J Mol Evol (1995) 40:293–307

JOURNAL OF MOLECULAR EVOLUTION

Specific Compositional Patterns of Synonymous Positions in Homologous Mammalian Genes†

Serguei Zoubak,[1]* Giuseppe D'Onofrio,[1] Simone Cacciò,[1] Giacomo Bernardi,[2] Giorgio Bernardi[1]**

[1] Laboratoire de Genetique Moleculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France
[2] University of California, Santa Cruz, Department of Biology, Santa Cruz, CA 95064, USA

Received: 15 July 1994

Abstract. All 69 homologous coding sequences that are currently available in four mammalian orders were aligned and the synonymous (*ie.,* third) positions of quartet (fourfold degenerate) codons were divided into three classes (that will be called conserved, intermediate, and variable), according to whether they show no change, one change, and more than one change, respectively. The three classes were analyzed in their compositional patterns. In the majority of GC-rich genes, the three classes of positions (but especially conserved positions) exhibited significantly different base compositions compared to expectations based on a ''random'' substitution process from the ''ancestral'' (consensus) sequence to the present-day (actual) sequences. Significant differences were rare in GC-poor genes.

An analysis of the present results indicates that natural selection plays a role in the synonymous nucleotide substitution process, especially in GC-rich genes which represent the vast majority of mammalian genes.

Key words: Homologous mammalian genes — Compositional patterns — Synonymous positions

Introduction

The recent discovery that synonymous substitution frequencies in mammals are gene-specific and correlated in frequency with nonsynonymous substitutions (Mouchiroud et al. 1995) prompted us to analyze the synonymous positions of quartet and duet (fourfold and twofold degenerate) codons in homologous genes from four mammalian orders with the aim of finding whether substitution frequencies were those expected from a random process or exhibited specific distribution patterns. Such an analysis (Cacciò et al. 1995) concerned synonymous positions as divided into three classes of conserved, intermediate, and variable positions (showing no change, one change, and more than one change, respectively) and revealed that the frequencies of the three classes (1) were different in different genes for both duet and quartet codons; (2) were significantly different from expectations based on a ''random'' synonymous substitution process in the majority of the genes for quartet codons and in a minority of genes for duet codons; (3) were correlated between duet and quartet codons. Moreover, the frequencies of the conserved positions in both duet and quartet codons were correlated with amino acid conservation. These results indicated a parallelism of substitutions in duet and quartet codons, as well as a gene specificity and a nonrandomness of the substitution process (Cacciò et al. 1995).

Here, we have extended the analysis of Cacciò et al. (1995) by investigating the base compositions of the synonymous positions belonging to the three classes. The rationale for such an approach was to develop evidence for (or against) negative selection in synonymous posi-

† This paper is dedicated to the memory of Dr. Motoo Kimura (1924–1994).

* *Permanent address:* Institute of Molecular Biology and Genetics, Zabolotnogo str. 150, 252627 Kiev, Ukraine

** *Permanent address:* Stazione Zoologica, Villa Comunale 1, 80121 Naples, Italy

Correspondence to: Giorgio Bernardi

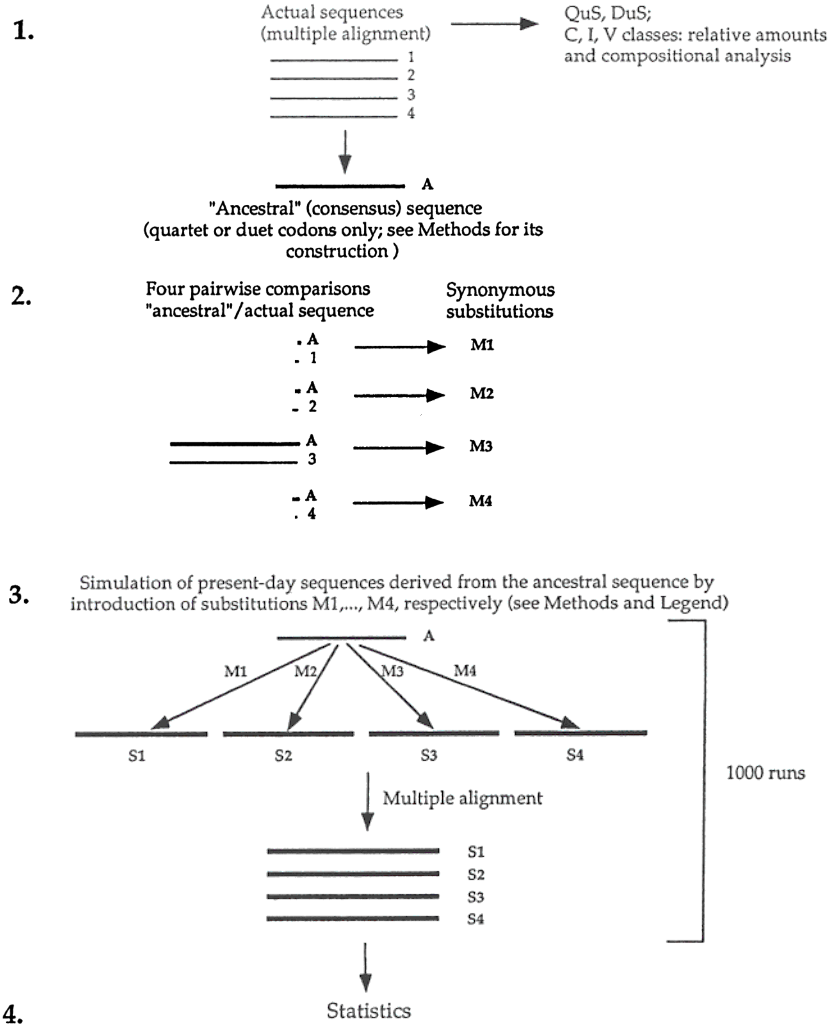
1.
Actual sequences
(multiple alignment)
1
2
3
4
QuS, DuS;
C, I, V classes: relative amounts
and compositional analysis
A
"Ancestral" (consensus) sequence
(quartet or duet codons only; see Methods for its construction)
2.
Four pairwise comparisons
"ancestral"/actual sequence
Synonymous
substitutions
A
1
M1
A
2
M2
A
3
M3
A
4
M4
3.
Simulation of present-day sequences derived from the ancestral sequence by introduction of substitutions M1,..., M4, respectively (see Methods and Legend)
A
M1
M2
M3
M4
S1
S2
S3
S4
1000 runs
Multiple alignment
S1
S2
S3
S4
4.
Statistics

Fig. 1. Scheme of the approach used in order to compare frequencies and base compositions of quartet codons of actual sequences with those of sequences generated by inserting random substitutions in a reconstructed ''ancestral'' (consensus) sequence. *QuS* and *DuS* stand for synonymous quartets and duets, *C, I,* and *V* for conserved, intermediate, and variable classes, respectively. In the study of frequency patterns (Cacciò et al. 1995), *M1* to *M4* are the numbers of synonymous substitutions found in step 2. In the study of compositional patterns (present work), *M1* to *M4* are the nucleotides that replaced in the actual sequences those present in the ''ancestral'' one. (See also Methods.)

tions along the line that led to the general acceptance of negative selection in nonsynonymous positions (see Li and Graur 1991), namely, the demonstration of the specific nature of amino acid changes, e.g., the lack of substitutions in functionally crucial amino acids and the conservative nature of replacements.

Methods

The frequencies and the compositions of each class of synonymous positions found in the actual coding sequences studied were compared with expectations based on a random substitution process taking place between the ''ancestral'' and the present-day actual sequence as follows (Fig. 1).

1. Construction of the ''Ancestral'' (Consensus) Sequence. First of all, the four homologous sequences of each gene from four mammalian orders were aligned in order to determine the number of synonymous quartets and duets, as well as the relative amounts and the base compositions of the conserved, intermediate, and variable classes of positions. (See Introduction and Cacciò et al. [1995] for a definition of these classes.) Then an ''ancestral'' sequence was constructed as the *consensus* of all synonymous quartet or duet codons. Conserved and intermediate positions obviously raised no problem, the positions being occupied by the same nucleotide in all four sequences or in three sequences out of four, respectively. In quartet codons, variable positions could be of three types, WWXY, WWXX, or WXYZ. In the first case, the choice for the consensus sequence was again obvious. In the second and third cases, which concern a minority of variable positions (themselves a minority of all synonymous quartet codons), using a random choice would lead to compositions that would deviate from those present in the actual sequence. The procedure used in order to avoid these problems was to approach the nucleotide frequencies as present in the actual sequences. (50:50 in the case WWXX, 25:25:25:25 in the case WXYZ).

2. A Pairwise Comparison of the ''Ancestral'' (Consensus) Sequence with Each One of the Actual Homologous Sequences. This comparison, also only comprising synonymous quartet or duet codons, was then performed in order to detect the nucleotide substitutions (M1 . . . M4) in third codon positions of the actual sequences compared to the ''ancestral'' one. The average base compositions of all third positions of synonymous quartet codons of the actual sequences were compared with those of the ''ancestral'' sequence. As shown by the difference histograms (Fig. 2A), in GC-rich genes, G and C were less abundant, and A and T were more abundant, in the actual sequence compared to the ''ancestral'' sequence. GC-poor genes tended to show a reverse pattern compared to GC-rich genes. Differences were, however, small (up to ±4–6%). The strongest differences and the deviations

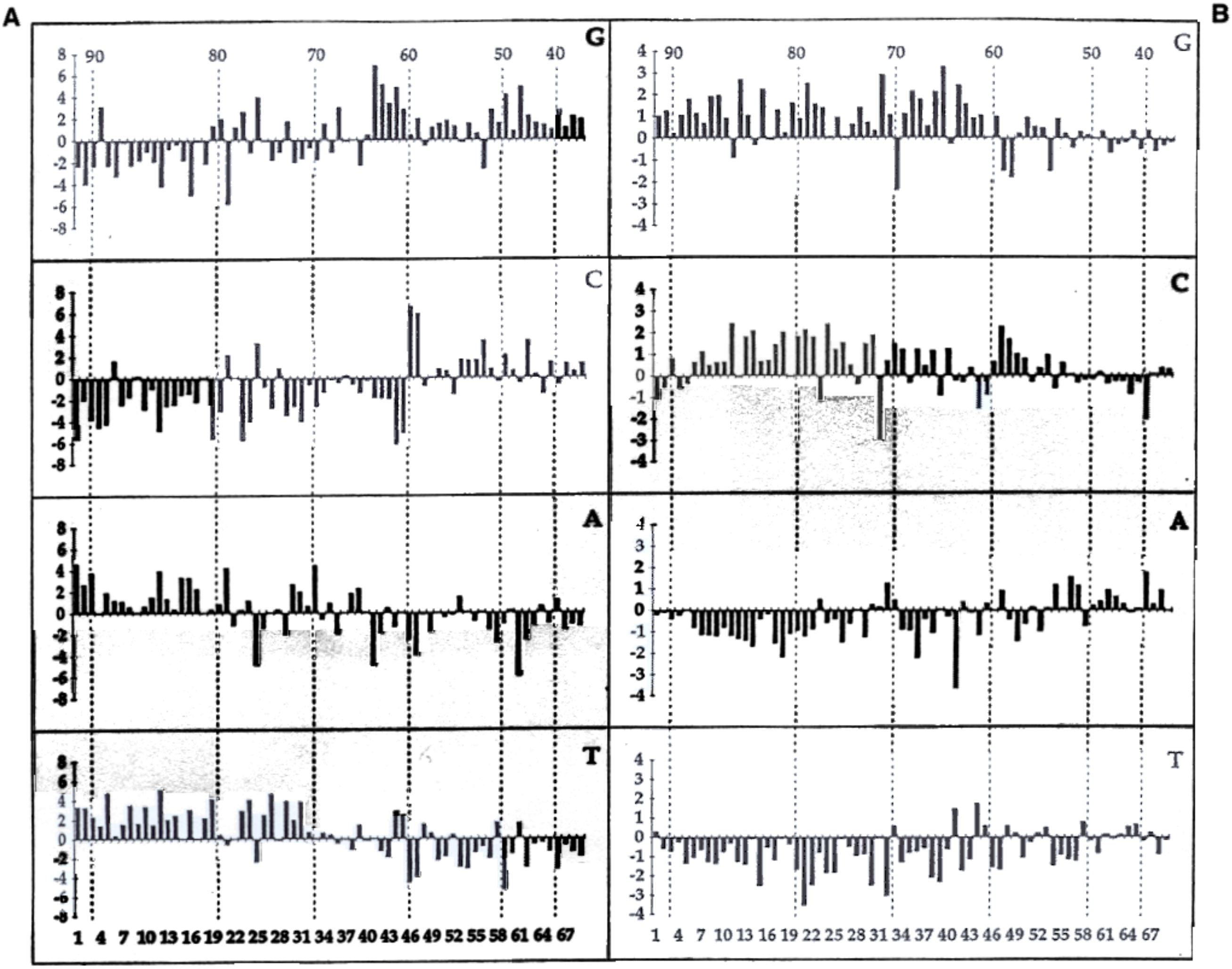

Fig. 2. Difference histograms of base composition of synonymous positions of all quartet codons in the actual sequences (average) and (**A; left**) in the "ancestral" sequence, or (**B; right**) in the (average) simulated sequence. The figures on the horizontal top line (*90 to 40*) correspond to GC_3 of the genes (dashed vertical lines); the figures on the horizontal bottom line (*1 to 67*) correspond to gene numbers (Table 1).

from the trends described largely disappeared when eliminating sequences comprising less than 50 quartet codons (not shown). Since the trends for G and C, and for A and T, respectively, were the same, a difference histogram for G+C was also constructed (Fig. 3A). As expected, this eliminated most of the "anomalies" present in the histogram for individual bases. The differences appearing in Figs. 2A and 3A can be understood because the majority rule was used in constructing the "ancestral" sequence from four sequences only.

3. Simulation of the Present-day (Actual) Sequences as Obtained by a Randomization of the "Ancestral" (Consensus) Sequence. In the case of quartet codons, the nucleotides (A, G, C, T) that were found to be substituted in the actual sequences compared to the "ancestral" sequence as judged from pairwise comparisons were pooled and redistributed at random in the third positions of quartet codons of the "ancestral" sequence. This generated four *simulated* sequences which were then compared in multi-alignments. This process was repeated 1,000 times. Substitutions of one given nucleotide by the same nucleotide were not allowed in order to keep the overall number of substitutions in the stimulated sequence equal to that in the actual sequence. In the case of duet codons, obviously, only G↔A and T↔C changes were allowed. Moreover, the situation in which a conserved third codon position was the result of a mutation followed by a reversion was neglected.

This procedure assumes a star phylogeny and an absence of common branches which are justified by the work of Bulmer et al. (1991). Different rates in the four lineages are at least largely taken into account by using the synonymous substitutions as found in pairwise comparison with the "ancestral" sequence and applying such substitutions in the simulation. The slight differences of Figs. 2B and 3B originate from the fact that the "ancestral" sequence from which the simulated sequences derived was a consensus sequence slightly enriched in G and C for GC-rich genes. (See also preceding section.)

The average base compositions of all third positions of synonymous quartet codons of the actual sequences were then compared with those of the simulated sequences obtained by randomizing the "ancestral" consensus sequence. Differences were very small (only up to ±2–3%) (Fig. 2B; note the different scale compared to Fig. 2A) and showed that, in GC-rich genes, G and C were slightly more abundant, and A and T slightly less abundant, in the actual sequences than in the average simulated sequences, as expected. In GC-poor genes, differences were essentially negligible. Again the strongest differences and the deviations from the trends described were due to sequences comprising less than 50 quartet codons. Figure 3B concerns G+C values and shows the differences which were expected from the data of Fig. 2B.

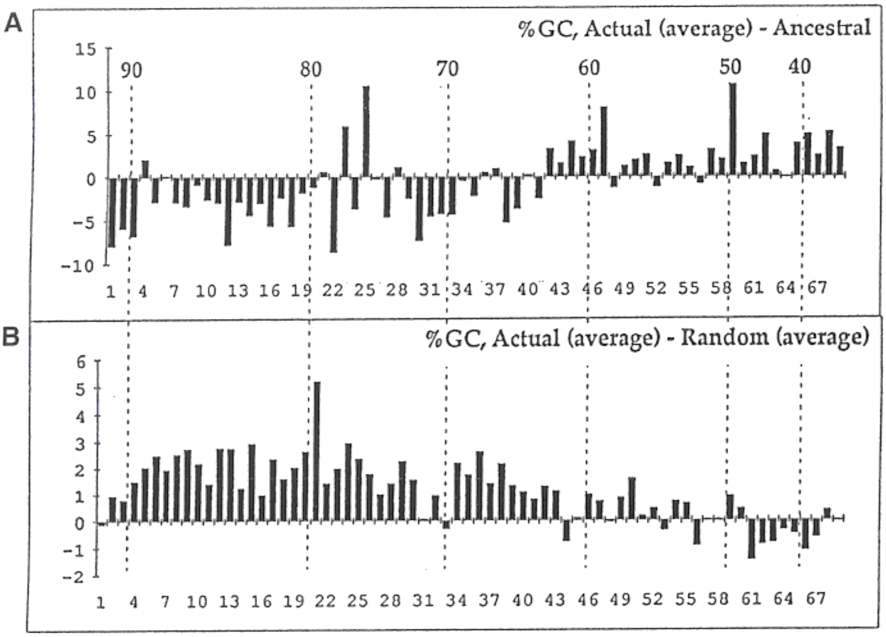

Fig. 3. Difference histograms of GC levels of synonymous positions of all quartet codons in the actual sequence (average) and **(A; top)** in the ''ancestral'' sequence, or **(B; bottom)** in the average simulated sequence. For other indications, see the legend of Fig. 2.

The smallness of differences of Figs. 2B and 3B indicates that the overall synonymous substitution process is well mimicked by the simulation procedure used. This involves random substitutions of third codon positions of quartet codons of the ''ancestral'' sequence using the pool of nucleotides found to be substituted in the actual compared to the (reconstructed) ''ancestral'' sequence, as defined in pairwise comparison. The compositional differences were so small and the overall process seems to be caused by a random process because the nucleotide substitutions reintroduced in the ''ancestral'' sequence are those which were actually found in the present-day sequences.

4. Significance of the Frequency Patterns and of Compositional Patterns. Histograms of the frequencies or of the base composition of conserved, intermediate, and variable positions, as found in the 1,000 alignments of the four simulated sequences, were constructed and compared with the results from the alignments of the four actual sequences for each gene. A χ^2 test was then carried out in order to assess the significance of the difference between the values derived from the multi-alignment of the actual sequences and the average value of each class of positions as detected in the histogram of the simulated sequences. χ^2 tests with two degrees of freedom were also performed for the three sets of frequencies of conserved, intermediate, and variable positions combined together and the significance of differences (*P*) was estimated (This approach was used in Cacciò et al. [1995].)

As far as base composition is concerned, the treatment described above has the problem that very low numerical values were often encountered when dealing with each individual base from positions of a given class. Since G and C, and A and T, respectively, showed the same trends, G+C values could be used more satisfactorily. χ^2 tests with two degrees of freedom were also performed for the G+C values in the three classes and the significance of deviation (*P*) was estimated. (This approach was used in the present work.)

5. An analysis of the codons following the synonymous positions. In order to study the nucleotide context of the synonymous positions of quartet codons, the first and second positions of the codons following the synonymous positions were analyzed (the two positions preceding them were obviously identical). The compositions of first and second positions following synonymous quartet codons were determined. Moreover, the percentages of conserved, intermediate, and variable first positions were assessed.

For other points, the reader is referred to the Methods section of the preceding paper (Cacciò et al. 1995).

Results

Conserved, Intermediate, and Variable Synonymous Positions of Homologous Genes (Especially of GC-Rich Genes) Are Characterized by Compositional Features That Are Generally Different from Expectations Based on a Random Substitution Process

As indicated in the Methods section, synonymous positions of quartet codons of homologous genes derived from four mammalian orders show compositional features that deviate only slightly from the ''ancestral'' (consensus) sequence and barely from random expectations. In contrast, when the molar ratios of nucleotides in the conserved and variable classes of positions from actual sequences were compared with those present in the corresponding classes of simulated present-day sequences (Table 1 and Fig. 4), very large differences (up to ±10–20%) were found in GC-rich genes (G and C values being higher in the conserved class and lower in the variable class). These trends diminished or reversed for GC-poor genes. The intermediate class resembled the variable class (not shown). Deviations from the trends just described were largely eliminated when neglecting sequences comprising less than 50 quartet codons (not shown). Figure 5 shows difference histograms for the

Table 1. Base composition of synonymous positions of fourfold degenerate codons in homologous genes from four mammalian orders[a]

		G	C	A	T
1 Apo E	T	40.3	47.7	7.3	4.7
	C	42.9	51.4	2.9	2.9
	I	40.6	44.5	8.6	6.2
	Ic	43.8	56.2	0.0	0.0
	V	28.1	43.8	21.9	6.2
2 Creatine kinase B	T	38.4	46.0	6.1	9.5
	C	44.0	49.0	3.0	4.0
	I	32.4	45.4	7.4	14.8
	Ic	38.9	51.9	3.7	5.6
	V	28.3	34.8	16.3	20.7
3 A1 adenosine	T	34.3	51.5	5.8	8.3
	C	33.7	61.6	2.3	2.3
	I	37.5	38.1	8.3	16.1
	Ic	42.9	40.5	4.8	11.9
	V	30.7	37.5	14.8	17.0
4 H,K ATPase (β)	T	36.1	44.3	12.5	7.1
	C	45.5	45.5	6.1	3.0
	I	32.1	49.1	13.4	5.4
	Ic	32.1	46.4	17.9	3.6
	V	28.7	38.0	19.4	13.9
5 α-globin	T	37.3	46.7	1.9	14.2
	C	42.9	50.0	0.0	7.1
	I	31.2	45.3	1.6	21.9
	Ic	37.5	50.0	0.0	12.5
	V	30.6	38.9	8.3	22.2
6 ApoA1	T	47.5	37.7	6.1	8.6
	C	55.6	41.7	0.0	2.8
	I	37.5	34.4	14.1	14.1
	Ic	12.5	56.2	18.8	12.5
	V	33.3	27.8	16.7	22.2
7 Na/H exchange protein	T	37.5	47.0	8.3	7.2
	C	42.3	52.6	2.8	2.3
	I	30.8	45.9	10.3	13.0
	Ic	28.8	46.2	8.7	16.3
	V	31.4	27.3	25.9	15.5
8 Serine-pyruvate aminotransferase	T	38.8	40.1	9.8	11.3
	C	47.5	49.2	1.6	1.6
	I	37.2	32.4	14.9	15.4
	Ic	38.3	40.4	14.9	6.4
	V	25.0	34.1	17.4	23.5
9 Dipeptidase	T	38.9	42.4	7.4	11.3
	C	46.8	50.0	1.6	1.6
	I	34.8	40.2	7.1	17.9
	Ic	41.3	43.5	2.2	13.0
	V	27.8	28.7	21.3	22.2
10 GMP-phosphodiesterase α	T	37.3	40.9	10.4	11.4
	C	44.9	47.1	4.4	3.7
	I	32.0	39.5	12.2	16.3
	Ic	43.0	33.7	10.5	12.8
	V	29.7	31.3	19.3	19.7
11 CDS α chain	T	39.2	40.2	9.3	11.3
	C	40.7	51.9	0.0	7.4
	I	37.5	30.0	20.0	12.5
	Ic	50.0	40.0	10.0	0.0
	V	37.5	25.0	19.6	17.9
12 Glutathione peroxidase	T	42.1	37.8	6.4	13.7
	C	56.8	37.8	0.0	5.4
	I	33.7	42.3	6.7	17.3
	Ic	34.6	42.3	3.8	19.2
	V	25.0	31.6	18.4	25.0
13 Retinol-binding protein	T	27.5	45.1	13.5	13.9
	C	25.0	59.4	6.2	9.4
	I	31.6	27.6	23.7	17.1
	Ic	31.6	26.3	26.3	15.8
	V	27.5	32.5	17.5	22.5
14 H,K ATPase α subunit	T	34.8	42.5	11.2	11.5
	C	40.9	53.5	2.2	3.5
	I	29.4	37.8	16.1	16.6
	Ic	27.2	41.1	16.5	15.2
	V	29.8	25.9	23.2	21.1
15 Glucose Glut3	T	33.7	43.9	9.7	12.7
	C	41.5	53.7	2.4	2.4
		30.4	34.7	16.7	18.3
	Ic	33.3	39.8	11.8	15.1
	V	16.2	34.5	16.2	33.1
16 Prostaglandin E receptor	T	41.4	40.0	6.2	12.3
	C	45.5	45.5	1.5	7.6
	I	38.2	36.8	11.3	13.7
	Ic	39.6	43.4	5.7	11.3
	V	36.9	31.0	8.3	23.8
17 Prolyl-4 hydroxylase β	T	35.6	36.9	13.5	14.0
	C	48.4	43.8	4.7	3.1
	I	27.7	39.5	15.2	17.6
	Ic	25.0	45.3	10.9	18.8
	V	29.4	25.0	22.5	23.0
18 Growth hormone	T	38.2	45.6	7.0	9.2
	C	36.4	57.6	0.0	6.1
	I	48.1	26.9	5.8	19.2
	Ic	46.2	38.5	7.7	7.7
	V	31.8	31.8	29.5	6.8
19 TNFα	T	24.1	50.0	13.0	13.0
	C	25.0	62.5	3.1	9.4
	I	26.6	45.2	17.7	10.5
	Ic	29.0	58.1	12.9	0.0
	V	17.2	34.4	23.4	25.0
20 Ferritin L	T	22.6	55.6	4.0	17.9
	C	27.0	64.9	0.0	8.1
	I	10.0	53.3	1.7	35.0
	Ic	0.0	46.7	0.0	53.3
	V	25.0	27.3	20.5	27.3
21 Myoglobin	T	38.8	38.3	10.6	12.2
	C	50.0	50.0	0.0	0.0
	I	43.8	33.3	18.8	4.2
	Ic	58.3	41.7	0.0	0.0
	V	20.0	26.7	18.3	35.0
22 Apo CIII	T	26.2	55.0	3.8	15.0
	C	25.0	75.0	0.0	0.0
	I	25.0	50.0	3.6	21.4
	Ic	14.3	71.4	0.0	14.3
	V	30.0	30.0	10.0	30.0
23 TNF β	T	28.6	40.9	13.3	17.2
	C	36.6	41.5	12.2	9.8
	I	19.3	42.0	15.9	22.7
	Ic	9.1	45.5	18.2	27.3
	V	19.6	37.5	12.5	30.4
24 Hydrophobic-surfactant-associated factor	T	27.2	43.7	13.1	16.0
	C	28.6	53.6	7.1	10.7
	I	27.2	45.7	14.1	13.0
	Ic	34.8	52.2	4.3	8.7
	V	25.0	23.4	21.9	29.7

Table 1 Continued

		G	C	A	T
25 Phospholipase A2	T	16.9	51.6	17.7	13.7
	C	16.7	66.7	8.3	8.3
	I	20.0	52.5	25.0	2.5
	Ic	30.0	40.0	30.0	0.0
	V	13.9	30.6	22.2	33.3
26 Phenyl-tRNA ligase	T	29.6	44.4	12.7	13.3
	C	30.3	57.6	3.0	9.1
	I	26.2	39.5	18.0	16.3
	Ic	20.9	37.2	23.3	18.6
	V	32.1	27.6	23.1	17.3
27 Tissue inhibitor of metalloproteinase	T	20.8	48.1	15.1	16.0
	C	17.4	60.9	8.7	13.0
	I	27.3	39.8	18.2	14.8
	Ic	31.8	45.5	18.2	4.5
	V	12.5	34.4	25.0	28.1
28 Guanine nt-binding protein	T	32.2	37.4	7.9	22.5
	C	36.2	39.7	6.9	17.2
	I	21.6	29.3	7.8	41.4
	Ic	27.6	24.1	6.9	41.4
	V	21.4	35.7	16.1	26.8
29 Polymeric Ig receptor	T	29.4	43.3	14.2	13.1
	C	34.4	51.6	7.8	6.2
	I	26.0	42.2	15.7	16.2
	Ic	31.4	45.1	13.7	9.8
	V	24.0	25.0	26.9	24.0
30 Colony-stimulating factor	T	16.9	54.1	16.2	12.8
	C	11.1	72.2	16.7	0.0
	I	28.3	36.7	13.3	21.7
	Ic	33.3	46.7	0.0	20.0
	V	0.0	37.5	25.0	37.5
31 CD4 antigen	T	41.3	30.8	11.3	16.6
	C	57.1	23.8	9.5	9.5
	I	27.0	39.0	9.0	25.0
	Ic	28.0	44.0	8.0	20.0
	V	25.0	35.5	18.4	21.1
32 Erythropoietin	T	26.4	37.9	20.7	15.0
	C	32.3	41.9	22.6	3.2
	I	23.4	33.9	21.0	21.8
	Ic	32.3	38.7	12.9	16.1
	V	15.6	37.5	12.5	34.4
33 Gastrin	T	23.2	29.5	29.5	17.9
	C	8.3	41.7	33.3	16.7
	I	32.1	19.6	26.8	21.4
	Ic	28.6	28.6	21.4	21.4
	V	50.0	25.0	25.0	0.0
34 Na^+nucleoside	T	28.4	37.6	19.1	15.0
	C	32.3	49.5	12.9	5.4
	I	25.9	31.1	21.3	21.6
	Ic	23.2	30.5	20.7	25.6
	V	25.0	25.5	27.1	22.4
35 D-amino-acid oxidase	T	29.0	33.9	20.8	16.3
	C	37.3	39.0	15.3	8.5
	I	25.0	27.2	24.5	23.4
	Ic	23.9	26.1	26.1	23.9
	V	14.3	34.5	28.6	22.6
36 Ferritin H	T	35.0	31.5	16.0	17.5
	C	48.1	37.0	3.7	11.1
	I	19.4	25.0	29.2	26.4
	Ic	11.1	22.2	38.9	27.8
	V	20.0	25.0	35.0	20.0
37 Protein kinase C (β type)	T	28.5	38.4	16.0	17.1
	C	29.0	45.8	13.0	12.2
	Ic	29.5	31.0	17.9	21.7
	Ic	23.8	33.3	19.0	23.8
	V	25.5	30.9	21.3	22.3
38 ANP	T	28.6	37.5	16.1	17.9
	C	33.3	45.8	8.3	12.5
		31.2	21.9	28.1	18.8
	Ic	37.5	25.0	25.0	12.5
	V	15.0	30.0	25.0	30.0
39 β-globin	T	34.7	37.5	2.3	25.5
	C	44.4	36.1	0.0	19.4
	I	15.9	40.9	6.8	36.4
	Ic	27.3	45.5	0.0	27.3
	V	14.3	39.3	7.1	39.3
40 Potassium channel	T	23.4	38.8	19.8	18.0
	C	22.2	47.2	17.6	13.0
	I	24.2	29.9	20.1	25.8
	Ic	22.7	27.3	22.7	27.3
	V	27.8	20.8	31.9	19.4
41 Cytochrome b5	T	21.9	33.1	17.5	27.5
	C	27.3	36.4	0.0	36.4
	I	21.2	32.7	13.5	32.7
	Ic	23.1	23.1	15.4	38.5
	V	18.8	31.2	32.8	17.2
42 Endothelin	T	20.5	36.5	23.7	19.2
	C	27.8	38.9	22.2	11.1
	I	10.4	39.6	29.2	20.8
	Ic	0.0	41.7	33.3	25.0
	V	19.4	27.8	19.4	33.3
43 Phagocytic glycoprotein-1	T	22.4	35.1	22.0	20.5
	C	25.6	38.5	23.1	12.8
		20.2	38.3	17.0	24.5
	Ic	14.9	36.2	19.1	29.8
	V	21.7	25.8	28.3	24.2
44 Prolactin	T	18.8	41.0	18.1	22.2
	C	18.2	54.5	0.0	27.3
		15.4	34.6	25.0	25.0
	Ic	15.4	30.8	30.8	23.1
	V	22.9	35.4	27.1	14.6
45 Interleukin 2-receptor	T	21.5	32.0	23.3	23.3
	C	21.1	36.8	15.8	26.3
		19.2	25.0	40.4	15.4
	Ic	15.4	23.1	53.8	7.7
	V	25.0	31.8	15.9	27.3
46 Tissue factor	T	25.0	25.0	32.1	17.9
	C	38.9	27.8	27.8	5.6
	I	6.7	21.7	43.3	28.3
	Ic	0.0	13.3	53.3	33.3
	V	26.6	25.0	26.6	21.9
47 β2 microglobulin	T	26.0	22.0	28.0	24.0
	C	11.1	33.3	33.3	22.2
	I	37.5	17.5	25.0	20.0
	Ic	40.0	10.0	30.0	20.0
	V	29.2	12.5	25.0	33.3
48 Na-K ATPase β1 subunit	T	18.5	40.2	19.8	21.5
	C	10.9	50.9	16.4	21.8
	I	25.0	28.4	21.6	25.0
	Ic	13.8	24.1	34.5	27.6
	V	32.8	25.0	28.1	14.1

Table 1. Continued

		G	C	A	T
49 CD3 ϵ antigen	T	20.2	31.0	22.0	26.8
	C	19.2	38.5	15.4	26.9
		21.4	17.9	33.9	26.8
	Ic	21.4	14.3	28.6	35.7
	V	25.0	25.0	25.0	25.0
50 Na-Ca exchanger	T	23.4	33.1	18.5	25.0
	C	26.4	38.9	13.4	21.3
	I	18.8	26.2	26.0	28.9
	Ic	16.5	19.0	28.1	36.4
	V	22.3	26.5	21.2	29.9
51 Ca-ATPase	T	22.8	27.8	19.1	30.4
	C	24.7	28.9	16.7	29.7
		18.9	25.3	21.7	34.1
	Ic	13.9	25.8	27.2	33.1
	V	24.4	29.0	21.0	25.6
52 Urate oxidase	T	20.5	32.8	23.0	23.7
	C	22.7	38.6	18.2	20.5
		14.9	29.1	24.3	31.8
	Ic	13.5	24.3	27.0	35.1
	V	26.4	26.4	31.9	15.3
53 Selectin	T	17.7	28.2	28.0	26.1
	C	10.2	32.7	28.6	28.6
		20.9	27.0	26.5	25.5
	Ic	10.2	26.5	34.7	28.6
	V	28.9	19.7	30.3	21.1
54 Prolactin receptor	T	19.0	28.6	23.8	28.6
	C	21.9	26.6	29.7	21.9
		12.8	32.3	15.9	39.0
	Ic	14.6	34.1	12.2	39.0
	V	22.6	27.4	21.4	28.6
55 Link protein	T	25.3	29.0	25.2	20.5
	C	25.5	30.9	25.5	18.1
	I	25.6	26.9	25.0	22.4
	Ic	25.6	30.8	25.6	17.9
	V	23.5	23.5	23.5	29.4
56 SOD Cu-Zn	T	21.9	29.8	24.6	23.7
	C	20.8	25.0	33.3	20.8
	I	20.7	38.0	16.3	25.0
	Ic	17.4	39.1	17.4	26.1
	V	27.5	22.5	22.5	27.5
57 Flavin-containing monoxygenase	T	20.2	29.6	27.7	22.4
	C	19.8	29.2	32.3	18.8
	I	20.3	31.0	24.1	24.6
	Ic	15.5	36.2	24.1	24.1
	V	21.7	28.3	20.0	30.0
58 Pancreatic triglyceride lipase	T	20.3	27.0	29.5	23.2
	C	22.0	28.8	25.4	23.7
	I	18.2	25.0	33.2	23.6
	Ic	12.7	20.0	40.0	27.3
	V	21.0	27.0	31.0	21.0
59 Osteopontin	T	10.6	29.8	22.3	37.2
	C	10.0	30.0	15.0	45.0
	I	11.9	28.6	31.0	28.6
	Ic	4.8	23.8	38.1	33.3
	V	8.3	33.3	16.7	41.7
60 Casein kinase II α subunit	T	18.2	22.7	32.2	27.0
	C	18.0	22.0	35.0	25.0
	I	18.1	23.6	25.7	32.6
	Ic	16.7	13.9	27.8	41.7
	V	19.6	25.0	28.6	26.8
61 Apo H	T	16.4	21.4	37.8	24.5
	C	9.8	22.0	41.5	26.8
	I	18.6	19.3	43.6	18.6
	Ic	8.6	17.1	54.3	20.0
	V	26.2	23.8	20.0	30.0
62 Calpastatin	T	12.8	24.6	29.9	32.7
	C	9.6	23.1	32.7	34.6
		10.7	24.4	27.4	37.5
	Ic	4.8	21.4	33.3	40.5
	V	21.6	27.6	28.4	22.4
63 Stem cell factor/ Kit ligand	T	15.9	26.3	24.7	33.1
	C	15.5	27.6	24.1	32.8
	I	8.3	16.7	33.3	41.7
	Ic	0.0	8.3	33.3	58.3
	V	32.1	32.1	14.3	21.4
64 Serum albumin	T	19.5	26.3	22.4	31.8
	C	21.4	23.2	21.4	33.9
	I	17.8	28.0	22.0	32.2
	Ic	10.2	25.4	28.8	35.6
	V	19.4	30.6	26.4	23.6
65 HSP 108	T	17.2	21.2	26.4	35.2
	C	15.8	19.2	26.0	39.0
	I	15.6	23.2	26.3	34.9
	Ic	16.7	18.8	27.1	37.5
	V	25.6	23.8	28.0	22.6
66 Macrophage scavenger	T	15.7	18.8	29.9	35.6
	C	16.0	6.0	38.0	40.0
	I	13.7	28.6	25.6	32.1
	Ic	14.3	28.6	19.0	38.1
	V	20.3	32.8	15.6	31.2
67 Protein phosphatase X catalytic	T	18.3	20.0	30.4	31.3
	C	16.7	19.2	30.8	33.3
	I	16.0	20.1	34.0	29.9
	Ic	16.7	16.7	30.6	36.1
	V	28.8	22.5	22.5	26.2
68 Rab 2	T	11.9	22.6	35.4	30.2
	C	8.2	22.4	38.8	30.6
	I	18.3	22.1	32.7	26.9
	Ic	11.5	26.9	38.5	23.1
	V	14.3	25.0	21.4	39.3
69 Rab 1	T	10.6	15.5	36.7	37.2
	C	9.1	15.2	36.4	39.4
	I	13.0	16.3	39.1	31.5
	Ic	0.0	21.7	52.2	26.1
	V	25.0	16.7	25.0	33.3

[a] Data concern the average of all synonymous positions (T) as well as the average of conserved (C), intermediate (I), and variable (V) positions. Ic refer to the three identical nucleotides present in the intermediate positions

three classes of positions using G+C values. In the case of GC-rich genes, GC was higher than expected in the conserved class, but lower than expected in the variable and intermediate classes.

Expectedly, strikingly different sets of codons were predominant in the conserved and in the variable class, the bias being very strong in the first case and very weak in the second. These results will be presented in detail elsewhere.

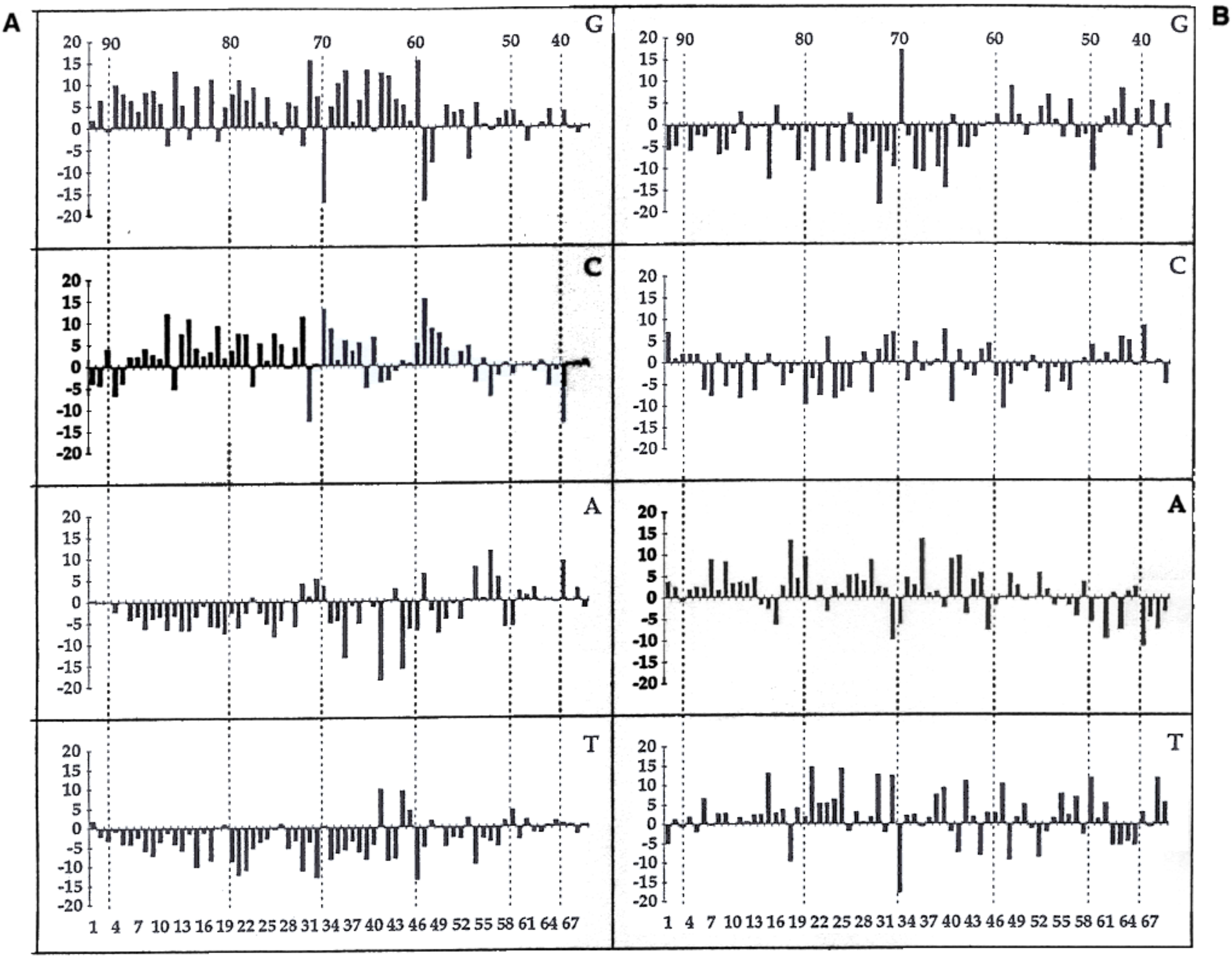

Fig. 4. Difference histograms of base compositions of (**A; left**) conserved and (**B; right**) variable synonymous positions from actual sequences and from the corresponding positions of simulated sequences. For other indications, see legend of Fig. 2.

Another comparison done (for reasons to be explained in the Discussion) was between the compositions of the Ic subclass—namely, of the three identical positions of the intermediate class—and those of the conserved positions. Difference histograms show that GC values were, in general, considerably lower than those of the conserved positions, whereas they were higher for GC-rich genes and lower for GC-poor genes compared to those of the variable class (Fig. 6).

χ^2 evaluations of the data for all genes examined are presented in Table 2. The percentage of genes showing significantly deviating GC_3 values was 38% (and mainly concerned conserved positions), but this value increased to 74% for the GC_3-richest third of the sample and decreased to 20% for the GC-poorest two thirds. It should be stressed that the first third comprises genes higher than 75% GC_3 and that the number of significantly deviating genes less than 50 synonymous positions in size was negligible. Nonsignificant values were obtained for genes from the same order (not shown).

The Base Compositions of the First Positions of Codons Following Conserved Positions Exhibit Strong Biases in GC-rich Genes

An analysis of the base compositions of the first and second positions of the codons following conserved and variable positions showed a number of interesting features (Table 3, left panel). While the second positions were quite similar whether following conserved or variable codons, the first positions next to variable positions were very different from those following conserved positions. Some of these features (like the scarcity of AT and TA doublets in positions 3-1) were simply due to the richness in G and C of conserved relative to variable synonymous positions (see below). A striking feature, however, was the extreme shortage of the 3-1 doublet CpG with C in a conserved position as opposed to the lack of shortage when C was in a variable position. It is of interest to note (Table 3, right panel) that first codon

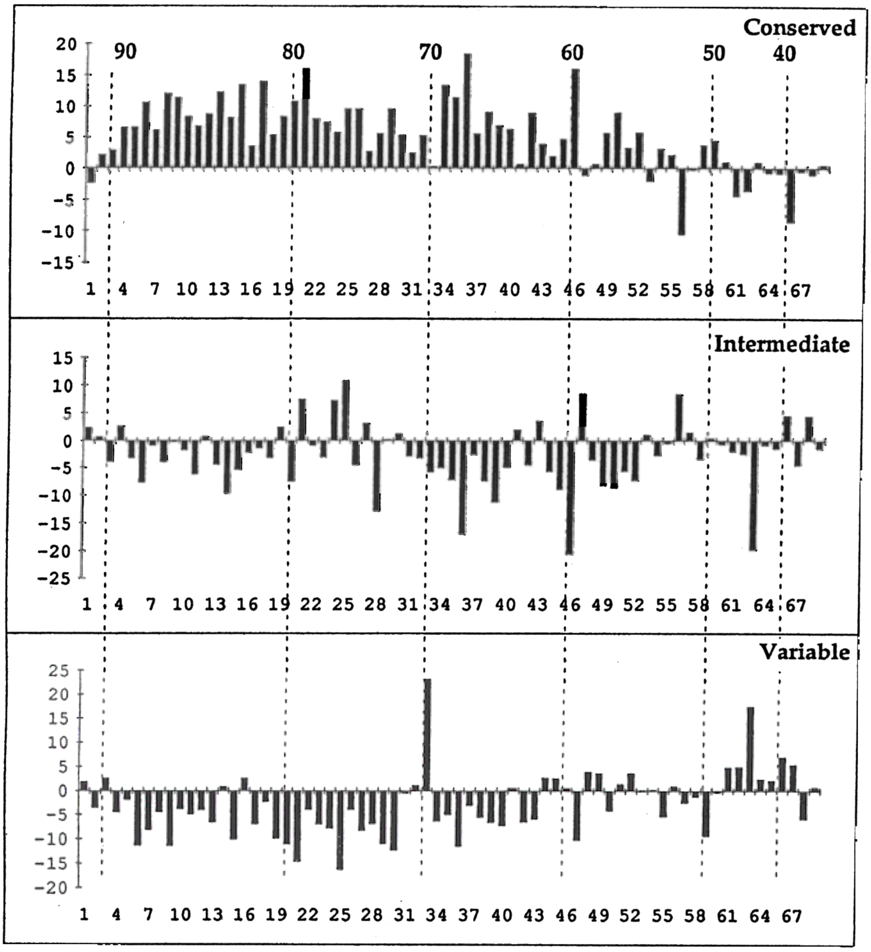

Fig. 5. Difference histograms of GC levels of conserved, intermediate, and variable synonymous positions from actual sequences and from the corresponding positions of simulated sequences. For other indications see legend of Fig. 2.

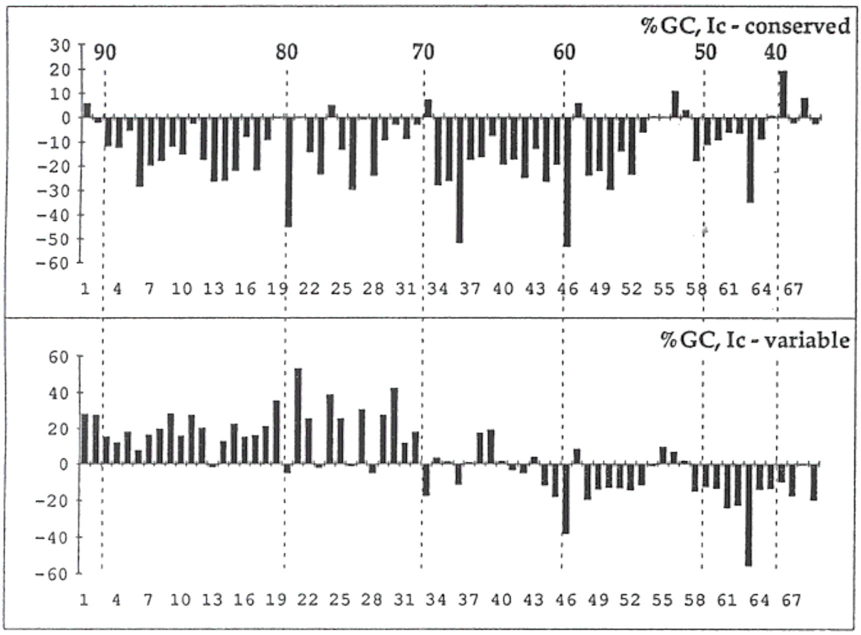

Fig. 6. Difference histograms of GC levels of Ic positions and conserved or variable positions. For other indications see legend of Fig. 2.

positions following third positions of synonymous quartet codons are highly conserved.

The results just mentioned concern the composition of the two classes of positions as averaged over all the genes analyzed. When the gene sample was split into three sets according to GC_3, it became even clearer that the scarcity of AT and TA in positions 3-1 was just due to composition since it also appeared in the GC_3-richest third of variable positions, but the difference in CpG frequencies between conserved and variable positions

Table 2 χ^2 values obtained by comparing the actual percentage of G+C in conserved (C), intermediate (I), variable (V), and all (CIV) positions from quartet codons with expectations based on random substitution process[a]

	χ^2				$P\leq$			
Gene	C	I	V	CIV	C		V	CIV
1* Apo E	1.15	0.88	0.11	0.09	0.300	0.500	0.750	0.800
2 Creatin kinase B	1.16	0.08	0.72	6.27	0.300	0.800	0.500	**0.025**
3 A1 adenosine	2.19	3.14	0.34	4.78	0.200	0.100	0.700	**0.050**
4* H,K ATPase β subunit	1.55	0.53	1.15	2.44	0.250	0.500	0.300	0.200
5* α-globin	2.16	0.62	0.08	5.76	0.200	0.500	0.800	**0.025**
6* Apo A1	7.58	3.02	1.78	9.39	**0.010**	0.100	0.200	**0.005**
7 Na-H exchange protein	15.65	0.36	7.12	41.05	**0.001**	0.700	0.010	**0.001**
8 Serine pyruvate aa transferase	10.91	1.99	1.57	15.35	**0.001**	0.200	0.250	**0.001**
9 Dipeptidase	10.39	0.01	7.33	18.21	**0.005**	1.000	**0.010**	**0.001**
10 GMP-phosphodiesterase α	9.09	0.82	2.46	21.89	**0.005**	0.500	0.200	**0.001**
11* CD8 α chain	1.34	1.77	0.50	1.68	**0.250**	0.200	0.500	0.200
12* Glutathione peroxidase	4.54	0.05	0.89	16.26	**0.050**	0.900	0.500	**0.001**
13 H,K ATPase α subunit	38.2	8.20	10.16	61.35	**0.001**	**0.005**	**0.005**	**0.001**
14* Retinol-binding protein	2.45	4.02	0.01	1.65	0.200	0.050	1.000	0.200
15 Glucose Glut3	28.95	5.91	8.88	43.74	**0.001**	**0.025**	**0.005**	**0.001**
16 Prostaglandin E receptor	1.58	0.75	0.35	3.73	0.250	0.500		0.100
17 Prolyl-4-hydroxylase β	9.80	0.29	5.39	11.53	**0.005**	0.700		**0.001**
18* Growth hormone	2.16	0.67	0.12	5.35	0.200	0.500		**0.025**
19* TNF α	2.68	0.35	3.53	5.12	0.200	0.700		**0.025**
20* Ferritin L	5.53	2.45	1.99	9.29	**0.025**	0.200		**0.005**
21** Myoglobin	6.52	1.82	5.22	16.14	**0.025**	0.200		**0.001**
22** Apo CIII	0.98	0.02	0.25	1.10	0.500	0.900		0.300
23* TNF β	2.08	0.37	1.03	3.91	0.200	0.700		**0.050**
24* Hydrophobic-surfactant-associated factor	0.81	2.08	2.19	7.99	0.500	0.200	0.200	**0.005**
25** Phospholipase A2	1.19	2.04	4.78	2.23	0.300	0.200	**0.050**	0.200
26 Phenyl tRNA ligase	5.54	2.24	1.08	5.85	**0.025**	0.200	0.300	**0.025**
27* Tissue inhibitor of metalloproteinase	0.18	0.34	1.20	0.75	0.700	0.700	0.300	0.500
28 Guanine-nt-binding protein	5.53	8.99	0.53	7.74	**0.025**	**0.005**	0.500	**0.010**
29 Polymeric Ig receptor	6.22	0.00	6.86	11.02	**0.025**	1.000	**0.010**	**0.001**
30** Colony-stimulating factor	0.75	0.04	1.78	1.84	0.500	0.900	0.200	0.200
31* CD4 antigen	0.26	0.53	0.01	0.01	0.700	0.500	1.000	1.000
32* Erythropoietin	0.81	0.39	0.03	0.90	0.500	0.700	0.900	0.500
33** Gastrin	0.00	0.45	3.66	0.04	1.000	0.700	0.100	0.900
34 Na$^+$ nucleoside	13.79	3.55	3.95	12.98	**0.001**	0.100	**0.050**	**0.001**
35 D-amino-acid oxidase	6.68	3.61	1.12	5.25	**0.010**	0.100	0.300	**0.025**
36** Ferritin H	9.43	6.17	1.22	5.30	**0.005**	**0.025**	0.300	**0.025**
37 Protein kinase C	3.59	1.06	0.81	6.74		0.500	0.500	**0.010**
38** ANP	1.58	1.25	0.39	2.95		0.300	0.700	0.100
39* β-globin	2.63	3.97	0.42	2.84		**0.050**	0.700	0.100
40 Potassium channel	4.31	1.97	1.74	3.36		0.200	0.200	0.100
41** Cytochrome b5	0.00	0.08	0.01	0.24		0.800	1.000	0.700
42** Endothelin	1.10	0.46	0.65	0.73		0.500	0.500	0.500
43 Phagocytic glycoprotein I	0.41	0.94	2.51	1.44		0.500	0.200	0.250
44** Prolactin	0.04	0.69	0.19	0.22		0.500	0.700	0.700
45** Interleukin 2 receptor	0.31	1.95	0.12	0.00		0.200	0.750	1.000
46** Tissue factor	4.38	11.60	0.01	0.48		**0.001**	1.000	0.500
47** β2 microglobulin	0.01	1.06	1.46	0.15		0.500	0.250	0.700
48 Na-K ATPase β-1 subunit	0.02	0.60	0.35	0.00		0.500	0.700	1.000
49** CD3 ϵ antigen	0.98	0.94	0.04	0.56		0.500	0.900	0.500
50 Na-Ca exchange protein	14.1	14.44	1.84	14.29		**0.001**	0.200	**0.001**
51 Ca-ATPase	2.16	8.22	0.22	0.17		**0.005**	0.700	0.700
52* Urate oxidase	1.22	2.97	0.43	0.26		0.100	0.700	0.700
53 Selectin	0.17	0.08	0.00	0.21		0.800	1.000	0.700
54 Prolactin receptor	0.49	0.49	0.00	0.81		0.500	1.000	0.500
55 Link protein	0.50	0.02	0.59	1.06		0.900	0.500	0.500
56* SOD Cu/Zn	2.03	2.07	0.02	0.69		0.200	0.900	0.500
57 Flavin-containing monooxygenase	0.01	0.21	0.33	0.00		0.700	0.700	1.000

Table 2 Continued

Gene	χ^2				$P \leq$			
	C		V	CIV	C		V	CIV
58 Pancreatic triglyceride lipase	0.60	0.89	0.09	0.00	0.500	0.500	0.800	1.000
59** Osteopontin	0.43	0.00	1.64	0.67	0.700	1.000	0.200	0.500
60 Casein kinase II α subunit	0.12	0.03	0.00	0.60	0.750	0.900	1.000	0.500
61* Apo H	0.65	0.22	0.94	2.79	0.500	0.700	0.500	0.100
62 Calpastatin	0.50	0.58	1.45	1.30	0.500	0.500	0.250	0.300
63* Stem cell factor/Kit ligand	0.08	8.54	1.17	1.36	0.800	**0.005**	0.300	0.250
64 Serum albumin	0.03	0.06	0.25	0.23	0.900	0.900	0.700	0.700
65 HSP 108	0.10	0.40	0.36	1.10	0.750	0.700	0.700	0.300
66 Macrophage scavenger	3.33	1.28	1.67	1.95	0.100	0.300	0.200	0.200
67 Protein phosphatase X catalytic	0.02	1.39	0.81	0.87	0.900	0.250	0.500	0.500
68* Rab 2	0.07	0.77	0.44	0.28	0.800	0.500	0.700	0.700
69* Rab 1	0.04	0.09	0.00	0.00	0.900	0.800	1.000	1.000

[a] Underlined bold figures refer to statistically significant P values ($P \leq 0.05$). Asterisks and double asterisks refer to genes with less than 100 QuS and less than 50 QuS, respectively

was maintained. Second positions did not show any feature that was significantly different in the three classes.

The same analysis was also performed (Table 4) for positions next to the conserved (Ic) nucleotides (see preceding section) and the variable (Iv) nucleotides of the intermediate class. First positions next to Ic and Iv were similar to each other and also similar to those found next to variable nucleotides. A moderate degree of CpG shortage was, however, to be found in Ic positions and following first positions.

The Base Compositions of Conserved Third Positions of Synonymous Quartet and Duet Codons Are Correlated

Cacciò et al. (1995) showed that the frequencies of the three classes of synonymous positions are correlated in quartet and duet codons. These results account for the agreement between the results obtained on quartet codons in comparisons of genes from four orders and those of Mouchiroud et al. (1995) on pairwise comparisons of all synonymous positions.

Figure 7 indicates, in addition, that the conserved positions show good correlations in the base compositions of synonymous positions of quartet and duet codons. These results account for the good compositional agreement (Mouchiroud and Bernardi 1993) between third codon position of homologous mammalian genes as studied in pairwise comparisons (except for those having undergone a minor shift).

Discussion

Base Composition of Synonymous Positions of Quartet Codons from the Conserved and Variable Classes Exhibit Specific Patterns

The results obtained when comparing the base composition of the conserved class of synonymous positions from the actual and the simulated sequences (Figs. 4 and 5) clearly show that G and C are in large excess in the GC-rich genes. In the same genes, variable positions also show large differences compared to the simulated sequences, but the trends are just opposite those found for conserved positions. GC-poor genes show either nonsignificant differences or reverse trends compared to GC-rich genes. The results of the difference histograms just described are paralleled by the fact that the significant values of Table 2 are much more frequent in GC-rich than in GC-poor genes. Because of the different features exhibited by GC-rich and GC-poor genes, they will be discussed separately in the following sections.

The Compositional Patterns of Conserved Positions of Quartet Codons from GC-Rich Genes

The compositional patterns shown by the conserved synonymous positions of quartet codons of GC-rich genes relative to the simulated sequences could be interpreted in three different ways.

1. The first, "stochastic," explanation is that the nucleotide synonymous substitution process simply randomized the composition of the positions that are seen as variable and intermediate in present-day (actual) sequences, whereas conserved positions were never touched by mutations and just represent the "ancestral" situation. In other words, the three classes of positions simply are the result of a stochastic nucleotide substitution process. This explanation implies that synonymous positions in GC-rich genes were GC-richer in the common ancestor of the mammalian orders under consideration than in present-day mammals.

This "stochastic" explanation is, however, contradicted by three findings. (a) The compositional dif-

Table 3. An analysis of base composition (left panel) and conservation (right panel) of first and second base 3′ of third codon positions of synonymous quartet codons

		First base 3′				Second base 3′					First base 3′				
	3rd	A	G	C	T	A	G	C	T	Class	A	G	C	T	Total
Total															
QuC	A	20.9	55.2	14.2	9.8	32.6	21.9	26.6	18.9	C	84.1	85.1	77.2	83.5	84.2
	G	25.4	41.1	22.7	10.9	35.9	14.6	21.6	27.9	I	8.8	9.0	12.1	7.7	9.0
	C	43.8	5.4	24.0	26.8	28.3	21.3	18.4	32.0	V	7.1	5.9	10.7	8.8	6.8
	T	6.8	46.4	30.7	16.1	25.8	21.0	23.1	30.1						
QuV	A	21.1	41.3	21.4	16.1	31.2	19.9	22.9	25.9	C	60.1	76.9	79.0	70.3	77.1
	G	19.4	40.9	23.9	15.7	27.2	18.6	22.3	31.9	I	17.8	14.2	12.2	8.2	13.4
	C	24.4	35.0	25.7	14.9	26.1	19.1	24.4	30.5	V	22.1	8.8	8.8	21.5	9.5
	T	15.7	47.6	23.7	13.0	25.3	18.1	25.9	30.7						
Genes 1–23															
QuC	A	8.6	73.8	10.7	6.9	30.4	33.2	24.7	11.7	C	83.6	84.5	82.4	85.0	84.7
	G	18.9	41.9	30.7	8.5	31.3	16.9	25.4	26.4	I	9.9	10.3	12.9	6.5	9.7
	C	40.2	8.4	25.1	26.3	25.3	20.1	18.1	36.5	V	6.5	5.1	4.6	8.5	5.5
	T	3.5	54.5	32.1	9.9	27.7	20.4	18.2	33.7						
QuV	A	14.7	52.0	23.9	9.4	30.0	17.1	25.6	27.3	C	52.9	81.0	80.4	70.0	78.4
	G	10.5	48.7	30.0	10.8	26.8	15.9	22.7	34.6	I	15.8	13.5	14.6	9.7	13.2
	C	17.3	45.1	25.6	12.1	25.0	18.7	20.5	35.8	V	31.3	5.5	5.1	20.3	8.5
	T	7.5	58.4	25.6	8.4	22.3	18.2	23.9	35.6						
Genes 24–46															
QuC	A	20.7	58.6	13.9	6.7	27.9	20.4	28.8	22.9	C	85.1	84.2	72.3	81.5	83.8
	G	23.3	44.7	19.0	12.9	40.4	11.3	19.9	28.3	I	9.1	9.3	13.9	7.9	9.5
	C	44.6	4.0	27.0	24.5	30.5	22.2	18.5	28.9	V	5.8	6.4	13.8	10.5	6.7
	T	6.4	37.0	37.4	19.2	25.4	22.4	27.4	24.8						
QuV	A	21.9	38.2	21.1	18.7	29.9	25.3	22.9	21.9	C	59.1	71.3	77.7	62.7	73.2
	G	20.7	37.3	26.4	15.6	24.5	21.9	22.3	31.3	I	20.2	17.9	13.2	9.3	16.3
	C	24.1	30.4	29.2	16.3	25.9	21.8	28.7	23.6	V	20.6	10.9	9.1	28.1	10.5
	T	17.8	42.2	27.1	12.9	29.5	20.0	24.5	26.1						
Genes 47–69															
QuC	A	29.0	40.0	16.6	14.4	37.9	15.9	26.0	20.2	C	83.8	86.6	76.9	84.1	84.0
	G	33.9	36.6	18.4	11.2	35.9	15.5	19.6	29.0	I	7.3	7.3	9.5	8.7	7.9
	C	46.6	4.0	19.8	29.6	29.1	21.5	18.7	30.8	V	9.0	6.1	13.6	7.3	8.1
	T	10.2	48.0	23.1	18.7	24.3	20.2	23.6	31.9						
QuV	A	26.7	33.8	19.3	20.2	33.8	17.3	20.3	28.6	C	68.3	78.3	79.0	77.8	79.6
	G	27.1	36.6	15.5	20.8	30.5	18.2	21.8	29.5	I	17.3	11.5	8.3	6.0	10.7
	C	31.7	29.5	22.4	16.4	27.3	16.7	23.9	32.0	V	14.3	10.2	12.7	16.2	9.6
	T	21.8	41.9	18.6	17.7	24.3	16.2	29.1	30.3						

ferences between the ''ancestral'' and the actual sequences are in fact very small. This is in contrast to the much stronger compositional heterogeneity in the ''ancestral'' compared to the present-day sequence that should be expected, according to the ''stochastic'' explanation. Even if more refined methods than the construction of a consensus sequence could be used to simulate the ancestral sequence, it is most doubtful that large compositional differences would be found. (b) The majority of nucleotides (three out of four) in intermediate positions (the Ic subclass) do not mimic the composition of the conserved positions, as expected according to the ''stochastic'' explanation. Instead, the composition of the Ic subclass of intermediate positions is very different from that of conserved positions (Fig. 6), being much poorer in G and C. The different base compositions of the first positions next to Ic nucleotides (see Table 4) compared to those of conserved positions confirms this difference. (c) The synonymous nucleotide substitution process was found not to be a purely stochastic one already when studying the frequencies of the conserved, intermediate, and variable classes, which were found to be significantly different in the actual sequences (especially in GC-rich genes) compared to expectations based on a random substitution process (Cacciò et al. 1995).

2. The second, ''mutationist,'' explanation is that there is a local mutation rate variation within each gene. This explanation is difficult to reconcile, however, with the following observations: (i) a GC excess is only found in conserved positions, whereas variable positions, which accumulate nucleotide substitutions, exhibit the opposite trend; (ii) there is no nucleotide context pointing to the fact that variable positions are ''hot spots'' for nucleotide substitutions; (iii) the GC excess in conserved positions is essentially only observed in GC-rich and not in GC-poor genes; (iv) the

Table 4. An analysis of first and second base 3′ of third codon positions of synonymous quartet codons beloging to the Ic and Iv classes[a]

		First base 3′				Second base 3′			
	3rd	A	G	C	T	A	G	C	T
Total									
QuI	A	23.5	39.4	19.6	17.6	41.9	16.6	20.6	20.9
	G	24.0	33.9	26.7	15.4	35.3	18.2	19.4	27.1
	C	32.7	13.6	32.4	21.3	27.6	18.9	19.2	34.3
	T	11.8	53.3	17.6	17.3	29.1	17.4	24.6	28.8
	A	22.8	38.1	21.6	17.5	39.4	18.2	21.9	20.6
	G	24.5	33.9	25.7	15.9	35.5	18.2	19.2	27.0
	C	28.2	18.6	32.6	20.6	27.2	18.4	17.7	36.8
	T	15.3	49.2	18.7	16.8	27.5	17.2	23.5	31.8
Genes 1–23									
QuI	A	12.6	51.6	21.5	14.4	40.2	13.6	31.6	14.6
	G	18.7	39.9	28.3	13.1	28.0	21.5	18.5	32.0
	C	25.7	19.9	35.3	19.0	30.2	13.3	17.9	38.6
	T	7.3	62.1	17.5	13.1	15.8	16.6	34.8	32.8
QuIv	A	8.7	49.7	23.8	17.8	33.4	15.2	37.0	14.4
	G	22.2	38.0	27.6	12.2	31.0	21.6	13.4	33.9
	C	22.0	21.8	36.9	19.4	31.4	13.0	16.8	38.9
	T	5.4	57.3	21.4	15.9	12.3	17.0	28.2	42.5
Genes 24–46									
QuI	A	26.9	33.7	18.1	21.3	42.1	19.0	17.6	21.3
	G	24.3	27.3	34.6	13.9	43.8	17.1	16.2	22.9
	C	33.2	9.6	34.6	22.6	24.9	22.2	19.7	33.2
	T	11.2	54.4	16.0	18.4	36.6	16.6	18.7	28.0
	A	28.1	32.8	18.8	20.4	39.1	22.2	18.1	20.6
	G	24.5	27.0	34.1	14.3	45.2	15.1	18.6	21.1
	C	28.8	16.7	34.8	19.8	24.1	22.2	17.6	36.1
	T	20.2	51.0	15.7	13.1	38.3	14.7	19.2	27.8
Genes 47–69									
QuI	A	27.8	36.3	19.6	16.3	42.9	16.4	15.9	24.8
	G	29.2	34.3	17.0	19.5	34.3	15.9	23.7	26.1
	C	39.1	11.4	27.4	22.2	27.7	21.2	19.9	31.2
	T	16.5	44.1	19.2	20.2	34.2	18.8	20.9	26.1
	A	27.5	35.1	22.8	14.6	43.8	16.4	14.9	24.9
	G	26.9	36.4	15.3	21.3	28.3	17.7	28.3	25.7
	C	33.9	17.4	26.1	22.7	26.2	19.9	18.6	35.3
	T	19.6	40.0	19.2	21.2	31.0	19.7	23.3	26.0

[a] Ic and Iv are the two subclasses of intermediate positions corresponding to the three identical and to the different nucleotides, respectively

levels of G and C in conserved synonymous positions of a given gene are, generally, quite different, stressing the specificity of the phenomenon under consideration. From a more general viewpoint, the correlation between mutation rate and compositional change in synonymous positions has been disproven by the work of Bernardi et al. (1993).

3. The third, ''selectionist'' explanation, is that the composition of the ''ancestral'' sequence was basically similar to the present-day sequence and conserved positions do not show changes because natural selection, at least in part, prevented them from doing so. This explanation is supported by the compositional similarity of the ''ancestral'' and present-day sequences, by the different composition of the Ic subclass of intermediate positions compared to the conserved ones, and by the nonstochasticity of the process shown by Cacciò et al. (1995). Obviously, this explanation implies that the conserved positions of the actual sequences in fact underwent a number of substitutions that were, however, lost by natural selection. A comparison of the frequency of the 3-1 CpG doublets (by far the preferred doublets for C methylation), in which C is in a synonymous conserved or in a synonymous variable position, shows a dramatic shortage in the first case but not in the second. This can be explained by the fact that C in conserved positions was methylated in ancestral sequences and had all the time to undergo deamination into T (with the consequence of an abundance of TpG which is actually observed). In contrast, C in variable positions being the result of a recent substitution event either was not methylated or, if methylated, did not have the time to undergo deamination to any appreciable extent. This analysis stresses the ''antiquity'' of C in synonymous conserved positions relative to the recent appearance of C in synonymous variable positions.

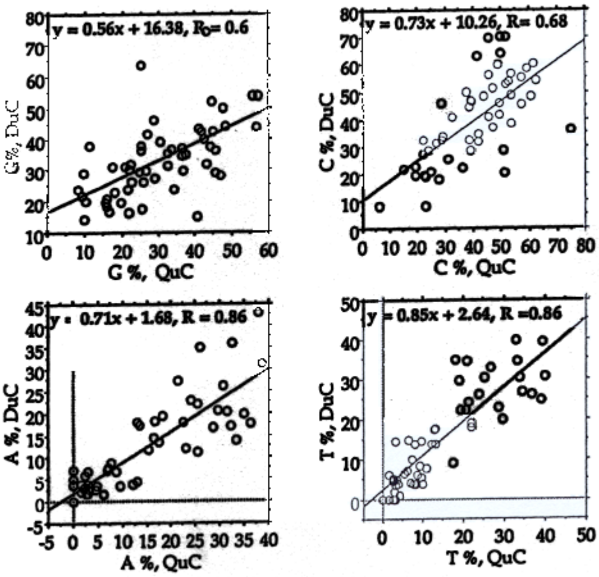

Fig. 7. Plot of the percentage of each base in duet and quartet codons. 15 genes comprising less than 150 synonymous codons were omitted in the plot (as they were in the corresponding plot of Fig. 3 of Cacciò et al. 1995).

It should be stressed that a four-order comparison, such as that carried out here because of the availability of sequences, is not ideal for several reasons. First, a set of sequences covering more orders would restrict conserved positions and enhance differences compared to expectations based on a random process. In other words, what is considered to be "conserved" at the four-order level in fact comprises not only "conserved" positions but also positions potentially belonging to the other two classes. Second, the number of conserved positions having escaped the substitution process simply because of its stochasticity (and not because of natural selection) will decrease when the number of orders examined is increased. Third, the necessity of using murid genes, which underwent a compositional shift (see Cacciò et al. 1995), can only reduce the number of GC-rich genes showing significant deviations. Finally, another reason (not related to the four-order analysis) for underestimating nucleotide substitutions in variable positions is the fact that no correction was done for multiple hits.

The Compositional Patterns of Variable Positions of Quartet Codons from GC-Rich Genes

As in the case of conserved positions, three possible explanations exist for the nucleotide substitutions accumulating in variable positions.

1. The first, "stochastic" explanation, is that variable positions are the result of random changes. This is, however, not supported by the comparison of the actual data with the simulations, since they show significant differences, even if they are less numerous than for conserved positions.
2. The second, "mutationist" explanation, is that variable positions are mutational "hot spots." However, no specific nucleotide features in the immediate neighborhood of variable positions support this viewpoint.
3. The third explanation is that these positions are not subject to any significant degree of negative selection and may accumulate mutations. As already mentioned, these are underestimated here because of the absence of correction for multiple hits.

The Compositional Patterns of Conserved and Variable Positions of Quartet Codons from GC-Poor Genes

As already mentioned, GC-poor genes do not show significant differences between actual sequences and simulated sequences, although it is not ruled out that such differences might appear (in the opposite direction of GC-rich genes) for extremely GC-poor genes. The features of GC-poor genes suggest that these genes are simply not subject to the same levels of negative selection as GC-rich genes and essentially drift.

A General Picture of the Compositional Evolution of Vertebrate Genes

The discussion presented above fits with the general picture that has emerged from the study of vertebrate genes (see Bernardi and Bernardi 1986; Bernardi et al. 1988; Bernardi 1993a,b). During the two independent transitions between reptiles and warm-blooded vertebrates (mammals and birds), nucleotide substitutions took place that increased GC levels in all codon positions, but especially in third codon positions. Such "directional" substitutions did not, however, affect all genes, but only those located in gene-dense compartments. These genes represent the vast majority (over 80%; see Bernardi 1993a; and S. Zoubak and G. Bernardi, paper in preparation) of mammalian genes and were themselves already slightly richer in GC than the other genes in cold-blooded vertebrates (Cacciò et al. in preparation). The enrichment in GC that occurred between the ancestral reptilian genomes and the genomes of present-day warm-blooded vertebrates can be visualized as a process (called the transitional mode of evolution; Bernardi et al. 1988) that was caused by positive selection and/or by a process of negative selection with a steadily increasing GC threshold. When an equilibrium was reached in warm-blooded vertebrates, as witnessed by the similar compositional patterns exhibited by mammalian and avian genomes (Sabeur et al. 1993; Mouchiroud and Bernardi 1993; Kadi et al. 1993; the minor shifts of some mammalian genomes are neglected here), codon positions of GC-rich genes were subjected to a certain degree of negative selection which was discussed in the previous sections (and which was called the conservative mode of evolution; Bernardi et al. 1988). Needless to say, the above explanation implies that GC-poor genes were generally subjected to a much lower degree of negative selection compared to GC-rich genes, as shown by their low compositional deviations relative to statistical expectations.

Some General Issues

A general conclusion of this work is that we can reject the widespread idea that almost all silent sites in mammalian genes are likely to be free to accept nucleotide substitutions, since an analysis of compositional patterns of the synonymous positions in the three classes of quartet codons disproves it. It should be stressed here that the simulation procedure used was precisely designed to test this idea, because it involved the reshuffling of synonymous nucleotide substitutions in quartet codons.

Indeed, the specific synonymous substitution patterns observed in the present work imply functional constraints and negative selection. As a consequence, the nonrandomness of the frequencies of substitutions over the genes (Cacciò et al. 1995), as well as the gene specificity of synonymous substitution frequencies (Mouchiroud et al. 1995), should also be understood as being due to functional constraints and negative selection.

The correlation of synonymous and nonsynonymous substitution frequencies is what should be expected, under a negative selection paradigm, if the conservation of the protein sequence is coupled with the conservation of the corresponding gene, including its synonymous sites which should be visualized as relevant for the regulation of gene function(s). In agreement with these conclusions, in the evolution genes for highly conserved proteins, like the actin genes, a few synonymous substitutions at third codon positions are forbidden or very rare (Britten 1993).

It is very striking that all above considerations essentially apply to GC-rich genes, which represent, however, the majority of human genes, and only to a much lesser extent to GC-poor genes. A more detailed discussion of the general issues raised by this work will be presented elsewhere.

Acknowledgments. We thank most warmly Laurent Duret for having provided us with the sequence alignments used in the present work and for very useful discussions, and Adam Eyre-Walker, Takashi Gojobori, Wen-Hsiung Li, Tomoko Ohta, and Ken Wolfe for critical reading of this paper.

References

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

Bernardi G (1993b) The isochore organization of the human genome, its evolutionary history: a review. Gene 135:57–66

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

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

Bernardi G, Mouchiroud D, Gautier C (1993) Silent substitutions in mammalian genomes and their evolutionary implications. J Mol Evol 37:583–589

Britten RJ (1993) Forbidden synonymous substitutions in coding regions. Mol Biol Evol 10:205–220

Bulmer M, Wolfe KH, Sharp PM (1991) Synonymous nucleotide substitution rates in mammalian genes: implications for the molicular clock and the relationship of mammalian orders. Proc Natl Acad Sci USA 88:5974–5978

Cacciò S, Zoubak S, D'Onofrio G, Bernardi G (1995) Nonrandom frequency patterns of synonymous substitutions in homologous mammalian genes. J Mol Evol 40:280–292

Kadi F, Mouchiroud D, Sabeur G, Bernardi G (1993) The compositional patterns of the avian genomes and their evolutionary implications. J Mol Evol 37:544–551

Li W-H, Graur D (1991) Fundamentals of molecular evolution. Sinauer, Sunderland, MA

Mouchiroud D, Bernardi G (1993) Compositional properties of coding sequences and mammalian phylogeny. J Mol Evol 37:441–456

Mouchiroud D, Gautier C, Bernardi G (1995) Frequencies of synonymous substitutions in mammals are gene-specific and correlated with frequencies of nonsynonymous substitutions. J Mol Evol 40:107–113

Sabeur G, Macaya G, Kadi F, Bernardi G (1993) The isochore patterns of mammalian genomes and their phylogenetic implications. J Mol Evol 37:93–108