Journal
of Molecular
Evolution

Compositional Properties of Coding Sequences and Mammalian Phylogeny

Dominque Mouchiroud,[1] Giorgio Bernardi[2]

[1] Laboratoire de Biométrie, Génétique et Biologie des Populations, U.R.A. 243, Université Claude Bernard, 69600 Villeurbanne, France
[2] Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France

Abstract. The compositional distributions of large DNA fragments reflect those of the isochores that make up vertebrate genomes and can provide novel phylogenetic insights in the case of mammalian genomes (see Sabeur et al. 1993). This approach has been complemented here by an analysis of the compositional patterns of coding sequences and their codon positions (which also reflect the isochore pattern) and by a comparison of the base compositions of codon positions from homologous genes in a number of pairs of species. The results obtained using these two approaches support the existence of a general compositional pattern for mammalian genomes and of a distinct pattern for Myomorpha. The other two "special" patterns identified in a megachiropteran and in pangolin could not be tested here.

Key words: Isochores — DNA — Mammals — Rodents — Myomorpha —Murids

Introduction

The discovery that mammalian genomes consist of long (>300 kb, on the average), compositionally homogeneous DNA segments belonging to a small number of families covering a broad GC range opened new avenues not only for studies of genome organization (Bernardi 1989) but also for investigations of genome evolution (Bernardi 1993) and mammalian phylogenesis (Sabeur et al. 1993).

Mammalian genomes are characterized by compositional patterns that can be most conveniently studied at the level of large DNA fragments and of coding sequences. Both of them reflect the compositional patterns of isochores. A recent analysis of DNA fragment patterns from the genomes of 20 species belonging to 9 out of the 17 eutherian orders revealed (Sabeur et al. 1993) the existence of a general pattern, which was found in the genomes of species belonging to eight orders, and three other patterns. One of the latter, characterized by a narrower distribution of DNA fragments, was found in pangolin, a species belonging to the only genus of the order Pholidota. Since this order (together with the order Edentata) is supposed to have diverged early from all other eutherians (Novacek 1990), we suggested that the compositional pattern of pangolin is a primitive one. The other two patterns which were identified, that of a megachiropteran and that of Myomorpha, may be primitive or may have derived from the general pattern.

In this work, we have approached the general problem of the compositional patterns of mammalian genomes through the analysis of the compositional distributions of coding sequences and of their different codon positions and through comparison of homologous coding sequences. These two approaches could only be applied to some of the species studied at the DNA level—namely, to those species for which enough coding sequences are available in data banks.

Materials and Methods

Complete coding sequences were obtained for all genes available from GenBank, release 74 (December 92) with the ACNUC re-

Presented at the NATO Advanced Research Workshop on *Genome Organization and Evolution*, Spetsai, Greece, 16–22 September 1992
Correspondence to: D. Mouchiroud

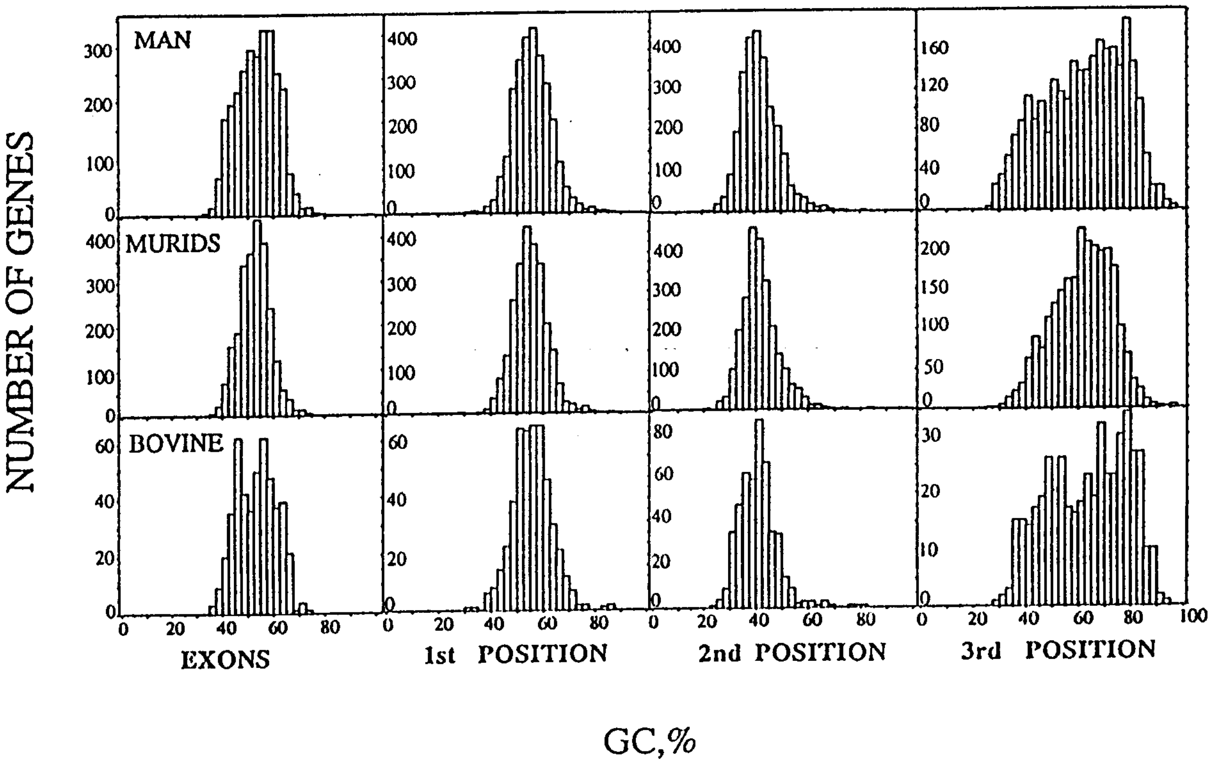

Fig. 1. Compositional distributions of coding sequences and codon positions of genes from man, murids (mouse + rat), and calf. In the case of murids, only one out of each pair of homologous genes (arbitrarily, the rat sequence) was taken into account. Each bar corresponds to a 2.5% GC interval. The number of genes analyzed is indicated in Table 1.

trieval system (Gouy et al. 1985). The search of homologous coding sequences was performed by using the BLAST software on protein sequences (Altschul et al. 1990). In the cases in which compositional distributions or comparisons concerned sequences pooled from different species, only one sequence from each homologous pair was used. Arbitrarily, rat sequences were preferred over mouse sequences and calf sequences over those from other artiodactyls. However, because of the existence of multigenic families, the pairwise association of homologous sequences from two species may be difficult and a strategy described elsewhere (Mouchiroud and Gautier 1990) was used to choose "orthologous" gene pairs. Homologous coding sequences were then aligned with the multiple alignment algorithm CLUSTAL (Higgins and Sharp 1989). The nucleotide alignment was obtained using the amino acid alignment and gaps were discarded. Computer analyses were performed by using ANALSEQ, a software for the statistical analysis of DNA sequences that was developed by the Laboratoire de Biométrie in Lyon. The relationships between GC levels for pairs of homologous genes were quantified by regression lines drawn using the least squares method in order to compare statistically the slopes. This would not have been possible if orthogonal regression (see D'Onofrio and Bernardi 1992) had been used.

Results

Histograms of compositional distributions of coding sequences and their codon positions are presented as follows. Figure 1 displays the data for the genes from the species most represented in data banks, namely man, and for genes from the mammals that are the next-best represented—namely murids (mouse and rat) and calf. Figure 2 presents the data for the sequences from rabbit and dog. Figures 1 and 2 show, therefore, data for genes from representative species belonging to five mammalian orders—namely primates, rodents, artiodactyls, lagomorphs, and carnivores. Table 1 summarizes the statistical analysis of all the data from Figs. 1 and 2, as well as some additional data concerning smaller samples from other species—hamster, guinea pig, cat, goat, sheep, and pig.

Comparisons of GC levels of exons and first-, second-, and third-codon positions from homologous genes are shown in Figs. 3–6. Comparisons concern the following pairs of species: man/calf (Fig. 3), man/rabbit (Fig. 4), man/rat (Fig. 5), and mouse/rat (Fig. 6). Figure 7 compares GC levels of third-codon positions for homologous genes from hamster and murids, guinea pig and murids, calf and pig, and man and dog.

Discussion

Compositional Patterns of Coding Sequences and Codon Positions

In this section, discussion will proceed from the more general to the more detailed points. An anal-

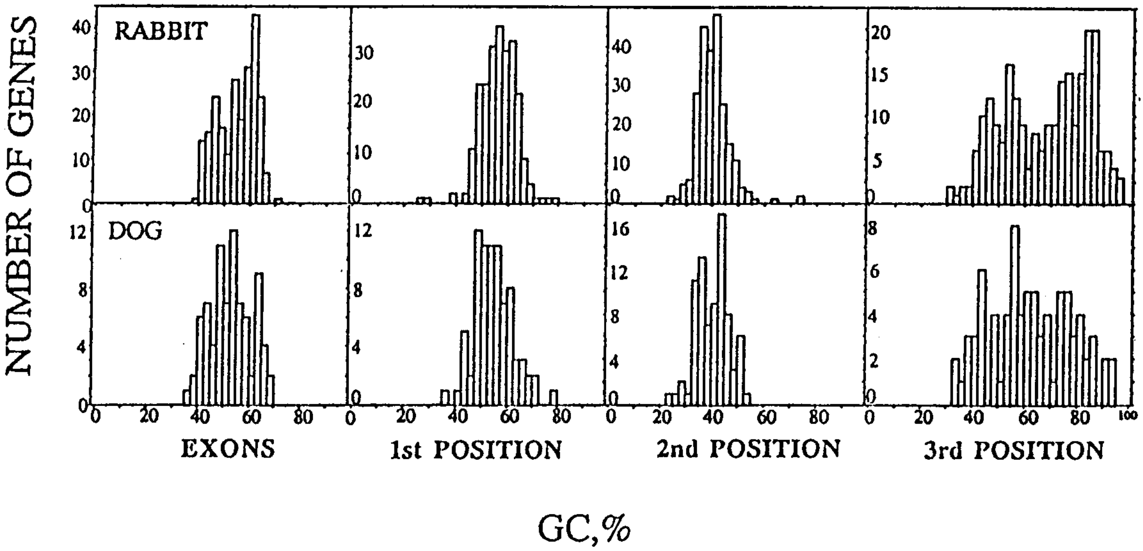

Fig. 2. Compositional distributions of coding sequences and codon positions of genes from rabbit and dog. For other indications see legend of Fig. 1.

Table 1. GC levels of coding sequences and codon positions from mammalian genes

Order and species	Number of genes	Coding sequence GC	Coding sequence Std	1st position GC	1st position Std	2nd position GC	2nd position Std	3rd position GC	3rd position Std
Primates									
Man	2772	53.6	7.7	56.0	7.1	42.2	7.2	62.5	15.5
Lagomorpha									
Rabbit	236	54.7	7.3	56.3	6.8	40.0	6.3	67.8	16.3
Rodents									
Mouse	1159	53.0	6.4	55.0	6.8	42.6	7.7	61.4	11.9
Rat	1350	52.6	5.9	55.0	6.5	40.8	6.5	62.0	10.7
Murids	2479	52.8	6.0	55.0	6.6	41.6	7.0	61.7	11.3
Hamster	61	52.0	5.8	52.6	6.9	40.1	8.2	63.1	11.5
Guinea pig	20	52.5	6.1	53.2	4.3	39.7	5.5	64.4	14.2
Carnivores									
Dog	80	52.8	8.1	54.9	7.5	40.5	6.4	62.9	15.8
Cat	14	53.8	7.5	56.4	7.1	42.4	5.5	62.6	15.6
Carnivores	84	53.0	7.9	55.1	7.5	40.8	6.3	62.8	15.7
Artiodactyls									
Calf	470	53.5	7.6	55.7	7.4	41.0	7.5	63.5	15.2
Goat	17	55.0	6.7	57.9	5.8	36.6	4.0	70.3	14.4
Sheep	87	55.1	6.9	53.9	8.0	43.6	12.0	67.6	11.8
Pig	152	54.7	7.2	55.7	7.5	41.5	8.2	66.7	13.5
Artiodactyls	693	53.9	7.4	55.7	7.3	41.3	8.2	64.7	14.5

ysis of the data from Table 1 shows (1) that GC levels increase from second to first to third position, the difference between second and third position being about 20%, and (2) that compositional heterogeneity (as indicated by standard deviations) is stronger, by a factor of about two, in third positions, compared to the other two positions. The heterogeneity of second-codon-position GC was comparable to that of first-codon-position GC.

The most significant results of Table 1 concern, however, the fact that murids and hamster from the rodent infraorder Myomorpha exhibit a lower variance in exons, and in first- and, especially, third-codon positions, compared to other mammals (including guinea pig, which belongs to the rodent infraorder Caviomorpha), whereas variances are similar for first- and second-codon positions. Expectedly, this difference is also very evident in the histograms of Figs. 1 and 2. The apparent deviations exhibited by GC values of goat, sheep, and pig compared to other mammals and even to another artiodactyl, calf, will be commented upon in the

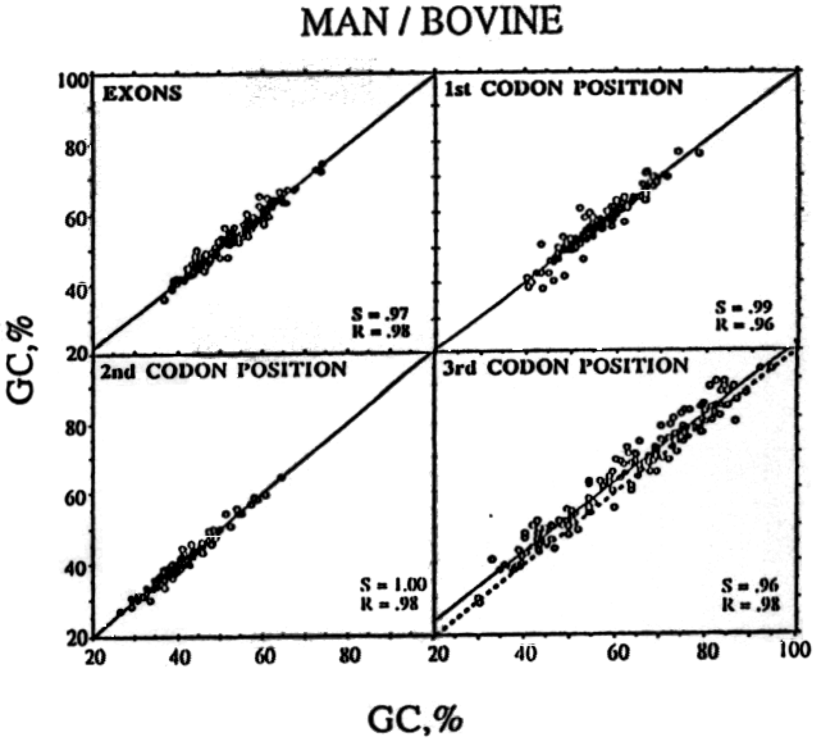

Fig. 3. Compositional correlations between pairs of homologous genes from man (abscissa) and calf (ordinate). Correlations are shown for coding sequences and codon positions. Slopes and correlation coefficients are indicated. The number of homologous gene compared was 141. In the case of third-codon positions, the *broken line* corresponds to the unity slope.

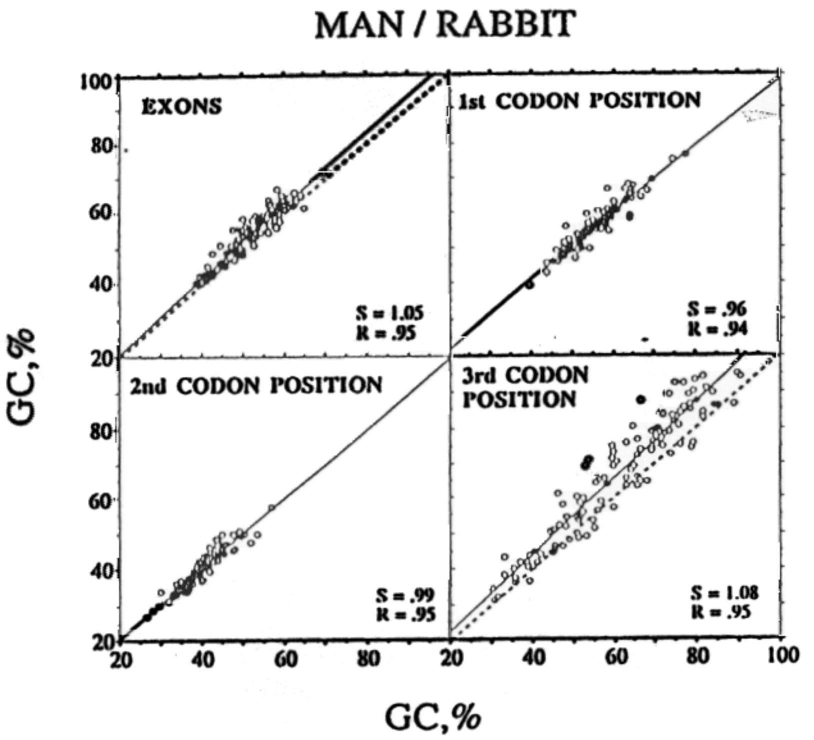

Fig. 4. Compositional correlations between pairs of homologous genes from man (abscissa) and rabbit (ordinate). For other indications see legend of Fig. 3. The number of homologous genes compared was 117.

next section. Figures 1 and 2 appear also to indicate a difference between rabbit, man, calf and dog, the latter showing a smaller percentage of GC-rich sequences and codon positions.

These results considerably extend previous results obtained on smaller-sequence samples and on a smaller number of species (Mouchiroud et al. 1987, 1988; Bernardi et al. 1988; D'Onofrio and Bernardi 1992). As expected, they are in agreement with results obtained on the compositional patterns of large DNA fragments (Sabeur et al. 1993) in that they fit into a general pattern and a murid (in fact, a myomorph) pattern. The differences noted in this section will be commented further in the following one.

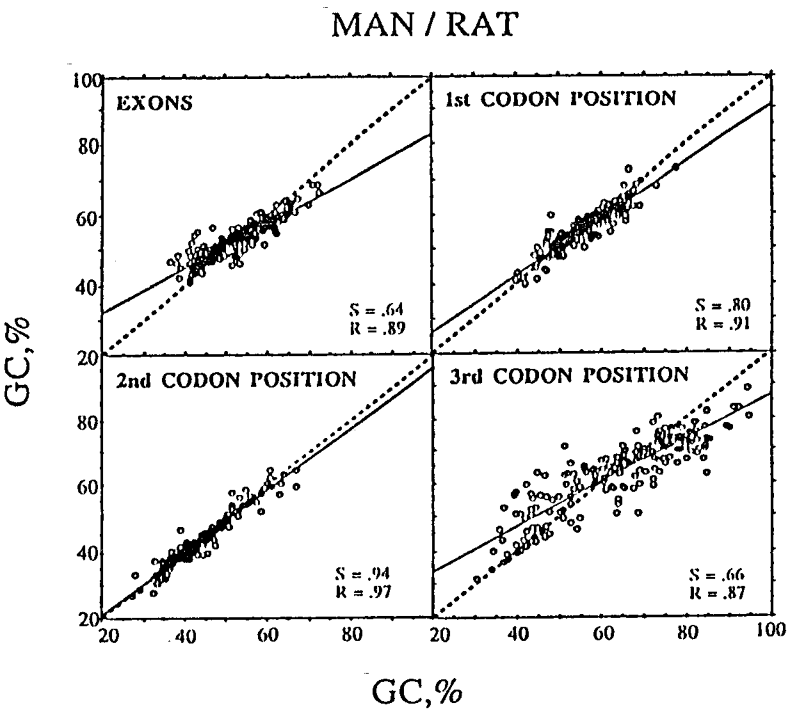

Fig. 5. Compositional correlations between pairs of homologous sequences from man (abscissa) and rat (ordinate). For other indications see legend of Fig. 3. The number of homologous genes compared was 210.

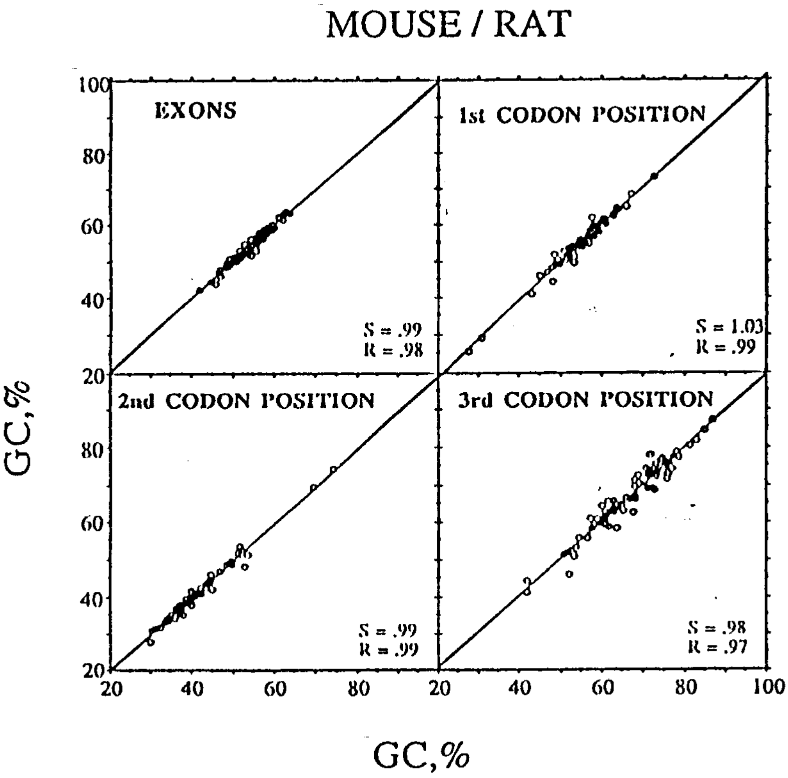

Fig. 6. Compositional correlations between pairs of homologous sequences from mouse and rat. For other indications see legend of Fig. 3. The number of homologous genes compared was 78.

A Comparison of GC Levels of Coding Sequences and Codon Positions from Homologous Mammalian Genes

The compositional comparison of coding sequences and codon positions from man and calf (Fig. 3) shows extremely high correlation coefficients (0.98, except for first-codon position, in which case the value was 0.96) and slopes ranging from 0.96 for third-codon positions to 1.00 for second positions. Because of the correlation between third-codon positions and the isochores containing the corresponding genes (Bernardi et al. 1985; Bernardi and Bernardi 1986; Mouchiroud et al. 1991), this implies a

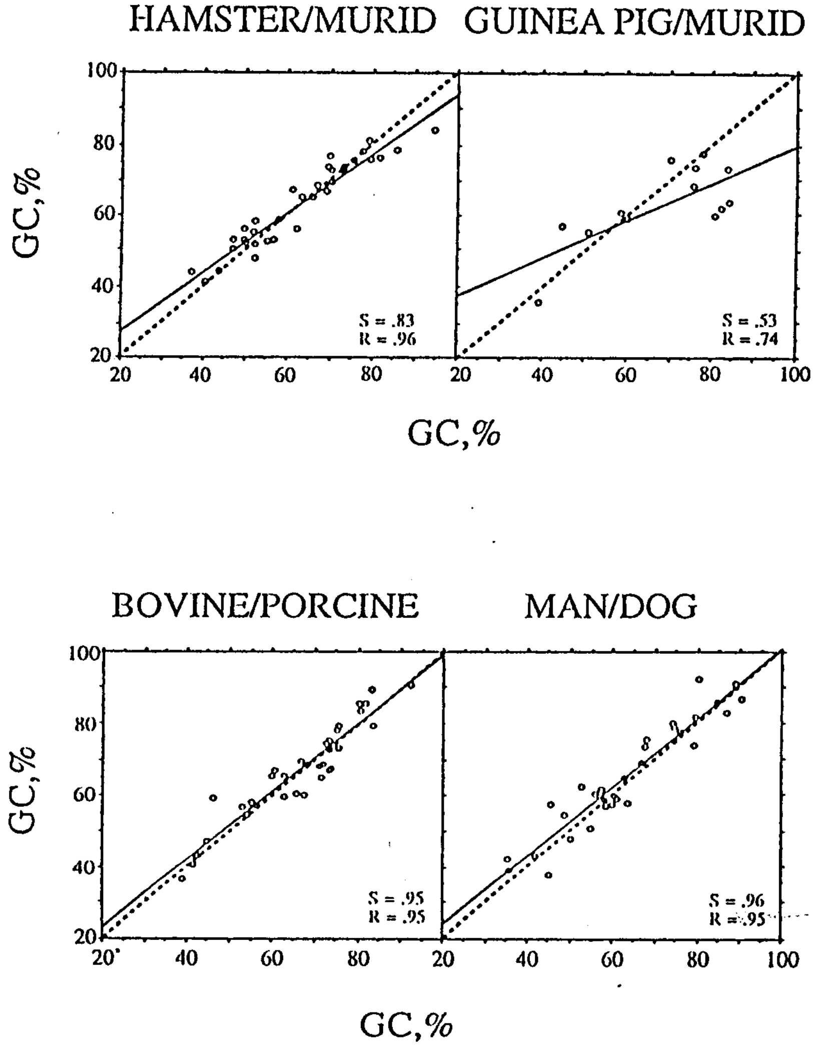

Fig. 7. Compositional correlations between third-codon positions of pairs of homologous sequences of murids and hamster, murids and guinea pig, pig and calf, and dog and man. For other indications see legend of Fig. 3. The numbers of homologous genes compared were 34, 13, 34, and 32, respectively.

very high level of similarity in the isochore patterns of human and bovine genomes. In particular, this indicates that the amounts of nonsatellite DNA above 1.710 g/cm^3 in modal buoyant density in calf and in man, studied at the DNA level by Sabeur et al. (1993), must be very close. It should be noted that a very slight systematic upward shift of bovine third-codon positions compared to human ones seems to suggest very slightly higher GC values for the former genome compared to the latter (in agreement with a slightly higher average GC in third-codon positions of 63.5 ± 15.2 vs 62.5 ± 15.5 for man), but this shift is not statistically significant at a threshold of 5% ($P > 0.2$).

The second comparison, between homologous genes from two other genomes exhibiting the general compositional pattern, man and rabbit (Fig. 4), shows small, but significant differences. The correlation coefficients are slightly lower (0.94–0.95), and the slope for third-codon positions deviate more from unity, 1.08, respectively, than in the case of man/calf. This difference (in agreement with a higher average value for third-codon position GC; see Table 1) is significant (at the 5% threshold; $P < 10^{-6}$) and is essentially due to systematically slightly higher values for GC-rich third-codon positions of rabbit. In turn, this suggests that GC-rich DNA reaches slightly higher GC values in rabbit compared to man. Since the estimates (Sabeur et al. 1993) of nonsatellite DNA having a modal buoyant density higher than 1.710 g/cm^3 are very similar in man and rabbit, an explanation for this apparent discrepancy is that some GC-rich DNA of rabbit may be hidden by the very abundant (9% of total DNA) satellite having a ρ_o value of 1.717 g/cm^3 (see Sabeur et al. 1993).

In contrast with the above results, the comparison of homologous genes from man and rat (Fig. 5) exhibited lower slopes ranging from 0.64 for exons to 0.94 for second-codon positions and correlation coefficients ranging from 0.87 to 0.97. These high correlation coefficients indicate a large extent of conservation in the ranking order of GC levels of third-codon positions from homologous genes in spite of the systematic differences affecting both high and low GC levels.

Mouse/rat comparisons (Fig. 6) showed slopes of 0.98 to 1.03 (the highest slope being found for first-codon position) and correlation coefficients very close to one, again confirming an extremely high compositional similarity of these coding sequences and codon positions and of the corresponding isochores.

A comparison of third-codon positions from murids and hamster, on the one hand, and murids and guinea pig, on the other, showed (Fig. 7) a slight difference in the first case and a large difference in the second. Both differences showed a trend similar to that found in the comparison of human third-codon positions with the homologous values for rat. The hamster/murid difference is in agreement with the slightly higher amount of GC-rich DNA in hamster (Sabeur et al. 1993), whereas that found in the guinea pig/murid comparison is the difference expected for a general pattern/myomorph pattern comparison. The slope is slightly different from the man/rat comparison (0.53 vs 0.66) (as is the correlation coefficient), but this can be understood as due to the much smaller number of points in the guinea pig/murid comparison.

The last two comparisons of third-codon positions of Fig. 7 are especially instructive. The first one, the bovine/porcine comparison, is of interest in two respects. Indeed, the extremely close GC levels indicate that a bias in the gene sample was in fact responsible for the differences of Table 1, and that two species from two different families of artiodactyls (Bovidae and Suidae) are characterized by the same compositional pattern. The lower correlation coefficient in this intraordinal comparison relative to the interordinal human/bovine comparison (0.95 instead of 0.98) is certainly due to the smaller number of sequences used in the first case. The man/dog comparison is also very interesting in that the identical values found indicate that the same general compositional pattern is present in primates and carnivores. This stresses the fact that the peculiar features of the distributions of GC levels of coding sequences and codon positions from dog genes (Fig. 1) obviously are the result of a biased gene sample.

Conclusions

The two approaches used in the present work complement the compositional analysis of large DNA fragments (Sabeur et al. 1993) and provide very stringent genome comparisons.

The compositional comparisons of codon positions from homologous genes show an extreme similarity in the two cases in which the largest number of sequences are available—namely man/calf and mouse/rat. Both of them strongly reinforce the conclusions drawn in the preceding paper (Sabeur et al. 1993) and the similar ones that one could draw from Figs. 1 and 2 of this paper.

It is of interest to note that the coding sequence comparisons done here are informative for a non-negligible part of the genomes under consideration. Because of the linear compositional correlations between coding sequences and their codon positions, on the one hand, and of the isochores containing them on the other, the sequence comparisons practically concern as many isochores as the sequences studied. Assuming an isochore size of 300 kb (which is, in all likelihood, a low estimate), when 200 homologous sequences are compared, as in the case of man and rat, 2% of these genomes are explored compositionally. It should be noted here that there are good reasons to believe that the data on the coding sequences explored can be extrapolated to all other coding sequences and to all the isochores containing them.

The sequence comparison approach is by far the most telling relative to comparisons of histograms concerning both homologous and non-homologous genes (the latter being predominant). This can be judged from the small yet significant differences detected here in the compositional patterns of man and rabbit or mouse and hamster. On the other hand, they provide evidence for close similarities of compositional patterns for genes from species belonging to different families in the same order (calf and pig) or to different orders (calf and man, man and dog).

Although already discussed in previous work (Mouchiroud et al. 1987; 1988; Bernardi et al. 1988; Mouchiroud and Gautier 1990), the man/murid comparison deserves some additional comments concerning the relationship between the general pattern and the myomorph pattern. It should be noted that the differences found are also seen in intronic sequences and in the 5′ and 3′ untranslated sequences (data not shown). As indicated in the preceding paper (Sabeur et al. 1993), at present no absolutely clear-cut conclusion can be drawn as to which pattern is primitive and which is derived. If the general pattern is primitive, the myomorph pattern could be derived from it by a release of the compositional constraints operating on the ends of the compositional distribution of coding sequences and third-codon positions. This would, indeed, push such ends toward 50% GC—namely, it would lower the high values and rise the low values. If the myomorph pattern is primitive, an increase in compositional constraints should have taken place between the myomorph and the general pattern, leading to an increased heterogeneity in the latter case. Interestingly, in such a case, the myomorph pattern would be intermediate between the low-heterogeneity pattern of cold-blooded vertebrates (Bernardi and Bernardi 1990a,b; 1991) and the high-heterogeneity pattern of most mammals, although very much closer to the latter. If the myomorph pattern is a derived one, the closer similarity of the hamster pattern (compared to the patterns of rat and mouse) to the general pattern would suggest that the former diverged from it earlier than the latter.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

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

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

Bernardi G, Bernardi G (1990a) Compositional patterns in the nuclear genomes of cold-blooded vertebrates. J Mol Evol 31: 265–281

Bernardi G, Bernardi G (1990b) Compositional transitions in the nuclear genomes of cold-blooded vertebrates. J Mol Evol 31: 282–293

Bernardi G, Bernardi G (1991) Compositional properties of nuclear genes from cold-blooded vertebrates. J Mol Evol 33:57–67

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

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

Gouy M, Gautier C, Attimoneli M, Lanave C, di Paola G (1985) ACNUC—portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. CABIOS 1:167–172

Higgins DG, Sharp PM (1989) Fast and sensitive multiple sequence alignments on a microcomputer. Comput Appl Biosci 5:151–153

Mouchiroud D, Fichant G, Bernardi G (1987) Compositional compartmentalization and gene composition in the genome of vertebrates. J Mol Evol 26:198–204

Mouchiroud D, Gautier C, Bernardi G (1988) The compositional distribution of coding sequences and DNA molecules in humans and murids. J Mol Evol 27:311–320

Mouchiroud D, Gautier C (1990) Codon usage changes and sequence dissimilarity between human and rat. J Mol Evol 31: 81–91

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

Novacek MJ (1990) Morphology, paleontology and the higher clades of mammals. In: Genoways HH (ed) Current Mammalogy vol 2, Plenum, New York, pp 507–543

Sabeur G, Macaya G, Kadi F, Bernardi G (1993) The isochore patterns of mammalian genomes and their phylogenetic implications. J Mol Evol (in press)