J Mol Evol (1988) 27:311–320

Journal of Molecular Evolution

The Compositional Distribution of Coding Sequences and DNA Molecules in Humans and Murids

Dominique Mouchiroud,[1] Christian Gautier,[1] and Giorgio Bernardi[2]

[1] Laboratoire de Biométrie, U.A. 243, Université Claude Bernard Lyon I, 69622 Lyon, France
[2] Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France

Summary. The compositional distributions of coding sequences and DNA molecules (in the 50–100-kb range) are remarkably narrower in murids (rat and mouse) compared to humans (as well as to all other mammals explored so far). In murids, both distributions begin at higher and end at lower GC values. A comparison of homologous coding sequences from murids and humans revealed that their different compositional distributions are due to differences in GC levels in all three codon positions, particularly of genes located at both ends of the distribution. In turn, these differences are responsible for differences in both codon usage and amino acids. When GC levels at first + second codon positions and third codon positions, respectively, of murid genes are plotted against corresponding GC levels of homologous human genes, linear relationships (with very high correlation coefficients and slopes of about 0.78 and 0.60, respectively) are found. This indicates a conservation of the order of GC levels in homologous genes from humans and murids. (The same comparison for mouse and rat genes indicates a conservation of GC levels of homologous genes.) A similar linear relationship was observed when plotting GC levels of corresponding DNA fractions (as obtained by density gradient centrifugation in the presence of a sequence-specific ligand) from mouse and human. These findings indicate that orderly compositional changes affecting not only coding sequences but also noncoding sequences took place since the divergence of murids. Such directional fixations of mutations point to the existence of selective pressures affecting the genome as a whole.

Key words: Genome composition — Coding sequences — Isochores — Humans — Murids

Offprint requests to: G. Bernardi

Introduction

The analysis of the compositional distribution of DNA molecules using density gradient centrifugation in the presence of sequence-specific ligands has revealed that vertebrate genomes consist of very long DNA stretches, the isochores, which are fairly uniform in base composition from 3 kb to over 300 kb. The compositional distribution of isochores, as seen at the level of DNA molecules 50–100 kb in size, is narrow and centered around low (≃40%) GC values for the vast majority of cold-blooded vertebrates. In contrast, the distribution found in warm-blooded vertebrates is broad. Although it covers the low GC range of the isochores from cold-blooded vertebrates, it also extends into a higher GC range. Roughly, one-third of the genomes from warm-blooded vertebrates are in a GC range that is not, or is very poorly, represented in cold-blooded vertebrates (see Bernardi et al. 1985 for a review).

Our understanding of the organization of the vertebrate genome can be pushed further if the approach at the DNA level that has just been outlined is supplemented (1) by reassociation kinetic studies (Cuny et al. 1978; Soriano et al. 1981), (2) by the localization of genes in DNA fractions (Cuny et al. 1978; Cortadas et al. 1979), or (3) by an analysis of the compositional distribution of coding sequences (Bernardi et al. 1985). The distribution of the latter (which only represent a few percent of vertebrate genomes) is again very different in the genomes of cold-blooded vertebrates, where most genes are GC-poor (as are most DNA molecules), compared to that of warm-blooded vertebrates, where the gene concentration attains the highest values in the GC-richest regions, which are the least represented in the genome.

Both approaches, at the DNA and at the coding sequence level, have recently progressed further. In-

deed, the known primary structures of an increasingly large number of coding sequences has allowed us to define their compositional distributions better (Mouchiroud et al. 1987). These results have not only confirmed the initial observations concerning cold-blooded and warm-blooded vertebrates, but have also revealed differences between the genomes of rat and mouse, on the one hand, and humans on the other. Human coding sequences have higher GC values than those of mouse and rat, which also show a narrower distribution on the GC-poor side. These findings account for a number of codon usage differences between humans and murids (Mouchiroud and Gautier 1988). Moreover, they are paralleled by previous results at the DNA level. Indeed, the human genome contains DNA components higher in GC than those of the mouse genome (Salinas et al. 1986; Zerial et al. 1986).

Here we have defined more clearly the differences between the genome of humans and those of mouse and rat, namely of the three vertebrates for which the information available at the coding sequence level is the most abundant. Moreover, we have analyzed these differences by investigating homologous coding sequences in these three genomes in detail.

Materials and Methods

Because too few homologous coding sequences are available for all three species under consideration here, human, mouse, and rat, we have built three different data files including coding sequences that are homologous for pairs of species. Listings of the genes used are available upon request. Complete coding sequences were obtained for all genes available from GenBank (Bilofsky et al. 1986) in release 50 (May 1987). The ACNUC retrieval system (Gouy et al. 1984) was used. Computer analysis was performed by using ANALSEQ, the software for the statistical analysis of DNA sequences that has been developed by the Laboratoire de Biométrie in Lyon.

The percentage of guanosine plus cytidine (called GC or GC level henceforth) of coding sequences was characterized by two parameters: (1) GC level in the third codon position; and (2) cumulative GC levels in the first + second codon positions. These parameters were used to indicate the extent of silent and nonsilent changes, respectively. It should be recalled, however, that a minority of third-position changes are nonsilent and that a minority of first-position changes are silent.

The differences in variance for each one of these two parameters were tested by the nonparametrical Siegel-Tukey test (Lehman 1975). This can be used when two samples are roughly centered about the same value. The numerical values of the samples (designated as A and B) are mixed, ordered, and replaced by rank values in such a way that the extreme figures get low values:

Numerical values of A	1	5	5	10	29				
Numerical values of B	7	20	25	15					
Values ordered	1	5	7	10	15	20	25	29	
Rank value	1	4.5	6	9	8	7	3	2	
Sample	A	A	B	A	B	B	B	A	

The sum, SR(A), of ranks for A is computed:

$$SR(A) = \sum_{i=1}^{M} Ri,$$

where M is the number of numerical values in sample A, and Ri is the rank value. The mean and variance of SR(A) are given by

$$E[SR(A)] = \frac{M(N + 1)}{2},$$

where N is the number of numerical values in samples A and B; and

$$V[SR(A)] = \frac{M(N/2)(N + 1)}{12} - \frac{M(N/2) \sum_{i=1}^{G} (di^3 - di)}{12N(N - 1)},$$

where di is the number of ties for the ith value, and G is the number of different values.

Modifications due to the existence of ties have been done according to Lehman (1975). The normal approximation is supported by a limit theorem, which states that the null distribution of $SR(A) - E[SR(A)]/\sqrt{V[SR(A)]}$ tends to the standard normal distribution when M and N tend to infinity.

If values in the A sample are more dispersed than in the B sample (A variance > B variance), the rank sum is higher than expected for a homogeneous dispersion and the value of the test is higher than 1.96 (5% threshold for a normal law).

For testing codon usage changes, human and rat homologous sequences were aligned and the nature of changes in third positions of synonymous codons was studied. The equality of the number of silent changes (n1) from A or U in sequence S1 to C or G in sequence S2 and that (n2) from C or G in sequence S1 to A or U in sequence S2 was tested by the chi-squared test with one degree of freedom. If the chi-squared test is significant, then the changes between the two sequences are not symmetric. This implies that the substitution process leading from the common ancestor to the first sequence does not have the same statistical properties as the one leading to the second sequence (Mouchiroud and Gautier 1988).

Results

Compositional Distribution of Coding Sequences and DNA in Humans and Murids

Figure 1A and B shows that striking differences exist in the compositional distribution of third codon positions of genes from humans and murids (rat and mouse). In humans the distribution is broader and a distinct peak is evident on the GC-poor side, whereas on the GC-rich side the distribution has higher values and comprises more coding sequences. Figure 1C shows that the compositional distribution of third codon positions of genes from mammals other than humans, murids, and hamster essentially resembles that of humans and not that of murids. The data of Fig. 1C correspond mainly to bovine and rabbit genes (64 and 20, respectively); dog, goat, horse, pig, and sheep genes are also represented at a lower level (5–8 per species).

At the DNA level, a reexamination of previous results concerning mouse and human (Salinas et al. 1986; Zerial et al. 1986) revealed that differences exist not only on the GC-rich side, as already noted,

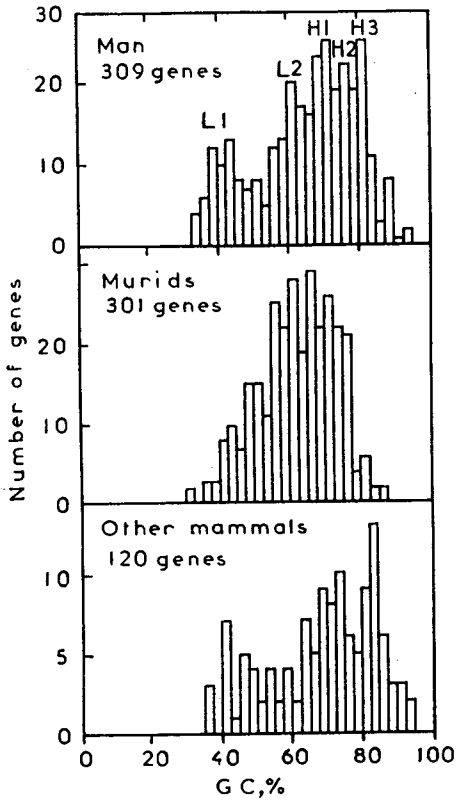

Fig. 1. Compositional distribution of third codon positions for genes from humans (A), murids (B; rat and mouse), and mammals other than humans, murids, and hamster (C). The number of genes under consideration is indicated. Genes belonging to homologous pairs in rat and mouse were considered only once. A 2.5% GC window was used. Tentative identifications of different compositional classes of coding sequences corresponding to the compartments of the human genome (L1, L2, H1, H2, and H3) are indicated following Mouchiroud et al. (1987; see Discussion).

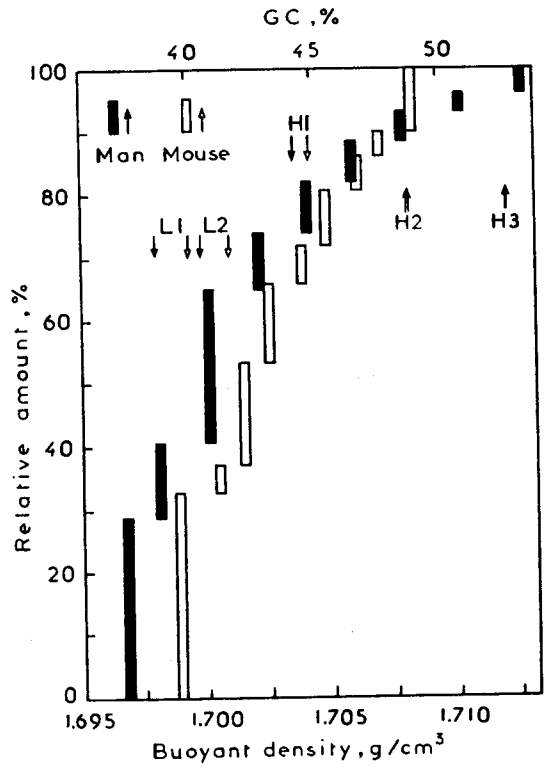

Fig. 2. Relative amounts and modal buoyant densities of DNA fractions from human (black bars) and mouse (white bars) as obtained by density gradient centrifugation in Cs_2SO_4/BAMD [BAMD is 3,6 bis (acetatomethyl-mercury) dioxane]. Data used to construct the histograms are those in Fig. 2c of Salinas et al. (1986) and of Fig. 1a of Zerial et al. (1986). In the first case, values were corrected for the presence of mouse satellite DNA. Data are plotted in a cumulative way in order to produce patterns of compositional distribution and to allow the construction of Fig. 7. Black and white arrows indicate the centers of distribution of major DNA components of man and mouse, respectively (from Cuny et al. 1981; Bernardi et al. 1985). GC levels were calculated from buoyant densities using the relationship of Schildkraut et al. (1962).

but also, and to a larger extent, on the GC-poor side of the distribution (Fig. 2). As a consequence, the average buoyant density of the mouse fractions is 1.7023 g/cm³, whereas that of human fractions is 1.7010 g/cm³. These values are in excellent agreement with mean buoyant densities calculated from the first moment of the band profile of unfractionated DNA in CsCl. Indeed $\langle\rho\rangle$ values of 1.7020 g/cm³ and 1.7010 g/cm³ were found for mouse and human, respectively (Thiery et al. 1976). Moreover,

Table 1. Comparison of GC levels in first + second and third codon positions for homologous genes from human, rat, and mouse

Sample	Gene pairs	First + second position				Third position			
		$\bar{x}$	V ($\times 10^3$)	r	S-T	$\bar{x}$	V ($\times 10^3$)	r	S-T
Human	66	48.7	3.50	1.42	−3.43*	65.3	28.42	2.25	−3.46*
Rat		48.8	2.47			65.5	12.61		
Human	55	48.3	2.56	1.22	−1.28	67.5	20.76	1.63	−1.02
Mouse		47.9	2.09			65.8	12.73		
Rat	28	47.2	1.64	1.04	−0.41	65.9	14.57	1.04	0.05
Mouse		47.3	1.57			66.9	13.96		

The mean ($\bar{x}$), variance (V), variance ratio (r), and Siegel-Tukey test value (S-T) are given for each sample. Significant values for the Siegel-Tukey test are asterisked at $P < 0.05$. All Student's t-tests for pairwise comparisons of $\bar{x}$ values were not significant

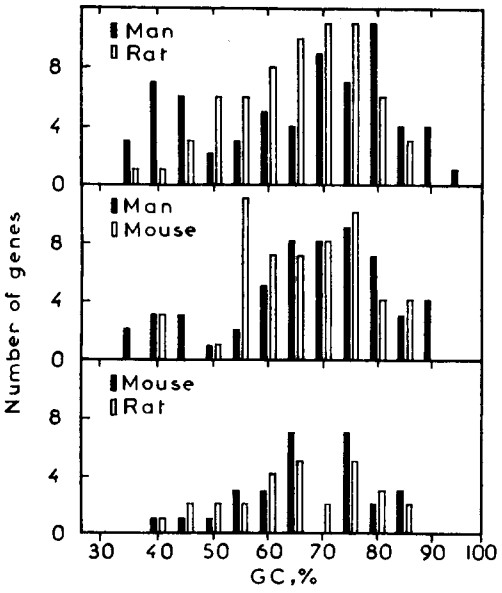

Fig. 3. Compositional distributions of third codon positions for homologous genes from human and rat, human and mouse, and rat and mouse. The numbers of gene pairs considered were 66, 55, and 28, respectively (see Table 1).

the standard deviations of the average buoyant densities from the fractions from mouse and human were 3.12 mg/cm^3 and 4.29 mg/cm^3, respectively.

A Comparison of Homologous Genes from Humans and Murids

Although the histograms of Fig. 1A and B provide an overall picture of compositional differences in coding sequences from murids and humans, a more detailed analysis of such differences can only be done on sets of homologous genes.

As shown in Table 1, GC levels of first + second codon positions and third codon positions, respectively, are not significantly different (by the Student's *t*-test; 5% threshold) for homologous genes of human and rat, human and mouse, and rat and mouse. As far as third positions are concerned, the average codon usage is very similar in humans and murids, with GC levels being high in all cases (65.3–67.5%).

In contrast, the variances of GC levels in third codon positions are very different for homologous genes of human and mouse, and, more so, of human and rat. Although the ratio of variances for homologous genes of mouse and rat is equal to one, this ratio is higher for both human/mouse and human/rat comparisons. In the latter case, but not in the former one, the variance ratio is significant according to the Siegel-Tukey test (see Table 1).

The differences just outlined become very evident when the compositional distributions of third codon positions are compared for homologous genes of humans and murids (Fig. 3). Indeed, such distribution is narrow for rat genes, whereas it is broad and at least bimodal for human genes. This difference is similar to that already observed for all (homologous + nonhomologous) genes (see Fig. 1A and B). The results obtained for human and mouse genes are similar to those just described for human and rat genes, but are less striking. Because the distributions for rat and mouse genes are very similar (Fig. 3), the differences seen in human/rat and human/mouse comparisons can only be attributed to differences in the gene samples studied. In view of the correlations found between mouse and rat coding sequences, some analyses, reported in the next two sections, were made only on the human/rat gene sample, which was larger in size and showed larger differences compared to the human/mouse sample.

Amino Acid Changes in Proteins Encoded by Homologous Genes Showing GC Differences between Human and Rat

A more detailed analysis of the GC differences exhibited by homologous human and rat genes is shown in Fig. 4 for first + second codon positions and third codon positions, respectively. In the third position, differences were concentrated at both ends of the distribution and were slightly larger for GC-poor genes (35–45% GC) than for GC-rich genes (80–90% GC). This makes sense because in the latter case, GC levels were close to the 100% limit. Third-position differences comprised 27 significant codon usage changes (chi-square test with $P < 5\%$) among the 66 gene pairs investigated (Fig. 4), namely 41% of the gene pairs tested.

Less striking GC changes, also concentrated at the two ends of the distribution and pointing in the same directions, were observed for first + second codon positions (Fig. 4). These led to amino acid changes. Examples of three genes (relaxin, apolipoprotein, and angiotensinogen genes) that are representative of low, medium, and high GC levels and that exhibit equal, lower, and higher GC levels in man compared to rat are shown in Table 2A. A comparison was made (by chi-square test) of the number of amino acid differences associated with higher GC levels in rat compared to human (R > H) with the number associated with a lower GC content (R < H). This showed that the two extreme GC-level genes exhibit modification of amino acid composition of the encoded protein, whereas the intermediate gene and its corresponding protein have the same base and amino acid composition, respectively, in human and rat. The amino acid changes are given in Table 2B.

The Nature of the Genes Showing GC Differences between Human and Rat

If the nature of the gene products is taken into consideration, one can see (Fig. 4 and Table 3) that the

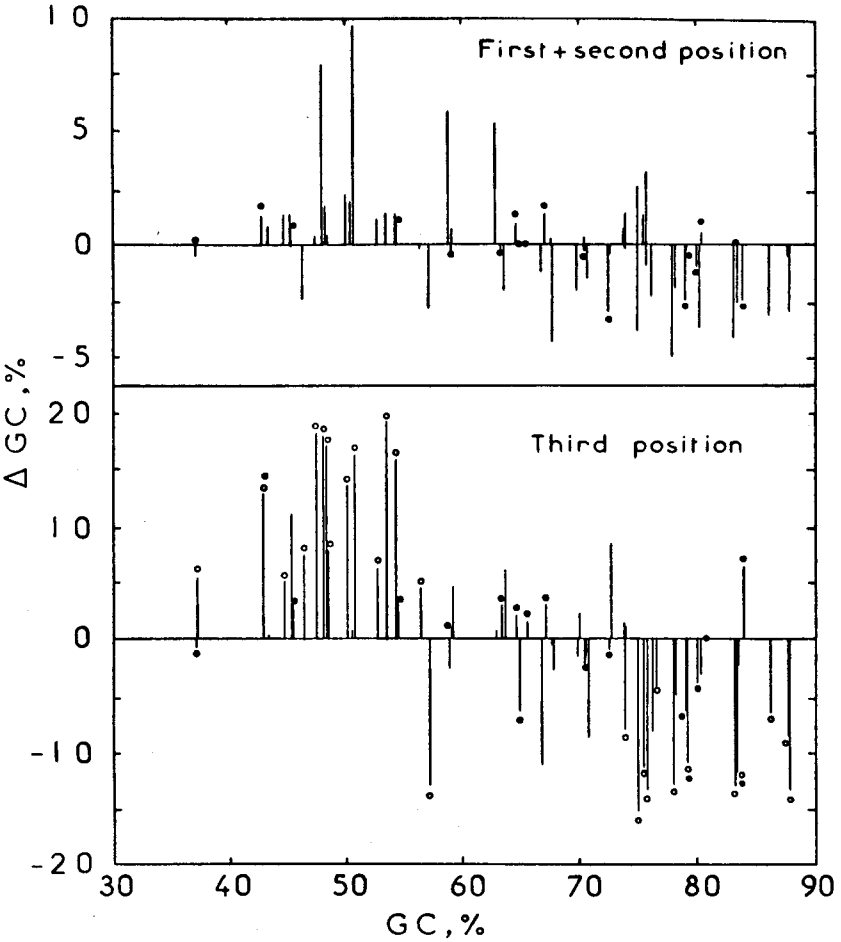

Fig. 4. GC differences for first + second and third codon positions of homologous human and rat genes are plotted (rat values less human values) against the average GC levels of third codon positions of human and rat genes. Open and closed circles indicate genes showing codon usage change and housekeeping genes, respectively.

largest differences in third-position GC are associated with genes expressed in a tissue-specific manner. These genes tend to be present at either end of the compositional distribution. They include genes coding for hormones, growth factors, proteins involved in cell regulation, and plasma proteins. In contrast, genes encoding housekeeping proteins and structural proteins showed the smallest differences, as well as a trend to be more frequent above 50% GC in the third codon position (see also Mouchiroud et al. 1987) compared to tissue-specific genes. The differences just discussed need to be strengthened, however, by data on a larger sample, particularly for housekeeping genes.

Correlations between GC Levels for Homologous Coding Sequences, for the Associated Noncoding Sequences, and for DNA Molecules from Humans and Murids

A question that was asked at this point concerned the correlations between GC levels in first + second codon positions and third codon positions, respectively, for homologous genes from humans and murids. Plots of such levels for human/rat and human/mouse genes (Figs. 5 and 6) revealed linear relationships with slopes that were very close to each other for the two pairwise comparisons and significantly different from the unity slope found for mouse/rat comparisons; the latter relationship, which passes through the origin of coordinates, is that predicted when identical GC levels are observed for homologous genes. Expectedly, the slopes in Fig. 6, concerning third codon positions (0.59 and 0.61), deviated more from unity slopes than those in Fig. 5, concerning first + second positions (0.76 and 0.81).

If the compositional distributions of DNA molecules from corresponding fractions (see Fig. 2) of human and mouse are compared as shown in Fig. 7, one can see that the two distributions intersect each other, with the mouse distribution starting at a higher and ending at a lower GC level compared to the human distribution. In this case the slope was 0.66, but this value is likely to be imprecise (see Discussion).

Discussion

Compositional Distributions of Coding Sequences and DNA in Humans and Murids

The results of Fig. 1A and B confirm, on larger samples, the differences detected by Mouchiroud et

Table 2. **A** GC differences in the two first codon positions of homologous genes from rat and human.[a] **B** Amino acid distribution in classes 0, 1, and 2

A GC differences

Relaxin gene

		Rat							
	Human	0	1	2	T		R > H[b]	R < H[b]	T[b]
ΔIII = −16.35	0	7	21	4	32	Obs.	46	16	62
ΔI + II = −9.66	1	3	16	21	40	Exp.	31	31	62
	2	0	13	5	18			$\chi^2 = 14.5$	
	T	10	50	30	90				

Angiotensinogen gene

		Rat							
	Human	0	1	2	T		R > H	R < H	T
ΔIII = −1.43	0	9	24	6	39	Obs.	64	57	121
ΔI + II = −0.66	1	20	39	34	93	Exp.	60.5	60.5	121
	2	4	33	4	41			$\chi^2 = 0.4$	
	T	33	96	44	173				

Apolipoprotein gene

		Rat							
	Human	0	1	2	T		R > H	R < H	T
ΔIII = 12.94	0	1	4	1	6	Obs.	18	38	56
ΔI + II = 4.15	1	14	23	13	50	Exp.	28	28	56
	2	5	19	5	29			$\chi^2 = 7.1$	
	T	20	46	19	85				

B Amino acid distribution in classes 0, 1, and 2

0: Asn-Ile-Leu(duet)-Lys-Met-Phe-Tyr
1: Arg(duet)-Asp-Cys-Gln-Glu-His-Leu(quartet)-Ser-Thr-Trp-Val
2: Ala-Arg(quartet)-Gly-Pro

[a] Only positions associated with amino acid changes are considered. 0, 1, and 2 indicate no G/C, one G/C, and two G/C, respectively, in first + second positions. T indicates totals. The three genes considered are representative for low, intermediate, and high GC genes, respectively. GC in the third codon positions of human genes was 42% for relaxin, 68% for angiotensinogen, and 90% for apolipoprotein genes, respectively. For each gene, GC differences between human and rat in third position (ΔIII) and first + second position (ΔI + II) are given

[b] A comparison was made by a chi-square test of the number of amino acid differences associated with higher GC levels in rat compared to human (R > H) with the one associated with a lower GC content (R < H). The 5% threshold for the test is 3.99. This showed that the two extreme GC-level genes exhibit modification of amino acid composition of the encoded protein, whereas the intermediate gene and its corresponding protein have the same base and amino acid composition, respectively, in human and rat. The amino acid changes are given in part B of the Table

Table 3. Distribution of homologous genes from human and rat according to the extent of GC differences in codon third positions (ΔIII) and the nature of the proteins encoded

	Genes coding for				
ΔIII	Hormones, growth factors, etc.	Plasma proteins	Enzymes	Structural proteins, etc.	Totals
>10%	7	7	2 (1)	3 (2)	19 (3)
5–10%	7	3	2 (1)	3 (2)	15 (3)
<5%	11 (1)	3	10 (7)	8 (4)	32 (12)
Totals	25 (1)	13	14 (9)	14 (8)	66 (18)

The number of housekeeping genes in each category is indicated in parentheses. The single housekeeping gene classified among growth factors is c-mos

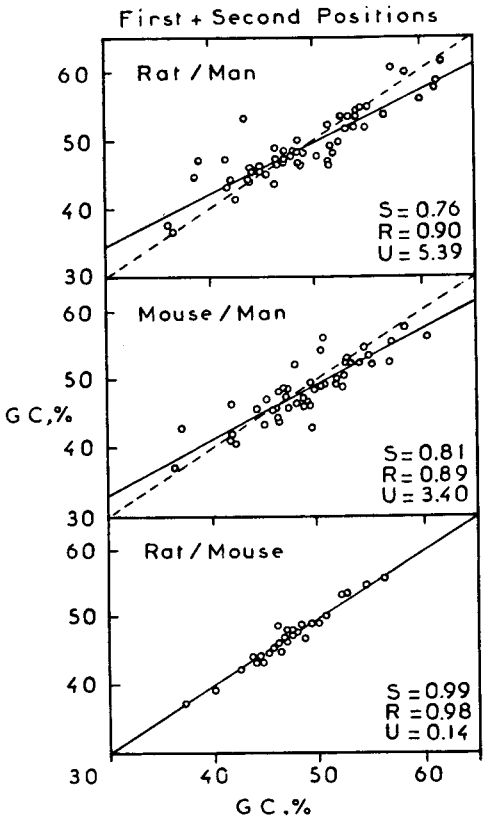

Fig. 5. GC levels of first + second codon positions of pairs of homologous genes are plotted against each other for rat and human, mouse and human, rat and mouse. In each case, data for the first species correspond to the ordinate; those for the second species to the abscissa. Lines were drawn using the least-squares method. Unity slopes are given by broken lines. Slopes (S) are given with their correlation coefficients (R) and the absolute value of a normal Σ test (U), which compares the observed slope with the unity slope (threshold at 5% is 2).

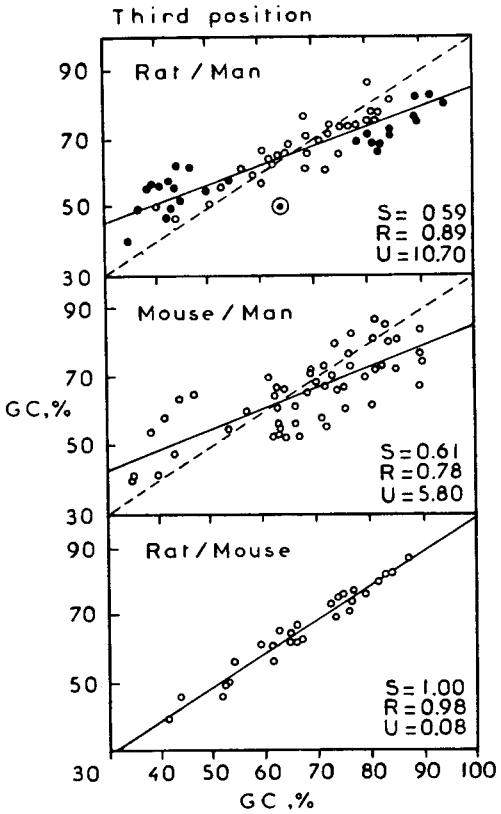

Fig. 6. GC levels of third codon positions for pairs of homologous genes are plotted against each other for rat and human, mouse and human, rat and mouse. Other indications are as in Fig. 5. Closed circles denote genes showing codon usage changes. The circled point denotes a gene (prolactin) that does not obey the relationship (see Discussion).

al. (1987) in the compositional distribution of third codon positions for genes from humans and two murids, rat and mouse. Increasing the human sample from 245 to 309 genes and pooling together the mouse and rat samples (also increased in size, in spite of the elimination of genes from homologous pairs; see Fig. 1) has not significantly changed the patterns nor the differences between humans and murids. More specifically, the larger human sample exhibits the same multimodal pattern previously described. This reinforces the notion that different compositional classes of coding sequences can be detected in the genome for which the information is more abundant. On the basis of the localization of human genes in different genome compartments (Bernardi et al. 1985; Zerial et al. 1986; Mouchiroud et al. 1987), it is possible (1) to support the tentative assignments of compositional gene classes to compositional DNA classes proposed by Mouchiroud et al. (1987) and indicated in Fig. 1A; and (2) to confirm the existence of a strong compositional gradient of gene concentration (Mouchiroud et al. 1987); the gene concentration in the H3 component is about 10 and 8 times higher than those in the L1–L2 and H1–H2 compartments, respectively.

A more detailed comparison of previous data on human and mouse DNAs has shown that the presence of two characteristic features of the composi-

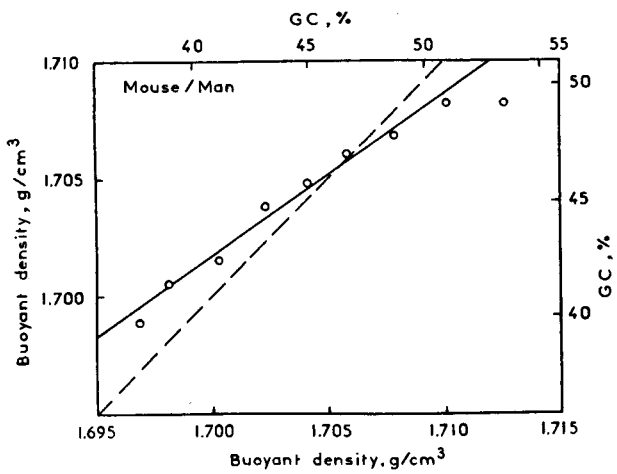

Fig. 7. Modal buoyant densities of DNA fractions from mouse and human are plotted against each other. Values corresponding to the same ordinate values of Fig. 2 were used. The broken line denotes a unity slope. GC levels were calculated as indicated in Fig. 2.

tional distribution of human coding sequences, namely the prominent GC-poor and GC-rich peaks corresponding to L1 and H3 components (Fig. 1A), is paralleled by a broader compositional distribution of human DNA molecules, which extend beyond both GC boundaries of the mouse DNA (Fig. 2). Incidentally, although the data used to construct Fig. 2 derive from the recent work of Salinas et al. (1986) and Zerial et al. (1986), previous results by Thiery et al. (1976) and Cuny et al. (1981), using fractionations in Cs_2SO_4/Ag^+ density gradients, had already led to the same conclusion.

It should be stressed here that the results used to construct Fig. 2 are rather imprecise in that modal buoyant densities of fractions were used. Moreover, they were obtained during investigations with a different purpose and are not ideally suited for the construction of the histogram. The data of Fig. 2 are, however, good enough to demonstrate the point of interest here and to construct the plot in Fig. 7 (see below).

The Compositional Pattern of Murid Genomes and Its Evolutionary Origin

The striking differences in the compositional distribution of both DNA molecules and coding sequences of humans and murids raises the question of the phylogenetic spread of these different distributions. Without ruling out the existence of minor differences between mammalian orders and families, the data so far available suggest that the human pattern is that generally found in mammals. This is indicated by the compositional distribution (1) of third codon positions of genes from mammals other than humans, murids, and hamster (Fig. 1C); and (2) of DNA molecules from the eight orders of mammals explored so far. Both distributions resemble that of humans and not that of murids.

If the murid compositional pattern is the exception and not the rule among mammals, a relevant question concerns the phylogenetic spread of the murid pattern and its evolutionary origin, respectively. The limited coding sequence data available so far indicate that hamster (which belongs to the cricetids, the rodent family that is closest to murids) shows a murid pattern.

As far as the evolutionary origin of the murid pattern is concerned, two possibilities should be taken into consideration. If murids arose relatively late in mammalian evolution, as it is generally believed, then it is difficult to escape the conclusion that their genome underwent a narrowing of its compositional distribution compared to the mammals from which they were derived, with, however, the average GC value being largely conserved. The alternative, much less orthodox, view would be that the compositional pattern of murids reflects an ancestral mammalian pattern that evolved further in other mammals, leading to a wider compositional distribution of both DNA molecules and coding sequences. Although much less likely in view of the narrow spread of the murid distribution, this possibility should be kept open because of our present ignorance of the phylogenesis and taxonomy of rodents (Eisenberg 1981). In any case, whatever the direction of changes, the causes leading to the differences between the human and the murid genomes are not yet understood.

A very remarkable feature of the murid pattern is how close it is to being identical in mouse and rat. Indeed, the alignment of points at the third codon position is very similar over a 50% GC range (from 40% to 90%) in spite of the divergence in sequence due to the accumulation of point mutations.

A Comparison of Homologous Genes from Humans and Murids

The compositional distribution of codon third positions from murid genes is narrower than that of homologous human genes (Fig. 5). This result is similar to that found for all genes compared (both homologous and nonhomologous; see Fig. 1), but the comparison of homologous coding sequences permits one to draw new conclusions concerning the meaning of the differences found.

First of all, differences in GC levels of homologous coding sequences essentially concern (Fig. 4) GC-poor and GC-rich genes, leaving the composition of coding sequences of intermediate GC level practically unaltered. GC-poor coding sequences are less GC-poor and GC-rich sequences are less GC-rich in murids compared to humans. Because the differences not only involve third codon positions, but also first and second codon positions, they lead to significant differences in the amino acids of the encoded proteins (see Table 2). Therefore, amino acids encoded by GC-poor or GC-rich codons, respectively, are favored not only at the two ends of the compositional distribution (as already shown for other genomes; see Bernardi and Bernardi 1986), but amino acid differences also exist at the two ends of compositional distributions between humans and murids.

Second, the linear relationships found when comparing GC levels in first + second codon positions and third codon positions for genes from humans and murids (Figs. 5 and 6) indicate that the order of GC levels for homologous coding sequences is conserved between the genomes that show the larg-

est compositional differences among mammals. Similar relationships were also found when GC levels of compositional classes of human and mouse DNAs were compared (Fig. 7). These findings are in keeping with the existence of the linear relationships (Bernardi et al. 1985) between the GC levels of coding sequences and the GC levels of the isochores in which coding sequences are embedded (or of the noncoding sequences flanking them, since these represent the large majority of mammalian DNA; see also Mouchiroud 1986). As expected, deviations of these linear relationships from unity slopes increased when going from first + second codon positions to third codon positions.

The linear relationships in Figs. 5–7 can only be due to ordered GC increases and decreases that have reduced the compositional gradient in the murid genome relative to other mammals (or enlarged the present-day compositional pattern of most mammals, if this is derived from a primitive one similar to the present-day murid pattern). In either case, the essential point is the ordered nature of the changes. This is so much more remarkable because it concerns isochores that, although clustered at either end of the compositional distribution, are physically scattered throughout the genome. If GC increases and decreases had taken place not in an orderly way, but at random, they would have led to changes in the order of GC levels of homologous genes, such as those associated with translocations of one gene from an isochore class to another one.

The case of the prolactin gene probably is an example of this situation. As far as first and second codon positions are concerned, this gene shows the same GC levels in human and rat, whereas the third position GC level is much lower in rat than in human (Fig. 6; circled point). A very similar case was found previously (Bernardi et al. 1985) for cardiac and skeletal α-actin genes from mouse. These genes show only 1% divergence in amino acid composition but 16% divergence in codon third position. The higher GC level of the skeletal α-actin gene was shown to be correlated with its presence in GC-rich isochores, whereas the cardiac α-actin gene was located in GC-poor isochores. Likewise, a translocation of the prolactin gene to a compositionally different isochore might have been at the origin of the situation described.

If the linear relationships discussed above hold for the mammalian genomes showing the largest difference in compositional patterns, namely those of humans and murids, it is reasonable to assume that they also exist (with slopes closer or identical to unity) among other mammalian genomes. In turn, this would lead to the conclusion that the GC levels of homologous coding sequences (or at least their order) are highly conserved in mammals. Because linear relationships exist between those GC levels and the GC levels of isochores in which they are embedded, it should also follow that the compositional pattern of the mammalian genome is highly conserved. These conclusions have been put to an experimental test and proven to be correct (G. Bernardi, D. Mouchiroud, and C. Gautier, submitted for publication).

Interestingly, the differences in karyotype and chromosomal banding exhibited by different mammals have obviously not affected their compositional pattern. This is easily understood if one considers that the number of chromosomal rearrangements that took place in divergent mammalian lines is very small (Dutrillaux 1979; Sawyer and Ozier 1986).

Under the circumstances described above, the data in Table 3 provide information on the distribution of particular gene classes in the isochores of the mammalian genome.

General Considerations

The results presented herein lead to some general considerations. First of all, differences in the compositional distributions of both isochores and coding sequences between murids and humans (as well as other mammals) are due to a directional fixation of mutations in genome compartments and in genes located at many different sites in the genomes.

Second, differences between the murid and the human genomes appear to correspond to essentially the same "compositional equilibrium," the higher GC level of GC-poor isochores and coding sequences from murids being compensated for by the lower GC level of GC-rich isochores and coding sequences, with almost no change in the average values. Moreover, differences between the human and the murid genomes are also characterized by a conservation of the order in the GC levels of homologous genes. Compositional differences had been observed previously between the genomes, as well as between homologous genes, of cold-blooded and warm-blooded vertebrates (Thiery et al. 1976; Perrin and Bernardi 1987). In this case, however, differences did not preserve the compositional equilibrium mentioned above, because there was a net gain of GC-rich isochores and GC-rich genes in warm-blooded vertebrates, and this did not conserve the order of GC levels of homologous genes. In other words, the compositional transition that occurred at the time of the divergence of murids, although not yet understood in its causes, is characterized by different features compared to that which occurred at the time of the divergence of warm-blooded vertebrates.

The third consideration concerns the conserva-

tion of the compositional distribution of homologous genes in mouse and rat in spite of the sequence divergence that exists between these genes. The molecular mechanisms allowing such compositional conservation against the accumulation of mutations will be discussed elsewhere.

All the above points are in keeping with the view (originally derived from the genome hypothesis of codon usage) that the genome is a unit of selection (Grantham and Gautier 1980; Grantham et al. 1980), with the existence of a "genome phenotype" (Bernardi and Bernardi 1986) and with the existence of selection pressures acting on the genome as a whole.

Finally, even if much more needs to be known in order to understand the functional aspects associated with the genome organization of vertebrates, the nonrandom genomic location of genes belonging in different functional classes and the amino acid changes due to the compositional constraints at the DNA level (see also Bernardi and Bernardi 1986; Gautier 1987a,b; Gouy 1987) strongly point to the existence of structure–function correlations.

Acknowledgments. We thank Giacomo Bernardi, Dominique Pontier, and Jean-Michel Gaillard for helpful comments.

References

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

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

Bilofsky HS, Burks C, Fickett JW, Goad WB, Lewitter FI, Rindone WP, Swindell LD, Tung CS (1986) The Genbank (R) genetic sequence database. Nucleic Acids Res 14:1–4

Cortadas J, Olofsson B, Meunier-Rotival M, Macaya G, Bernardi G (1979) The DNA components of the chicken genome. Eur J Biochem 99:176–186

Cuny G, Macaya G, Meunier-Rotival M, Bernardi G (1978) Some properties of the major component of the mouse genome. In: Boyer HW, Nicosia G (eds) Genetic engineering. Elsevier/North-Holland, Amsterdam, pp 109–115

Cuny G, Soriano P, Macaya G, Bernardi G (1981) The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity. Eur J Biochem 115:227–233

Dutrillaux B (1979) Chromosome evolution in primates: tentative phylogeny from *Microcebus murinus* (prosimian) to man. Hum Genet 48:251–314

Eisenberg JF (1981) The mammalian radiations. University of Chicago Press, Chicago

Gautier C (1987a) Changements d'usage du code génétique au cours de l'évolution. Le cas des mitochondries animales. C R Acad Sci 304:123–128

Gautier C (1987b) Analyse statistique et évolution des séquences d'acides nucléiques. Thèse Doctorat d'Etat, Lyon

Gouy M (1987) Origine et fonction de l'utilisation de la dégénérescence du code génétique chez *Escherichia coli.* Thèse Doctorat d'Etat, Lyon

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

Grantham R, Gautier C (1980) Genetic distances from mRNA sequences. Naturwissenschaften 67:93–94

Grantham R, Gautier C, Gouy M (1980) Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. Nucleic Acids Res 8:1893–1912

Lehman EL (1975) Non-parametric statistical methods based on ranks. Holden-Day, San Francisco, pp 32–40

Mouchiroud D (1986) Relation entre la composition en base de l'ADN no codant du gène et la composition en codon. C R Acad Sci Paris 303:743–748

Mouchiroud D, Gautier C (1988) High codon usage change in mammalian genes. Mol Biol Evol 5:192–194

Mouchiroud D, Fichant G, Bernardi G (1987) Compositional compartmentalization and gene composition in the genome of vertebrates. J Mol Evol (in press)

Perrin P, Bernardi G (1987) Directional fixation of mutations in vertebrate evolution. J Mol Evol 26:301–310

Salinas J, Zerial M, Filipski J, Bernardi G (1986) Gene distribution and nucleotide sequence organization in the mouse genome. Eur J Biochem 160:469–478

Sawyer JR, Ozier JC (1986) High resolution of mouse chromosomes: banding conservation between man and mouse. Science 232:1632–1635

Schildkraut CL, Marmur J, Doty P (1962) Determination of the base composition of deoxyribonucleic acid from its buoyant density in CsCl. J Mol Biol 4:430–443

Soriano P, Macaya G, Bernardi G (1981) The major components of the mouse and human genomes. 2. Reassociation kinetics. Eur J Biochem 115:235–239

Thiery JP, Macaya G, Bernardi G (1976) An analysis of eukaryotic genomes by density gradient centrifugation. J Mol Biol 108:219–235

Zerial M, Salinas J, Filipski J, Bernardi G (1986) Gene distribution and nucleotide sequence organization in the human genome. Eur J Biochem 160:479–485

Received November 17, 1987/Revised and accepted March 20, 1988