J Mol Evol (1993) 37:131–136

Journal of Molecular Evolution

Dinucleotides and G+C Content in Human Genes: Opposite Behavior of GpG, GpC, and TpC at II-III Codon Positions and in Introns

G. Gutiérrez,[1] J.L. Oliver,[2] A. Marín[1]

[1] Departamento de Genética, Universidad de Sevilla, Apartado 1095, E-41080 Sevilla, Spain
[2] Departamento de Genética, Universidad de Granada, Campus de Fuentenueva, E-18071 Granada, Spain

Abstract. We have studied the behavior of the dinucleotide preferences under G+C content variation in human genes. The doublet preferences for each dinucleotide were compared between two functionally distinct zones in genes, the II-III codon positions, and the introns. The 16 dinucleotides have been tentatively classified in three groups:

- AA, AC, CC, CT, and GA, doublets showing no difference between introns and II-III codon positions in the full range of G+C variation
- TG and TA, which differ in the full range of G+C variation
- AT, AG, GT, TC, TT, GG, GC, CG, and CA, which show differences in regions over 50% G+C

A remarkable pattern observed concerns the behavior of GG, GC, and TC, which showed opposite trends in II-III codon positions and in introns. If codon positions and introns are under the same structural requirements and the same mutational bias, our results indicate that the differences observed could be related to post-transcriptional constraints acting on mRNA.

Key words: Human genome — Dinucleotides — Introns — II-III codon positions — G+C content

Introduction

Extensive work carried out in a variety of genomes has revealed that the frequencies of occurrence of dinucleotides (nearest-neighbor base pairs or "doublets") is non-uniform. Rules accounting for doublet preferences have been developed (Nussinov 1980, 1981a,b, 1984a,b), and it has been proposed that the excess of CT and TG doublets accompanied by complementary deficiency of CG and TA dimers can be considered as the universal rule of coding-sequence construction (Ohno 1988). The origin of doublet preferences could be related to requirements for advantageous DNA structure (Nussinov 1984a), and to mutational biases in the replication/repair of DNA machinery which may be influenced by neighboring base effects (Bulmer 1986).

The individual doublet frequencies in a DNA segment are related to its base composition or G+C content. The variation in G+C content throughout the human genome covers a wide range; this variation is particularly pronounced at third-codon posi-

Presented at the NATO Advanced Research Workshop on *Genome Organization and Evolution,* Spetsai, Greece, 16–22 September 1992

Correspondence to: A. Marín

tions. A high correlation is found between the G+C content at third-codon positions of exons and the neighboring introns (Bernardi et al. 1985; Bernardi 1989; Aota and Ikemura 1986). The variation in doublet preferences with G+C content in human genes has been extensively studied by Hanai and Wada (1988) and Wada et al. (1991). These authors have confirmed the mosaic structure of the human genome and drawn consequences for doublet preferences in relation to G+C content—most notably, that as G+C content increases CG disfavor is attenuated at both II-III and III-I codon positions, while TA disfavor is enhanced.

An interesting observation in both prokaryotic and eukaryotic DNA sequences was that doublet preference in noncoding regions is correlated with that in coding regions (Nussinov 1981a,b). The correlation coefficients are decreasing from II-III to III-I to I-II (Hanai and Wada 1990), thus meaning that the doublet patterns are not dependent only on the coding capacity of the nucleotide chain, and that selective constraints decrease from amino acid composition (I-II) to codon context or amino acid juxtaposition (III-I) and lastly to codon usage (II-III).

This note intends a further step in the characterization of the mosaic domains of human DNA by disclosing which doublets in which compositional circumstances show a different behavior at II-III codon positions and in introns. In a first approach, it may be considered that introns and exons are subjected to the same mutational bias effects and DNA structure requirements. If this is true, the differences found between doublet preferences at II-III codon positions and within introns could be related to posttranscriptional constraints (RNA processing and translation/codon usage), and the changes that due take place are, of course, on DNA.

Data and Methods

The human sequences used in this study were retrieved from release 29 (December 1991) of the EMBL Nucleotide Sequence Data Library on CD-ROM (Stoehr and Cameron 1991). Analyses have been carried out on a random sample consisting of 1,208 complete coding sequences over 600 bp and all the 651 introns over 400 bp identified in entries tagged with the key "CDS" and the operator "join" in the Features Table. We have not concerned ourselves here either with redundancy of the database or with similarity between sequences due to homology.

To express the nature of the relationships between doublet preference and G+C content we proceeded as follows. In the first place, we computed in introns and at II-III codon positions the doublet preference (DP) of each doublet as defined by Hanai and Wada (1990):

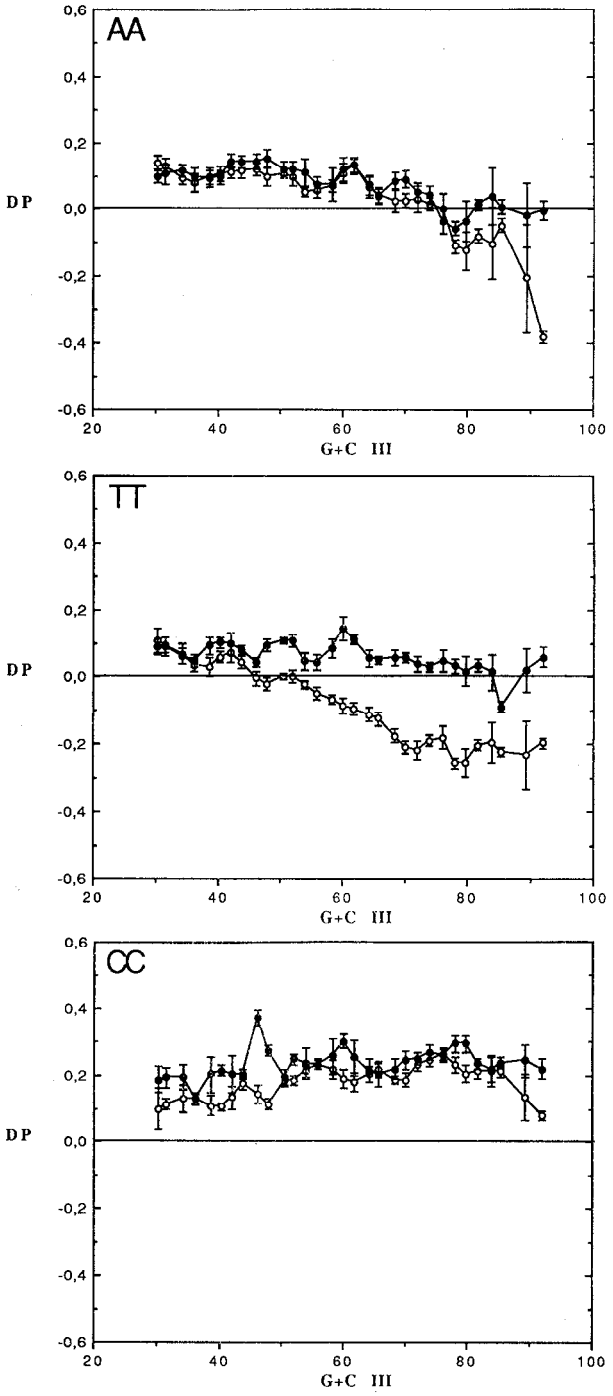

Fig. 1. Plots relating the variation of *DP* (vertical axis) to G+C content (G+CIII) measured at third-codon positions (horizontal axis) for the 16 nucleotides. (○) II-III codon position, (●) introns. Vertical bars indicate the standard error of the mean within each interval. For more information see Data and Methods. Continued on pages 133–135.

$$DP = ((\text{Observed frequency}) - (\text{random expectation}))/(\text{random expectation})$$

Random expectations were estimated on the basis of single nucleotide frequencies.

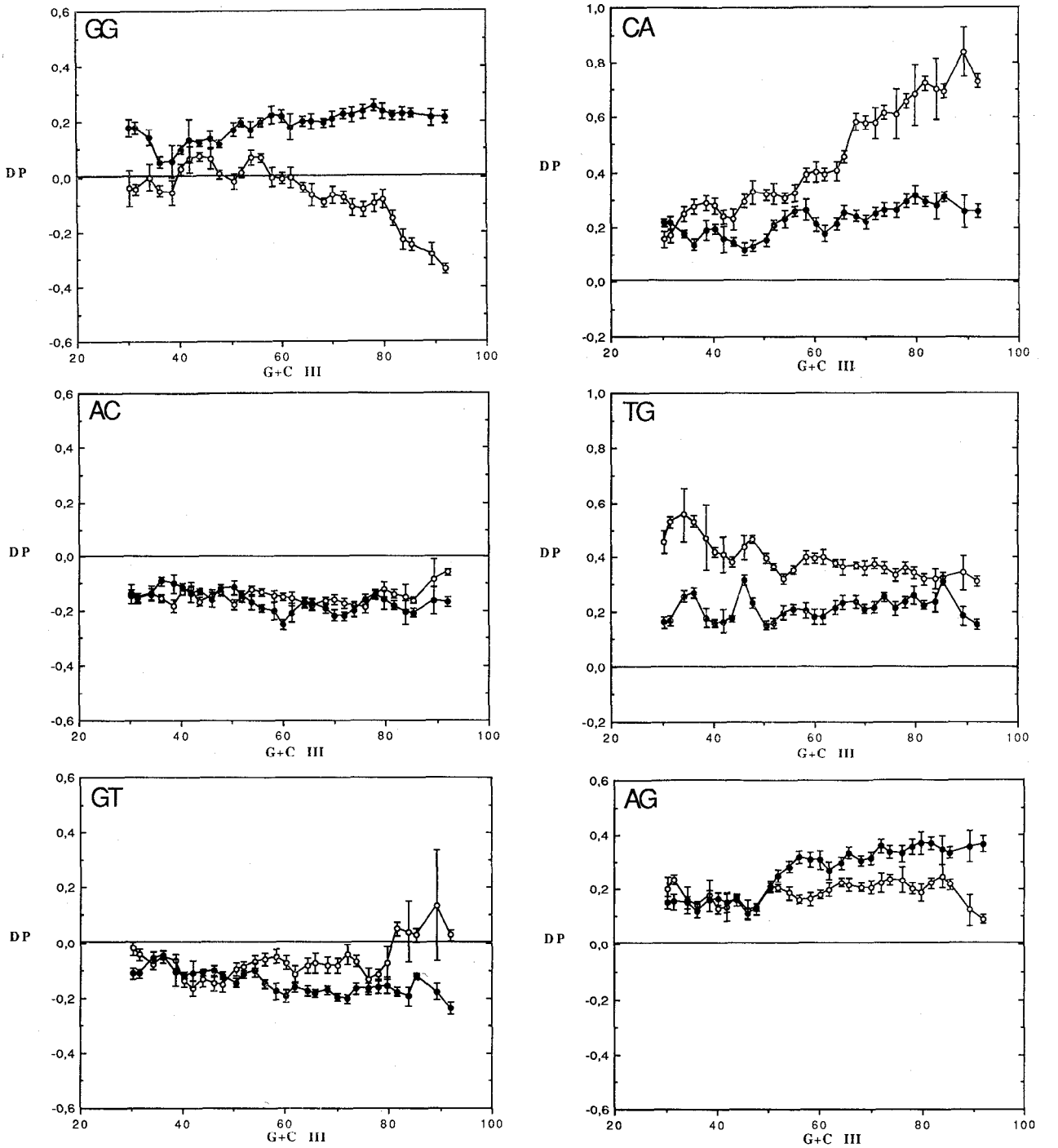

Fig. 1. Continued.

In the second place, for each doublet we made a diagram relating *DP* in introns and at II-III codon positions (vertical axis) to G+C content (horizontal axis). For this purpose, introns were sorted according to the G+C content of the third codon positions of genes harboring them. A window of 4% G+C size was slid across horizontal axis by steps of 2%. In each step the averaged values of *DP* and G+C content of the included segments were used as coordinates.

The program BMDP3S from the BMDP statistical package (Dixon and Brown 1979) was used to carry out Mann-Whitney non-parametric tests in comparing *DP* at II-III codon positions and in introns.

Results and Discussion

Plots of DP against G+C are given in Fig. 1. Our results relating II-III doublet preferences to G+C variation closely agree with those of Hanai and

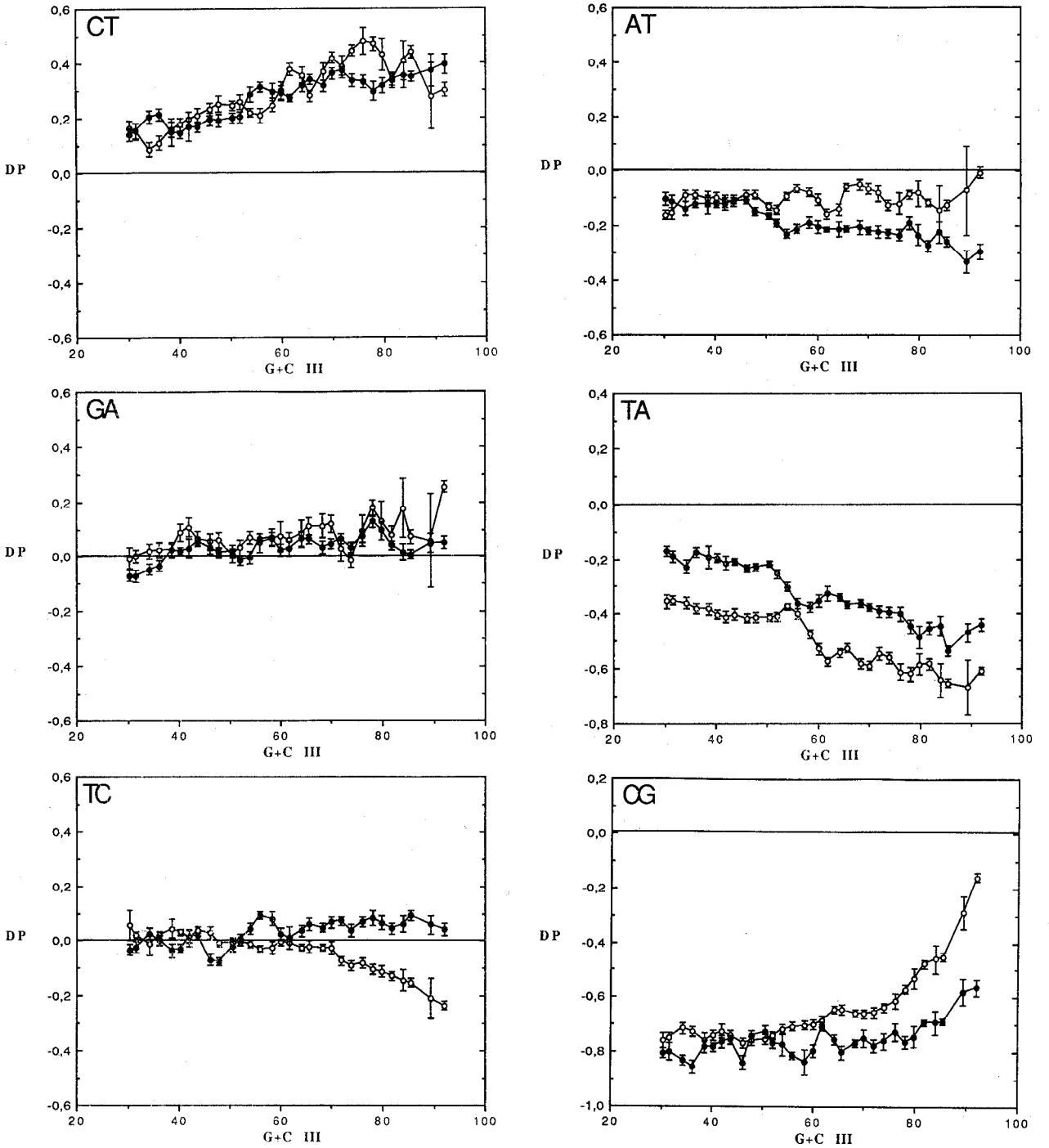

Fig. 1. Continued.

Wada (1988), while those for introns are first reported here.

In order to quantify the differences in *DP* between coding (II-III) and noncoding sequences, we have tested the statistical significance of *DP* differences between these subsets in intervals of 4% G+C content (Table 1).

According to the behavior through G+C variation and the number of G+C intervals with nonsignificant differences, the 16 doublets can be tentatively classified into three groups:

1. Doublets that do not show any difference in the full range of G+C variation: AA, AC, CC, CT, and GA (10 or more nonsignificant differences).
2. Doublets differing in the full range of G+C variation: TG and TA (one or two nonsignificant differences).

It is a general rule in nuclear DNA of vertebrates that TG is favored while TA is disfavored (Nussinov 1984b). Our results indicate that the excess of TG is more apparent at II-III than in introns, and TA is more disfavored at II-III.

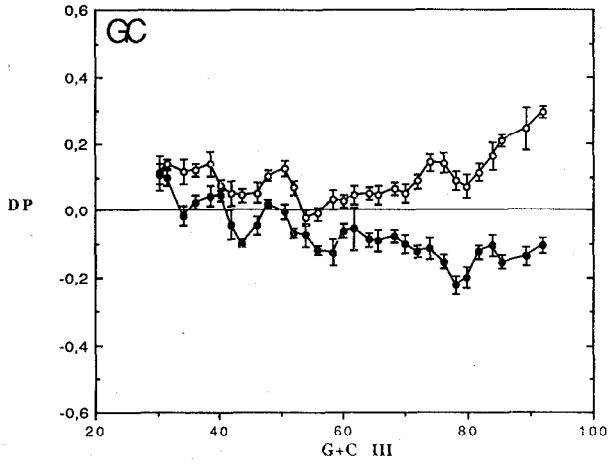

Fig. 1. Continued.

3. Doublets mostly differing in regions over 50% G+C: AT, AG, GT, TC, TT, GG, GC, CG, and CA (three or more nonsignificant differences clustered in G+C intervals under 56%). Within this group a clear pattern emerges distinguishing purine-purine (RR) and pyrimidine-pyrimidine (YY) doublets; thus RR and YY doublets (AG, GG, TC, TT) are favored in introns over II-III, and RY and YR doublets (AT, GT, GC, CG, CA) are favored in II-III over introns.

It has been noted that with the exception of TG, CA, and sometimes GC, RR and YY are more frequent than YR and RY in eukaryotes (Nussinov 1984a,b). Our results agree indicating that this effect is more pronounced in introns than in coding (II-III) sequences. However, it is noteworthy that on the one hand, at high G+C content there is a clear disfavor for GG, TC, and TT at II-III which does not operate in introns, and there is a disfavor for GC in introns which does not occur at II-III. On the other hand, the rule of TG and CA excess and TA deficiency is emphasized in coding sequences, while CG disfavor is emphasized in noncoding sequences. The differences here shown could be useful to identify introns and exons in nucleotide sequences.

Conclusions

We have found that similarities between II-III and introns in doublet preferences mostly occur at low G+C content. This is unlike the suggestion by Hanai and Wada (1988) that high G+C noncoding regions should have doublet preferences resembling those revealed in the II-III and III-I doublets at high-third-letter G+C.

It has been suggested that complementary doublets have in general similar frequencies of occurrence (Nussinov 1981b, Nussinov 1984a,b). This characteristic, which is found in the sample of introns, does not further hold at II-III. Particularly, CA/TG show opposite trends in the full range of G+C variation, and at high G+C content, three out of six complementary doublets—CC/GG, AA/TT, and GA/TC—also have different behavior.

Finally, it is worth noting that GG, GC, and TC show opposite trends in II-III and in introns. Although GG and GC are clearly more frequent in

Table 1. Statistical significance of comparisons between doublet preferences at II–III position and in introns using a Mann-Whitney nonparametric test [I = intervals of G + CIII; Ncs = number of coding sequences; Nin = number of introns. NS = nonsignificant; (*) = $P < 0.05$; (**) = $P < 0.01$]

			Group															
			1					2		3								
I	Ncs	Nin	AA	AC	CC	CT	GA	TA	TG	AT	AG	GT	TC	TT	GG	GC	CG	CA
<32	27	20	NS	NS	NS	NS	NS	**	**	NS	NS	**	NS	NS	**	NS	NS	NS
32–36	35	9	NS	NS	NS	*	*	**	*	NS	NS	NS	NS	NS	NS	NS	**	NS
36–40	63	39	NS	*	**	NS	NS	**	**	NS	*	NS	**	**	*	*	NS	*
40–44	90	26	NS	NS	*	NS	*	**	**	NS	NS	NS	NS	NS	NS	*	**	NS
44–48	62	13	NS	NS	**	NS	NS	**	*	NS	NS	NS	NS	NS	NS	NS	**	**
48–52	59	32	NS	NS	NS	NS	NS	**	**	NS	NS	NS	NS	**	**	**	NS	**
52–56	93	39	*	*	NS	NS	NS	*	**	**	**	NS	NS	**	**	NS	**	*
56–60	102	49	NS	*	NS	NS	NS	**	**	**	**	**	**	**	**	**	**	**
60–64	99	37	NS	NS	NS	*	NS	**	**	NS	NS	NS	NS	**	**	**	NS	**
64–68	101	79	NS	NS	NS	NS	NS	**	**	**	**	**	**	**	**	**	**	**
68–72	126	75	*	**	NS	NS	*	**	**	**	**	**	**	**	**	**	**	**
72–76	109	63	NS	NS	*	**	NS	**	**	**	**	**	**	**	**	**	**	**
76–80	115	72	NS	NS	*	**	NS	**	**	**	**	NS	**	**	**	**	**	**
80–84	75	67	**	*	NS	NS	NS	**	**	**	**	**	**	**	**	**	**	**
84–88	40	18	NS	NS	NS	NS	NS	NS	NS	NS	*	**	**	*	**	**	**	**
>88	12	13	NS	NS	*	NS	NS	NS	*	**	**	**	**	*	**	**	**	**

G + C-rich regions (Wada et al. 1991), after normalizing by the expected values it becomes apparent that some selective constraint is operating at high G + C content against GG at II-III and against GC in introns. TC at II-III is negatively selected as well. These facts are adequately reflected when codon usage is analyzed in relation to variation in G + C content. Thus the use of G as ending base in arginine (quartet) and glycine codon group and the use of C as ending base in leucine (quartet) and valine codon groups are poorly correlated to G + C content while the respective use of C and G ending codons is highly correlated to the G + C content of genes (Marín et al. 1989).

Acknowledgments. We thank Josep Casadesús for helpful comments on the manuscript.

References

Aota SI, Ikemura T (1986) Diversity in G + C content at the third position of codons in vertebrate genes and its cause. Nucleic Acids Res 14:6345–6355

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

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

Bulmer M (1986) Neighboring base effects on substitution rates in pseudogenes. Mol Biol Evol 3:322–329

Dixon WJ, Brown MB (1979) BMDP-79 biomedical computer programmes P series. University of California Press, Berkeley

Hanai R, Wada A (1988) The effects of guanine and cytosine variation in 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

Marín A, Bertraneptit J, Oliver JL, Medina JR (1989) Variation in G + C-content and codon choice: differences among synonymous codon groups in vertebrate genes. Nucleic Acids Res 17:6181–6189

Nussinov R (1980) Some rules in the ordering of nucleotides in the DNA. Nucleic Acids Res 19:4545–4562

Nussinov R (1981a) Eukaryotic dinucleotide preference rules and their implications for degenerate codon usage. J Mol Biol 149:125–131

Nussinov R (1981b) The universal dinucleotide asymmetry rules in DNA and the aminoacid codon choice. J Mol Evol 17:237–244

Nussinov R (1984a) Strong doublet preferences in nucleotide sequence and DNA geometry. J Mol Evol 20:111–119

Nussinov R (1984b) 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

Stoehr PJ, Cameron GN (1991) The EMBL data library. Nucleic Acids Res 19 Supl:2227–2230

Wada K-N, Watanabe I, Tsuchiya R, Ikemura T (1991) G + C% mosaic structures of the higher vertebrate genome and distribution of dinucleotide frequencies. In: Kimura M, Nakahata T (eds) New aspects of the genetics of molecular evolution. Japan Sci Soc Press, Tokyo/Springer-Verlag, Berlin, pp 195–210