10

Isochores and synonymous substitutions in mammalian genes

GIORGIO BERNARDI, DOMINIQUE MOUCHIROUD, and CHRISTIAN GAUTIER

1. Introduction

The mammalian genome is a mosaic of isochores, long DNA segments (>300 kb on average) that are remarkably homogeneous in base composition and that can be subdivided into a small number of families characterized by different GC levels (1–4) (GC is the molar ratio of guanine + cytosine). In the human genome, which is representative of the majority of mammalian genomes (5, 6), isochores cover a broad GC range, 30–60% (4, 7). The third codon positions of human genes are compositionally correlated with the isochores in which the corresponding genes are located (1, 8), but they cover a much broader GC range, 25–97.5% (6).

The generation and maintenance of the large compositional heterogeneity of the isochores forming the mammalian genomes, and the genomes of warm-blooded vertebrates in general, have been the subject of contrasting views. In the traditional view of molecular evolution, the rate of point mutation is uniform over the genome of an organism and variation in the rate of nucleotide substitution among DNA regions reflects differential selective constraints (9–11).

In 1989, it was claimed (11) that the mutation rate at synonymous positions varied significantly among regions in the mammalian genome and was correlated with the base composition of genes and their flanking DNA. It was further proposed (11) that the differences arise because mutation patterns vary with the timing of replication of different chromosomal regions in the germline and that this hypothesis could account for both the origin of isochores in the mammalian genomes (1) and the observation (12) that synonymous nucleotide substitutions in different mammalian genes do not have the same molecular clock. More recently, it was claimed (13) that patterns of either damage or excision repair rates along the genome will produce patterns of mutation rates and that such patterns cause the DNA sequence patterns (i.e. the isochore patterns) which exist along the genome.

The specific point which will be discussed here, essentially following (14), is the evidence that synonymous substitution rates vary over different isochore families and over the genes contained in them. Obviously, this question is of crucial importance for the models (11, 13) just mentioned in connection with the generation and maintenance of isochores.

Recent developments in this area are described in refs 40–43.

2. Methods

The analyses concerning pairwise comparisons of homologous sequences carried out in (14) were performed on sequences retrieved from GenBank Release 76 (March 1993) using the database management system ACNUC (15). The procedure used to select the homologous pairs was described by Mouchiroud and Gautier (16). Comparisons comprised the mammalian genome pairs for which a large enough number of homologous sequences was available, namely, the human/other primates, human/artiodactyls, human/rabbit, mouse/rat, and human/rat pairs (6). The features exhibited by human/mouse homologous gene pairs were similar to those of the human/rat pairs.

To quantify dissimilarity between homologous sequences, the synonymous difference frequency (SDF) was used. SDF is the percentage divergence in third codon positions of synonymous codons. SDF does not rely on any hypothesis concerning the nature of the substitution process, in contrast with K_s, the substitution rate per synonymous site, as calculated according to Li *et al.* (17). However, when relatively small ranges are considered, K_s (or K_4, the substitution rate per fourfold degenerate site) and SDF show a strong linear correlation (14). This means that comparisons between our SDF data with K_s or K_4 data of other authors, are valid. In every case, the average SDF values were estimated not only for all points, but also for those characterized by low (<47%), intermediate (47–74%), and high (>74%) GC levels, in order to detect possible differences. These GC values roughly correspond to the borders between genes that are located in human isochore families L1, L2–H2, and H3, respectively (7).

3. Results

Plots of frequencies of substitution in third codon positions of synonymous codons (SDF) against GC levels of those positions show no significant correlation in the comparisons human/other primates, human/sheep, human/rabbit (14), rat/mouse (*Figure 1*), and human/rat (*Figure 2*). Such lack of correlation holds, therefore, both in the cases characterized by a conserved base compo-

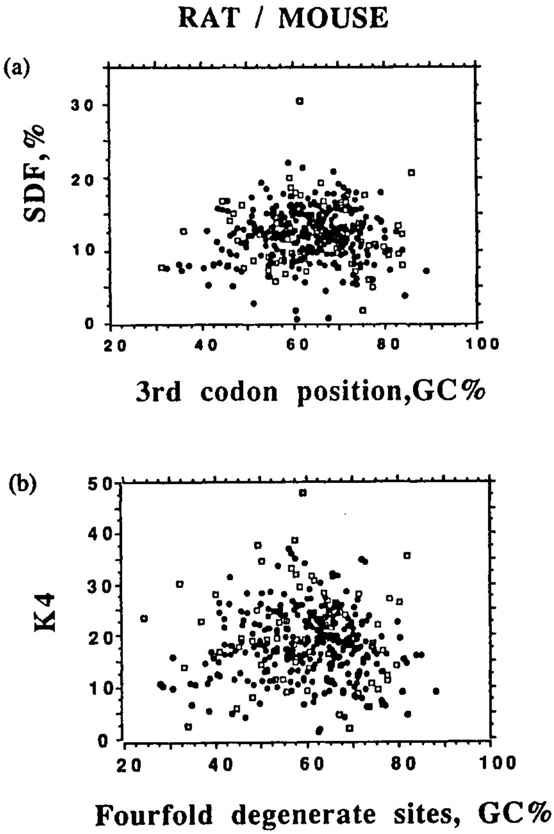

Figure 1. (a) SDF between homologous genes of mouse and rat is plotted against the average GC level of third codon positions. Filled circles correspond to coding sequences longer than 180 codons. For correlation coefficients and average SDF values for the L (< 47% GC), M (47–74% GC), and H (> 74% GC) sections see *Table 1.* (b) K4, the frequency of substitutions per fourfold degenerate sites, is plotted against the average GC level of those sites. Filled circles correspond to coding sequences longer than 180 codons.

sition in third codon positions (rat/mouse) and in those characterized by a different base composition (human/rat). A correlation which is weak, yet significant, was found in the comparisons man/calf (*Figure 3*) and human/pig (*Table 1*), in which SDF showed a slight increase with increasing GC. However, this increase appears tǒ be due to slightly lower SDF values for low GC third codon positions (*Table 1*). Since this phenomenon is not found in any other case, in particular in the two comparisons (mouse/rat and human/rat) involving the largest number of genes, an explanation which can be offered at present is that it is due to the particular, small sample of genes present in the low GC section of the plot. An alternative explanation, not exclusive of the

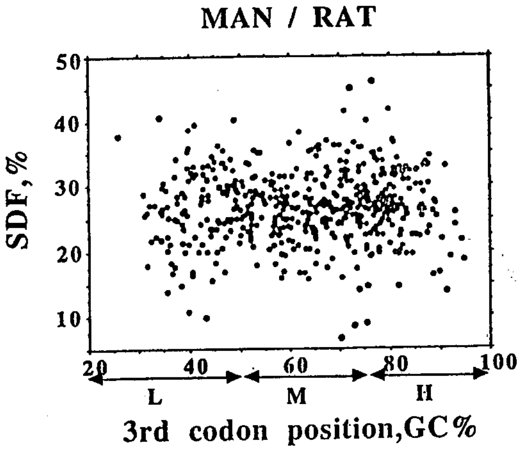

Figure 2. SDF between homologous genes of man and rat is plotted agains third codon position of human genes. Other indications as in *Figure 1* (14).

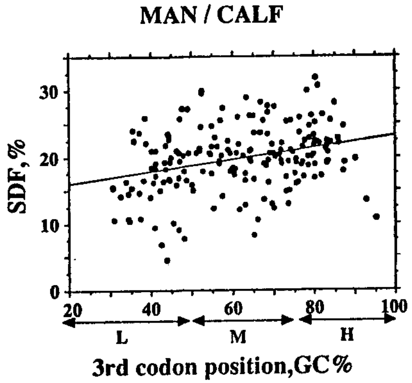

Figure 3. SDF between homologous genes of man and calf is plotted against GC of third codon positions of human genes. Other indications as in *Figure 1* (14).

former one, is that some changes become very frequent at the two ends of the spectrum of synonymous GC levels and they are underestimated. In the case of the man/calf comparison, A↔T and A→G are very frequent at the low GC end, G↔C are very frequent at the high GC end (*Table 2* and Figure 4). Under these circumstances, it is possible that back mutations become very frequent leading to an underestimate of the frequency of changes. A similar, yet weaker, phenomenon is probably responsible for the higher SDF value of the middle GC section of the human/rabbit comparison (*Table 1* and ref. 14).

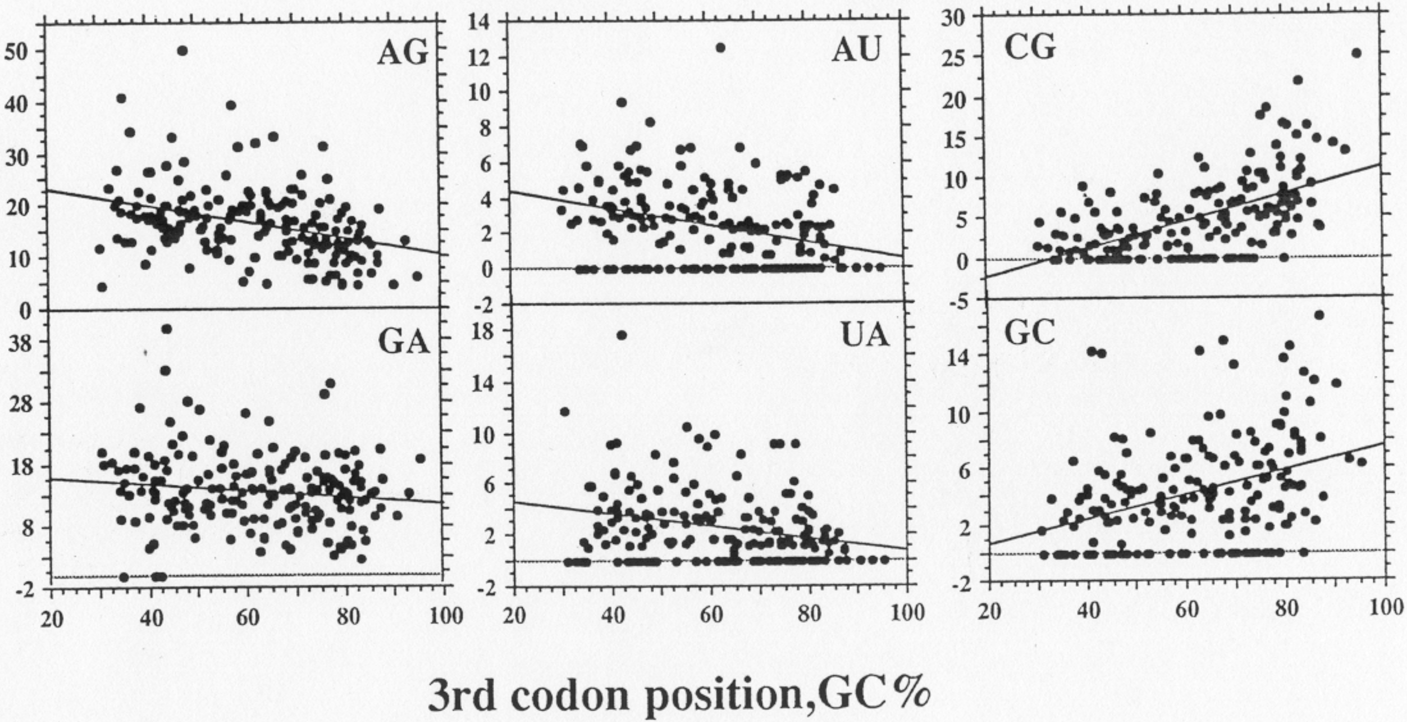

Figure 4. Individual changes between bovine and human genes (G → C, A → T, etc.) as percentages of all changes in synonymous third codon positions are plotted against GC levels of human third codon positions.

Table 1. Statistical analyses of synonymous difference frequencies in mammalian genes

Plots	Human/primate	Human/calf	Human/sheep	Human/pig	Human/rabbit	Rat/mouse	Human/rat
Number of genes	44	194	39	79	97	386	523
SDF *vs* GC_3 (human)[a]	$R = 0.07$	$R = 0.32$**[b]	$R = 0.16$	$R = 0.29$**	$R = 0.16$	$R = 0.045$	$R = 0.03$
Average SDF							
All sequences	5.2 (2.3)[c]	19.7 (4.9)	21.4 (4.8)	19.9 (4.3)	20.8 (5.3)	12.5 (3.7)	26.8 (5.4)
L sequences (< 47% GC)	4.2 (2.1)	17.5 (5.3)	19.7 (6.3)	16.5 (5.1)	18.2 (5.8)		26.7 (6.1)
M sequences (47–74% GC)	5.5 (2.5)	21.6 (4.5)	21.4 (4.8)	20.7 (3.6)	22.7 (4.6)		26.7 (5.3)
H sequences (> 74% GC)	5.2 (2.3)	21.3 (4.1)	22.6 (3.7)	20.6 (4.4)	17.8 (4.1)		27.1 (4.9)

[a] GC_3 is the third codon position GC of human sequences. In the mouse/rat comparison the average between the third codon position GC of the sequence pairs was used.

[b] Asterisks refer to statistical significancy (*, 5%; **, 1%).

[c] Values in parentheses are standard deviations.

Table 2. Frequencies of individual synonymous substitutions in homologous genes from human/calf and their correlations with GC levels of synonymous positions

Changes[a]	Human/calf	
	%	R
A→C	4.1	−0.19
A→G	**16.5**	**−0.37**
A→U	**2.4**	**−0.35**
C→A	3.2	+0.03
C→G	**4.8**	**+0.57**
C→U	19.2	+0.22
G→A	13.8	−0.14
G→C	**4.3**	**+0.39**
G→U	2.3	+0.10
U→A	**2.7**	**−0.12**
U→C	23.8	−0.12
U→G	2.9	+0.16

[a] Changes are indicated in the human → calf direction. Bold-type changes correspond to *p*-values equal to 10−4.

4. Discussion

This discussion will be divided in three parts. First, we discuss the previous reports in the light of our present findings. Then we analyse the interpretations claimed to account for previous reports. Finally, we discuss some general issues.

4.1 The frequencies of synonymous substitutions do not exhibit differences related to regions of the mammalian genome

The results of *Figures 1–3* unambiguously show that, although the frequencies of synonymous differences exhibit relatively large fluctuations in different genes (covering up to a 20-fold range in the mouse/rat case), they do not show any significant trend over the very extended range of synonymous position GC under consideration (explanations for the trend of *Figure 3* were provided in the Section 3). The above findings raise the question why discrepant results on the dependence of synonymous position GC level were previously reported.

1. The data by Miyata *et al.* (18), showing a lack of differences in synonymous substitution rates, are in apparent agreement with the results by Bernardi *et al.* (14). They fail, however, to prove the point made by the latter authors because of the small size of the samples studied. Indeed, the data only concern a total of 17 comparisons of genes from human, rat, rabbit, and monkey. Although some of the coding sequences investigated did show differences in synonymous position GC levels, the lack of differences in synonymous mutation rates may well be due to the small size of the sample studied,

the largest pairwise comparison, human/rat, comprising only nine genes. No definite conclusion can, therefore, be drawn from those data (and none was drawn by the authors). Indeed, if a systematic variation existed, it would have been missed. The same conclusion applies to the 12 comparisons of primate genes made by Filipski (19), in which no correlation was found, and also to the 13 human/Old World monkeys comparisons made by Wolfe *et al.* (11).

2. The second set of results concerns the rat/mouse comparison in which variations of mutation rates were reported, even if the results were different in two series of data. Indeed, in one case a strong increase with decreasing GC was found (19), whereas in another case a peak at 50% GC was reported (11). It is clear that these discrepant results were both due to the fact that the sizes of the gene samples used (30 genes in (19), 23 large genes in (11)) were still too small. Indeed, the variation in synonymous divergence being relatively large even for genes having exactly the same GC levels in synonymous codon positions (see *Figures 1–3*), any correlation may be found when using a small sequence sample.

3. Ticher and Graur (20) reported a correlation between silent substitutions rate and the percentage of different nucleotides at silent positions. This correlation was positive for A and T, negative for C, and non-significant for G. It concerned 42 homologous genes from human and rat having GC levels in third codon positions higher than 45%. This correlation could not be confirmed in the present work.

4. The last set of data (21) indicated no significant rate difference for 17 human/artiodactyl gene pairs that showed no GC differences in synonymous positions. This is, however, again a small sample from which no general rule can be drawn (see paragraph 1 above). Human or artiodactyl/rodent comparisons showed that some nine pairs of genes, with a small or no difference in silent position GC, exhibited lower rates than another nine or so pairs of genes, which showed large differences. The significance of these differences in rates is, however, doubtful in view of the present results on the human versus rat (or mouse) comparisons. Indeed, since in such a case differences in silent position GC exist for both GC-poor and GC-rich genes, the so-called 'minor shift' (5, 6, 16), one should notice higher numbers of substitutions for those 'extreme' genes compared with genes having a more balanced composition, which is not the case.

4.2 Differences in repair efficiency do not cause differences in the rates of synonymous substitutions of genes located in different isochore families

The higher rate of accumulation of mutations in GC-poor sequences in rodents compared with primates (and their compositional bias) was explained (19) as being due to less-efficient DNA repair in the GC-poor regions of the rodent genome. Obviously, the non-existence of such a higher rate does not question the existence of a less efficient repair of DNA lesions in rodent cells than in human cells (22), nor the evidence for between-gene differences in efficiency of DNA repair (23). In fact, the latter do exist among genes located in the same or in different isochore families. The lack of rate differences over genes located in different families of isochore indicates, therefore, that such less efficient or differential repair does not influence, on average, the rates of silent substitutions, even if between-gene differences do exist. This is an important conclusion, because DNA repair has been repeatedly considered to be a cause for differences in rates (and biases) in the mutation process (24–27). Indeed, this conclusion implies that there is no 'organization of mutation along the genome' which would be a 'prime determinant of genome evolution' as recently claimed (13) and that it is plainly wrong to consider (13) that R-bands and G-bands are characterized by a slow and by a fast molecular clock, respectively. Incidentally, these claims ignore the strong compositional heterogeneity of R-bands (see also Section 4.3) and the fact that repair efficiency is very different in transcribed versus non-transcribed sequences.

4.3 Differences in the process of mutation associated with replication timing do not affect the rates nor the biases of synonymous substitutions of genes located in different isochore families

The main conclusion drawn by Wolfe *et al.* (11), was 'that much of the intragenomic variation in silent substitution rate and base composition in mammals results from variation in the process of mutation, rather than from natural selection (28, 29). According to Wolfe *et al.* (11) 'the variation in both silent substitution rate and base composition is due to systemic differences in the rate and pattern of mutation over regions of the genome, the differences arising because mutation patterns vary with the timing of replication of different chromosomal regions in the genome', the proposal being that 'isochores arise as a result of the synchronous replication of megabase stretches of DNA under varying dNTP pool conditions'.

The fact that the conclusion of Wolfe *et al.* (11) 'that the substitution rate and the base composition of silent sites vary together in a systematic way' is wrong has two important consequences.

First, if changes in the nucleotide pools in the germline do exist (as assumed on the basis of what happens in somatic cells), the fact that mutation patterns do not vary with the timing of replication (because mutation rates are the same on average for genes which replicate early or late) means that changes in nucleotide pools do not cause biases in mutation patterns. In fact, it had already been pointed out that in the somatic tissues of mammals, late replicating DNA, like satellites and the inactive X chromosome, may be either GC-rich or GC-poor (30), and it has also been shown that early and late replicating genes may be either GC-poor or GC-rich (31). This finding can be understood (32–35) because of the abundance of GC-poor isochores in R-bands (which replicate early) and, to a much lesser extent, of GC-rich isochores in G-bands (which replicate late). Finally, early and late replication patterns also exist in cold-blooded vertebrates (2, and papers quoted therein), which never developed strong compositional differences in their genomes (36–38).

Secondly, if mutation patterns do not vary with the timing of replication of different chromosomal regions in the germline, the explanation of Wolfe *et al.* (11) for the origin of isochores in mammalian genomes no longer holds.

5. Conclusions

In conclusion, differences in average mutation rates (and in mutational biases) of synonymous codon positions of genes located in different isochore families of mammalian genomes have been claimed by several authors (11,19–21), but they could not be confirmed (14). The differences under consideration appear to be due to three reasons:

(a) The existence of relatively large individual fluctuations from gene to gene.

(b) The use of small, non-representative gene samples.

(c) The underestimate of A↔T and G↔C transversions which become quite frequent at the two ends of the compositional spectrum of synonymous positions.

A lack of correlation between synonymous site divergence and GC levels in the mouse/rat comparison (as well as between K_4 and GC_4 (the GC levels at fourfold degenerate sites)) was also independently reported by Wolfe and Sharp (39). However, these authors found a variation in K_4 (more specifically, a peak of K_4 values at 60% GC), but only when K_4 was averaged over genes within each 1% interval of GC_4. This effect might be simply due to sampling effects. Alternatively, it might be real and due to an underestimate of the rate associated with extreme compositions, as already discussed. In any case, this effect is admittedly small. In fact, small enough not to support anymore the claim 'that the substitution rate and the base composition of silent sites vary together in a systematic way' and the ensuing speculations on the maintenance and origin of isochores (11).

Under these circumstances, explanations other than differences in mutation rates and in mutational biases have to be taken into consideration in order to account for the generation and maintenance of the large differences in GC levels of third codon positions of genes located in different isochore families from mammalian genomes (3, 28, 30).

References

1. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., *et al.* (1985). *Science,* **228**, 953.
2. Bernardi, G. (1989). *Annu. Rev. Genet.,* **23**, 637.
3. Bernardi, G. (1993). *Mol. Biol.Evol.,* **10**, 186.
4. Bernardi, G. (1993). *Gene,* **135**, 57.
5. Sabeur, G., Macaya, G., Kadi, F., and Bernardi, G. (1993). *J. Mol. Evol.,* **37**, 93.
6. Mouchiroud, D. and Bernardi, G. (1993). *J. Mol. Evol.,* **37**, 109.
7. Mouchiroud, D., D'Onofrio, G., Aïssani, B., Macaya, G., Gautier, C., and Bernardi, G. (1991). *Gene,* **100**, 181.
8. Aïssani, B., D'Onofrio, G., Mouchiroud, D., Gardiner, K., Gautier, C., and Bernardi, G. (1991). *J. Mol. Evol.,* **32**, 493.
9. Kimura, M. (1985). *The neutral theory of evolution.* Cambridge University Press.
10. Sharp, P. M. and Li, W.-H. (1987). *Mol. Biol. Evol.,* **4**, 222.
11. Wolfe, K. H., Sharp, P. M., and Li, W.-H. (1989). *Nature,* **337**, 283.
12. Li W.-H., Tanimura, M., and Sharp, P. M. (1987). *J. Mol. Evol.,* **25**, 330.
13. Holmquist, G. P. and Filipski, J. (1994). *Trends Ecol. Evol.,* **9**, 65.
14. Bernardi, G., Mouchiroud, D., and Gautier, C., (1993). *J. Mol. Evol.,* **37**, 583.
15. Gouy, M., Gautier, C., Attimonelli, M., Lanave, C., and di Paola, G. (1985). *Comput. Appl. Biosci.,* **1**, 167.
16. Mouchiroud, J. D., and Gautier, C. (1990). *Mol. Evol.,* **31**, 81.
17. Li, W. -H., Wu, C. I., and Luo, C. C. (1985). *Mol. Biol. Evol.,* **2**, 150.
18. Miyata, T., Hayashida, H., Kikuno, R., Hasegawa, M., Kobayashi, M., and Koike, K. (1982). *J. Mol. Evol.,* **19**, 28.
19. Filipski, J. (1988). *J. Theor. Biol.,* **134**, 159.
20. Ticher, A. and Graur, D. (1989). *J. Mol. Evol.,* **28**, 286.
21. Saccone, C., Pesole, G., and Preparata, G. (1989). *J. Mol. Evol.,* **29**, 407.
22. Hart, R. W. and Setlow, R. B. (1974). *Science,* **71**, 2169.
23. Bohr, V. A., Philips, D. H., and Hanawalt, P. C. (1987). *Cancer Res.,* **47**, 6426.
24. Filipski, J. (1987). *FEBS Lett.,* **217**, 184.
25. Sueoka, N. (1988). *Proc. Natl. Acad. Sci. USA,* **85**, 2653.
26. Sueoka, N. (1992). *J. Mol. Evol.,* **34**, 95.
27. Sueoka, N. (1993). *J. Mol. Evol.,* **37**, 137.
28. Bernardi, G. and Bernardi, G. (1986). *J. Mol. Evol.,* **24**, 1.
29. Gillespie, J. H. (1986). *Genetics,* **113**, 1077.
30. Bernardi, G., Mouchiroud, D., Gautier, C., and Bernardi, G. (1988). *J. Mol. Evol.,* **28**, 7.
31. Eyre-Walker, A. (1992). *Nucleic Acids Res.,* **20**, 1497.
32. Gardiner, K., Aïssani, B., and Bernardi, G. (1990). *EMBO J.,* **9**, 1853.
33. Pilia, G., Little, R. D., Aïssani, B., Bernardi, G., and Schlessinger, D. (1993). *Genomics,* **17**, 456.
34. Saccone, S., De Sario, A., Della Valle, G., and Bernardi, G. (1992). *Proc. Natl. Acad. Sci. USA,* **89**, 4913.
35. Saccone, S., De Sario, A., Wiegant, J., Raap, A. K., Della Valle, G., and Bernardi, G. (1993). *Proc. Natl. Acad. Sci. USA,* **90**, 11929.
36. Bernardi, G. and Bernardi, G. (1990). *J. Mol. Evol.,* **31**, 265.
37. Bernardi, G. and Bernardi, G. (1990). *J. Mol. Evol.,* **31**, 282.

38. Bernardi, G. and Bernardi, G. (1991). *J. Mol. Evol.,* **33**, 57.
39. Wolfe, K. H. and Sharp, P. M. (1993). *J. Mol. Evol.,* **37**, 441.
40. Mouchiroud, D., Gautier, C., and Bernardi, G. (1995). *J. Mol. Evol.*, **40**, 107.
41. Cacciò, S., Zoubak, S., D'Onofrio, G., and Bernardi, G. (1995). *J. Mol. Evol.*, **40**, 280.
42. Zoubak, S., D'Onofrio, G., Cacciò, S., Bernardi, G., and Bernardi, G. (1995). *J. Mol. Evol.*, **40**, 293.
43. Bernardi, G. (1995). *Annu. Rev. Genet.*, **29**, 445.