1991) The compositional properties of human genes. J Mol Evol 32:493–503

Barnes RD (1977) Zoología de los invertebrados, 3rd ed. Nueva Editorial Interamericana, México

Bernardi G (1989) The isochore organization of the human genome. Ann Rev Genet 23:637–661

Bernardi G (1993) The vertebrate genome: isochores and evolution. Mol Biol Evol 10:186–204

Bernardi G, Bernardi G (1985) Codon usage and genome composition. J Mol Evol 22:363–365

Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. J Mol Evol 24:1–11

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 228:953–956

Bernardi G, Mouchiroud D, Gautier C, Bernardi G (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol 28:7–18

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

D'Onofrio G, Bernardi G (1992) A universal compositional correlation among codon positions. Gene 110:81–88

D'Onofrio G, Mouchiroud D, Aïsani B, Gautier C, Bernardi G (1991) Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J Mol Evol 32:504–510

Ikemura T (1981a) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J Mol Biol 146:1–21

Ikemura T (1981b) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translation system. J Mol Biol 151:389–409

Ikemura T (1982) Correlation between the abundance of yeast tRNAs and the occurrence of the respective codons in protein genes. J Mol Biol 158:573–597

Isacchi A, Bernardi G, Bernardi G (1993) Compositional compartmentalization of the nuclear genes of *Trypanosoma brucei* and *Trypanosoma equiperdum.* FEBS Letters 335:181–183

Karlin S, Blaisdell BE, Sapolsky RJ, Cardon L, Burge C (1993) Assessments of DNA inhomogeneities in yeast chromosome III. Nucleic Acids Res 21:703–711

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

McCutchan T, Dame J, Gwadz R, Vernick K (1988) The genome of *Plasmodium cynomolgi* is partitioned into separable domains which appear to differ in sequence stability. Nucleic Acids Res 16:4499–4510.

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

Mouchiroud D, D'Onofrio G, Aïssani B, Macaya G, Gautier C, Bernardi G (1991) The distribution of genes in the human genome. Gene 100:181–187

Musto H, Rodríguez-Maseda H, Alvarez F, Tort J (1994) Possible implications of CpG avoidance in the flatworm *Schistosoma mansoni.* J Mol Evol 38:36–40

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

Schorderet D, Gartler S (1992) Analysis of CpG suppression in methylated and nonmethylated species. Proc Natl Acad Sci USA 89: 957–961

Strickberger MW (1990) Evolution. Jones and Barlett, Boston

Sharp P, Lloyd A (1993) Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure. Nucleic Acids Res 21:179–183

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