J Mol Evol (1991) 32:493–503

Journal of
Molecular Evolution

The Compositional Properties of Human Genes

Brahim Aïssani,[1] Giuseppe D'Onofrio,[1]* Dominique Mouchiroud,[2] Katheleen Gardiner,[3]
Christian Gautier,[2] and Giorgio Bernardi[1]

[1] Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France
[2] Laboratoire de Biométrie, Génétique et Biologie des Populations, U.R.A. 243, Université Claude Bernard, 69600 Villeurbanne, France
[3] Eleanor Roosevelt Institute for Cancer Research, 1899 Gaylord Street, Denver, CO 80202, USA

Summary. The present work represents the first attempt to study in greater detail previously proposed compositional correlations in genomes, based on a body of additional data relating to gene localizations as well as to extended flanking sequences extracted from gene banks. We have investigated the correlations that exist between (1) the GC levels of exons of human genes, and (2) the GC levels of either intergenic sequences or introns associated with the genes under consideration. In both cases, linear relationships with slopes close to unity were found. The similarity of the linear relationships indicates similar GC levels in intergenic sequences and introns located in the same isochores. Moreover, both intergenic sequences and introns showed GC levels 5–10% lower than the corresponding exons. The above findings considerably strengthen the previously drawn conclusion that coding and noncoding sequences (both inter- and intragenic) from the same isochores of the human genome are compositionally correlated. In addition, we find linear correlations between the GC levels of codon positions and of the intergenic sequences or introns associated with the corresponding genes, as well as among the GC levels of codon positions of genes.

Key words: Human genome — Gene localization — Isochores — Coding sequences — Introns — Codon positions

* On leave from Stazione Zoologica, Villa Comunale, 80121 Naples, Italy
Offprint requests to: G. Bernardi

Introduction

Density gradient centrifugation in the presence of sequence-specific DNA ligands (Corneo et al. 1968; Cortadas et al. 1977; Macaya et al. 1978) led to the fractionation of high molecular weight (30–100-kb) nuclear DNA fragments from vertebrates according to their base composition (Filipski et al. 1973; Macaya et al. 1976; Thiery et al. 1976; Cuny et al. 1981). The DNA fragments so fractionated derive (by the mechanical and enzymatic degradations that occur during DNA preparation) from much longer (>300-kb) segments that are remarkably homogeneous in base composition and belong to a number of families characterized by different GC levels (Macaya et al. 1976). These segments were later termed isochores for equal regions (Cuny et al. 1981).

Compositional DNA fractions were used to demonstrate (1) that the GC-poor beta-globin genes are localized in GC-poor isochores in human, rabbit, and mouse, whereas the GC-rich alpha-globin genes are localized in the GC-rich isochores of the same species, and that similar differences in localization exist for the GC-rich globin genes and for the GC-poor ovalbumin genes of chicken (Bernardi 1979, 1984) [similar findings were reported by Ikemura (1985) for GC levels of third codon positions and flanking sequences of the same genes]; (2) that within each isochore family, interspersed repeated sequences and unique sequences do not differ in GC levels (Soriano et al. 1981); (3) that the interspersed, mobile, repeated sequence families, LINES and SINES (Singer 1982), match in GC level the long DNA segments in which they are located (Meunier-

Rotival et al. 1982; Soriano et al. 1983); (4) that integrated viral sequences also match the genomic environments in which they are inserted (Kettman et al. 1979; Zerial et al. 1986a; Salinas et al. 1987); and (5) that the GC levels of genes, exons, third codon positions, and introns from the nuclear genomes of vertebrates are linearly correlated with the GC levels of the large DNA fragments that contain them (Bernardi et al. 1985; Bernardi 1989). These findings have important functional and evolutionary implications that were discussed elsewhere (Bernardi and Bernardi 1985, 1986, 1990a,b; Mouchiroud et al. 1987, 1988; Perrin and Bernardi 1987; Bernardi et al. 1988).

In the original work (Bernardi et al. 1985), the compositional correlations referred to above were established on the basis of results from a total of 24 loci (defined here as genes or gene clusters) from five vertebrates, *Xenopus* (1 locus), chicken (5 loci), mouse (7 loci), rabbit (2 loci), and human (9 loci). The fact that only one locus was explored for a single cold-blooded vertebrate left open, however, the question of whether a common relationship held for all vertebrates. Moreover, as far as warm-blooded vertebrates are concerned, the small number of loci and species analyzed hindered the precise definition of the parameters (slope, intercept, and correlation coefficient) for the relationships under consideration. Obviously, this definition was totally impossible for individual genomes.

The compositional correlations between genes (exons, codon positions, and introns) and isochores were studied here using a set of 21 human loci that had been localized in DNA fractions characterized by different GC levels, as well as a second set of 32 loci (4 of which were also represented in the first set) that were found in the long (>10 kb) DNA segments available in sequence banks.

The primary reason for this work was simply to understand better the compositional correlations linking the genes with their genomic environments (namely with the isochores in which they are located), exons with introns, and codon positions among themselves.

The second reason was that the existence of compositional correlations between coding sequences and isochores (as represented by large DNA fragments) permits, in principle, assignment of an isochore location to genes of known coding sequence that were not experimentally localized in fractionated DNA fragments. This is an important point, because the distribution of genes in the genome is known to be highly nonuniform (Bernardi et al. 1985; Bernardi 1989) (genes are concentrated in the GC-richest isochores of the human genome) but needs to be defined better. Moreover, the isochore distribution of genes is correlated with their chromosomal distribution (Bernardi et al. 1985; Aota and Ikemura 1986; Bernardi 1989; Gardiner et al. 1990; and unpublished) as, in general, GC-poor genes are located in Giemsa-positive bands, GC-rich genes in Giemsa-negative bands, and the GC-richest genes in T-bands [the most denaturation-resistant bands that are mainly located in the telomeres of certain chromosome arms (Dutrillaux 1973; Ambros and Sumner 1987)]. It is obvious that an assessment of gene distribution in isochores and chromosomal (Giemsa-positive, Giemsa-negative, and telomeric) bands needs to be based on compositional correlations that are as precise as possible.

Needless to say, a prerequisite for assigning an isochore (and chromosomal band) localization to human genes is that the sample of localized genes used as a reference is representative of human genes. This can be demonstrated by comparing gene properties that are independent of the gene location in the genome, such as the correlations between GC levels of third and first + second positions, and between GC levels of exons and introns. These correlations were determined here in order to compare them with those found for the 1400 or so human genes available in gene banks (D'Onofrio et al. 1991), namely for the largest set of genes from a single genome that have known coding sequences.

Materials and Methods

A set of 21 human loci (defined here as genes or gene clusters; see Tables 1 and 2) were localized in DNA fractions or DNA components. Localizations were carried out by hybridizing appropriate probes on human DNA that had been first fractionated by preparative ultracentrifugation in a Cs_2SO_4 density gradient in the presence of BAMD [bis-(acetato mercurimethyl) dioxane], a sequence-specific DNA ligand, and then digested with either EcoRI or HindIII (see Bernardi et al. 1985 and papers quoted therein). This procedure allows estimation of the GC levels of DNA segments that are about twice the size of the DNA fragments that were fractionated (see Bernardi 1989), namely about 70 kb for alpha-globin, c-mos, c-myc, c-Ha-ras1, proopiomelanocortin, and c-sis genes, and 200–300 kb for all other genes. In early work, DNA fractions from the preparative ultracentrifugation were recentrifuged once or twice, and fractions and subfractions were pooled according to their modal buoyant density to provide the so-called major DNA components (Cuny et al. 1981), namely families of DNA fragments characterized by very similar modal buoyant densities in CsCl. Gene localization data are derived from Bernardi et al. (1985), Zerial et al. (1986b), Gardiner et al. (1990), and from as yet unpublished work on X chromosome genes carried out in collaboration with R. Little and D. Schlessinger (St. Louis, MO, USA).

Another set of 32 genes or gene clusters (four of which, alpha-globin, beta-globin, coagulation factor IX, and HPRT, were also present in the first set) were found in all the sequences >10 kb that were available (see Table 1). Only one gene, the dystrophin gene was not taken into consideration here because it is an exceptional case in which the gene size is about 2.3 Mb, larger than many bacterial genomes, whereas the coding sequence is only 11 kb. This case was studied in a separate investigation (T. Bettecken, B. Aïssani, C. Müller, and G. Bernardi, unpublished).

Table 1. List and properties of human genes and gene clusters investigated

Genes			GC %					Size (bp)			
No.	Name	Mnemonic	Exon	I + II	III	Intron	Fraction/ Sequence	Exon	Intron		Sequence
	Beta-globin cluster	HBB	54.0	48.3	65.5	38.5	40.3	2220	4969		
2	Coagulation factor IX	FIXG	41.3	44.3	35.3	38.9	39.7	1386	29,954		
3	HPRT	HPRTB	41.1	41.8	39.7	39.0	41.5	657	39,168		
4	SOD	SODG1	49.7	51.9	45.1	40.1	42.0	465	878		
5	APP	PRA401	52.0	48.8	58.2	36.2	42.0	2088	3104		
6	IFN-alpha receptor	IFNRA	36.8	39.8	30.8		42.0	1671			
7	GART	Schild et al. 1990	44.8	47.1	40.1		43.0	906			
8	c-mos	CMOS	61.0	55.0	72.9		43.7	1041			
9	Vimentine	VIM	56.4	49.0	71.1	41.8	42.0	1401	1212		
10	ETS2	ETS2A	52.8	46.1	66.4		46.0	1410			
11	ERG2	ERG2	54.9	49.5	65.9		46.0	1389			
12	MX cluster	Aebi et al. 1989	50.6	44.0	63.7		46.5	4131			
13	Collagen cluster	COLTHA + B	69.4	73.1	60.5		47.0	2013			
14	c-myc	MYCB1	58.8	50.0	76.4	50.7	46.7	567	669		
15	Breast cancer induced protein	PS2	57.6	52.9	67.1	54.6	48.0	255	856		
16	G6PD	G6PD	59.4	46.0	86.2		52.4	1089			
17	c-Ha-ras1	RASH	59.1	48.1	81.1	69.0	53.7	570	1114		
18	Alpha-globin cluster	HBA4 + 1	64.9	51.9	90.7	73.3	53.7	1287	1598		
19	Proopiomelanocortin	POMC	67.7	56.9	89.2	48.0	53.7	804	6592		
20	c-sis	CSIST	62.3	54.5	77.7	59.5	53.7	726	287		
21	MCF2	MCF2PO	41.6	38.0	48.8		37.7	1704			
22	HPRT	HPRTB	41.1	41.8	39.7	39.0	40.3	657	39,168	(69)	56,536
23	Fibrinogen gamma	FBRG	41.2	42.2	39.7	35.9	37.1	1314	6956	(66)	10,564
24	Gamma crystallin B-C	CRYGBC	58.1	47.6	79.2	42.9	44.8	1053	3520	(15)	22,775
25	Protein C	PRCA	61.9	51.2	83.2	57.6	56.9	1386	7589	(65)	11,725
26	Adenosine deaminase	ADAG	57.2	49.4	72.8	53.0	53.9	1092	30,542	(83)	36,741
27	Growth hormone	GHCSA	56.3	46.1	76.3	56.7	49.2	3216	4186	(6)	66,495
28	Beta-globin cluster	HBB	54.0	48.3	65.5	38.5	39.5	2220	4969	(7)	73,326
29	Alpha-1-antitrypsin	A1ATP	52.0	43.6	68.7	50.2	51.4	1257	8839	(72)	12,222
30	Alpha-fetoprotein	AFP	42.0	43.9	38.2	33.8	41.7	1830	17,460	(79)	22,166
31	Serum albumin	ALBGC	42.9	44.9	38.8	34.3	35.7	1830	16,349	(86)	19,002
32	Alpha-globin cluster	HBA4	64.8	52.8	88.8	73.0	58.8	858	520	(4)	12,847
33	Placental tissue factor	TFPB	47.6	44.9	53.0	43.5	44.7	888	10,282	(74)	13,865
34	Haptoglobin	HPARS1	49.4	47.1	54.0	46.6	46.2	1221	6920	(60)	11,551
35	Aldolase B	ALDB1	53.5	50.8	58.9	41.2	41.5	1095	8626	(84)	10,239
36	Cytochrome P450IIE1	CYPIIE	52.5	44.9	67.6	52.4	50.3	1482	9742	(66)	14,776
37	Tissue plasminogen activator	TPA	58.3	51.0	72.8	48.3	48.8	1689	30,068	(82)	36,594
38	Prothrombin	THB	58.0	50.5	73.0	49.7	50.6	1869	18,244	(88)	20,801
39	Cytosolic adenylate kinase	AK1	57.5	49.6	73.2	60.6	60.0	570	8984	(73)	12,229
40	Plasminogen activator inhibitor-1	PAIA	57.0	47.8	75.4	50.7	50.2	1209	1147	(7)	15,867
41	Pulmonary surfactant protein	SPBAA	62.5	54.5	78.5	57.1	57.3	1146	7334	(70)	10,476
42	Thymidine kinase	TKRA	60.4	51.2	78.7	52.0	53.3	705	11,532	(85)	13,500
43	Coagulation factor VII	CFVII	62.5	53.5	80.5	61.7	60.7	1401	9737	(76)	12,850
44	Na, K-ATPase	ATP1A2	57.1	48.6	73.8	50.8	51.4	3063	19,403	(73)	26,668
45	ATP synthetase beta subunit	ATPSYB	51.3	52.5	48.7	43.6	45.8	1590	6054	(59)	10,186
46	c-fes/fps protooncogene	FESFPS	61.9	52.7	80.2	57.6	59.0	2469	8529	(69)	12,263
47	Coagulation factor IX	FIXG	41.3	44.3	35.3	38.9	39.0	1386	29,954	(79)	38,059
48	HLA-DP-beta 1 & alpha 1	HLADPB	58.3	50.7	73.6	48.0	46.0	876	9710	(66)	14,782
49	HLA-SB-alpha	HLASBA	56.4	48.2	71.1	40.3	38.9	882	7427	(51)	14,646
50	Alpha-D-galactosidase	GLA	48.6	47.2	51.4	47.4	44.1	1290	9799	(79)	12,436
51	Glucagon	GLUCG2	46.2	44.7	49.2	33.4	34.3	543	8267	(82)	10,050
52	Interleukin-1-alpha	IL1AG	43.9	39.9	51.8	39.3	39.4	816	8178	(68)	11,970
53	Int-2 Protooncogene	INT2	67.1	58.8	83.8	60.1	60.3	720	7941	(68)	11,608

Entries 1–21 refer to genes localized in DNA fractions; entries 22–53 to genes present in sequences longer than 10 kb, as available in GenBank (or EMBL for point 21). Mnemonics are given for each gene (except for entries 7 and 12, which are not yet available). GC% values of exons (weight average values of all exons of any given gene; see Materials and Methods), first + second codon positions, introns and total sequences are given. Sizes of exons, introns, and total sequences are also indicated. The relative amounts of introns in the sequences are given in parentheses

Table 2. List and properties of human clustered genes

Gene clusters		GC%[a]				Size (bp)[b]	
Name	Mnemonic	Exon	I + II		Intron	Exon	Intron
Beta globin							
Beta	HBB	56.3	51.3	66.2	32.9	444	978
Delta	HBB	53.6	49.7	63.5	34.6	444	1024
Gamma A	HBB	53.6	47.0	66.2	44.5	444	986
Gamma G	HBB	53.6	47.0	66.2	44.5	444	1006
Epsilon	HBB	52.9	46.3	65.5	36.0	444	975
		54.0	**48.3**	**65.5**	**38.5**	**2220**	**4969**
Alpha globin							
Alpha-1	HBA4	64.8	52.8	88.8	73.3	429	263
Alpha-2	HBA4	64.8	52.8	88.8	72.6	429	257
Zeta	HBA1	65.0	50.0	94.4	73.5	429	1078
		64.9	**51.9**	**90.7**	**73.1**	**1287**	**1598**
Interferon-induced protein							
MXA	Aebi et al. 1989	51.1	44.6	64.1		1986	
MXB	Aebi et al. 1989	50.1	43.4	63.4		2145	
		50.6	**44.0**	**63.7**		**4131**	
Alpha collagen							
Alpha-1 (VI)	COLTHA	68.8	72.9	60.7		1008	
Alpha-2 (VI)	COLTHB	70.0	73.3	60.3		1005	
		69.4	**73.1**	**60.5**		**2013**	
Growth hormone							
GH-1	GHCSA	56.2	46.2	76.2	56.6	654	814
CS-5	GHCSA	56.5	46.2	77.0	56.3	600	883
CS-1	GHCSA	56.2	46.0	76.6	56.7	654	835
CH-2	GHCSA	56.2	46.0	76.6	56.4	654	824
CS-2	GHCSA	55.8	46.1	75.2	57.4	654	830
		56.3	**46.1**	**76.3**	**56.7**	**3216**	**4186**
Gamma crystallin							
Gamma-B	CRYGBC	56.3	46.4	76.1	46.8	528	2459
Gamma-C	GRYGBC	60.0	48.8	82.3	33.8	525	1061
		58.1	**47.6**	**79.2**	**42.9**	**1053**	**3520**

[a] Bold values are weighted averages
[b] Bold values are total base pairs

Weight-average GC levels of exons, third codon positions, and introns of any given gene or gene cluster were used. Exons (but not introns) preceding the initiation codon AUG were disregarded. All available 5′ flanking sequences >1 kb preceding the first exons were also studied (see Table 3). Gene sequences were obtained from Release 63 (March 15, 1990) of GenBank, and from Release 23 of EMBL for one gene (MCF2, point 21 of Table 1) using the ACNUC retrieval system (Gouy et al. 1985).

Results

The compositional correlations of human exons and third codon positions with intergenic sequences and introns, as well as the compositional correlations among codon positions of human genes were investigated as shown in Figs. 1–5.

Figure 1A displays a plot of GC levels of exons from a set of human genes or gene clusters (see Table 1, entries 1–21, for gene numbering and properties) against the GC levels of DNA fractions in which the genes or gene clusters were localized by hybridization with appropriate probes. Values for exons from interferon receptor (point 6) and alpha (VI) collagen genes (point 13) are shown in the plot but were disregarded in calculating the correlation coefficient and the slope for reasons given in the Discussion. The least-square line through the points exhibits a slope of 1.19 and a correlation coefficient of 0.80. This line is about 5–10% above the line of unity slope passing through the origin. (The diagonal line corresponds to identical GC values for the exons and the DNA fractions in which the exons are located.) Because clustered genes exhibited coding sequences that were very similar in composition (see Table 2), as expected (Bernardi et al. 1985; Bernardi 1989), average GC values for all exons in each cluster were used to plot the data in Fig. 1A. This was also done in the other plots presented.

Figure 1B shows a plot of GC levels of exons from another set of human genes (see Table 1, en-

Table 3. List and properties of 5′-flanking sequences from human genes

No.	Name	Mnemonic	Intron GC%	Flanking GC%	
1	Gamma-interferon	IFNINI	36.4	37.5	2133
2	Serum albumin	ALBGC	34.4	32.5	1736
3	Proline-rich protein	PRPH1	40.5	39.7	1560
4	HPRT	HPRTB	43.0	52.3	1708
5	Fibrinogen gamma	FBRG	35.9	39.5	1748
6	Interleukin 2	IL2A	31.8	34.7	1362
7	Nuclear antigen	PCNA	41.5	47.7	1267
8	SomatostatinI	SOMI	44.2	41.6	1125
9	Cytochrome P450IIE1	CYPIIE	52.5	43.5	2788
10	Steroid 17 alpha-hydroxylase	P45C17	52.1	45.5	1778
11	ATP/ADP translocator	ANT1	47.2	50.0	1497
12	Alpha-tubulin	HA44G	57.9	52.0	1088
13	S-protein	SPRO	57.3	56.9	1635
14	Alpha-1-acid glycoprotein	A1GLY2	55.8	51.1	1593
15	4F2 glycosylated antigen	4F2HG1	62.4	53.5	1110
16	Adenosine deaminase	ADAG	52.9	51.6	3935
17	Keratin 18	KER18	52.8	55.2	2532
18	Growth hormone-1	GHCSA	56.6	50.0	5162
19	Gamma-b-crystallin	CRYGBC	46.8	47.0	2115
20	Pulmonary surfactant protein	SPBAA	57.1	58.0	1039
21	Myoglobin	MGI1	48.9	48.3	1894
22	Insulin	INSO1	65.1	67.1	2185
23	Steroid 21-hydroxylase	MHCP42	62.8	52.1	1560
24	c-Ha-ras 1	RASH	70.1	75.7	1663
25	r-ras gene	RASR1	55.3	63.5	1173
26	Protein C	PRCA	57.6	53.5	2200
27	Na,K-ATPase	ATP1A2	50.8	53.6	1576
28	ATP synthetase beta subunit	ATPSYB	43.6	48.4	2098
29	Coagulation factor IX	FIXG	38.9	38.9	2965
30	Interleukin-1-alpha	IL1AG	39.3	42.3	2160
31	Haptoglobin	HPARS1	46.6	39.6	1040
32	Epsilon-globin	HBB	36.0	40.4	19,288
33	Alpha-2-globin	HBA4	72.6	57.1	2500

tries 22–53) against the GC levels of the corresponding extended (> 10 kb) sequences in which they were present. Values for exons 24, 27, 28, 32, and 40 are shown, but were not taken into account in calculating the correlation coefficient and the slope for reasons given in the Discussion. The slope of the least-square line through the points was 0.81 and the correlation coefficient 0.85. This line was about 5–10% higher than the diagonal line.

Figure 2A shows a plot of the GC levels of third codon positions from the same set of human genes shown in Fig. 1A against the GC levels of DNA fractions. The least-square line through the points showed a slope of 2.61 and a correlation coefficient of 0.82. Again, values for genes 6 and 13 are shown but were not taken into account in calculating the slope and the correlation coefficient.

Figure 2B shows a plot of the GC levels of third codon positions from the set of genes of Fig. 1B against the GC levels of the sequences in which the genes were present. The slope was 1.65 and the correlation coefficient was 0.84. Values for points 24, 27, 28, 32, and 40 are shown, but were not taken into account for calculating the correlation coefficient and the slope.

Figure 3A and B shows cumulative plots corresponding to Fig. 1A and B and to Fig. 2A and B, respectively, except that this time points 24, 27, 28, 32, and 40 were taken into account. In contrast, points 1, 2, 3, and 18 were disregarded because they were already represented in the second set of genes (they correspond to points 28, 47, 22, and 32, respectively). In the first case, the least-square line through the points, which was 5–10% higher than the diagonal line, had a slope of 0.81 and a correlation coefficient of 0.79. In the second case, the slope was 1.69 and the correlation coefficient 0.78.

Figure 4A presents a plot of GC levels from all introns listed in Table 1 against all corresponding exons (except for points 1, 2, 3, and 18, which concerned genes present in both sets; see above). A slope of 1.0 was obtained with a correlation coefficient of 0.77. Figure 4B shows a plot of GC levels of introns against the GC levels of 5′ flanking sequences preceding the first exons (see Table 3 and Materials and Methods). A slope of 0.94 and a correlation coef-

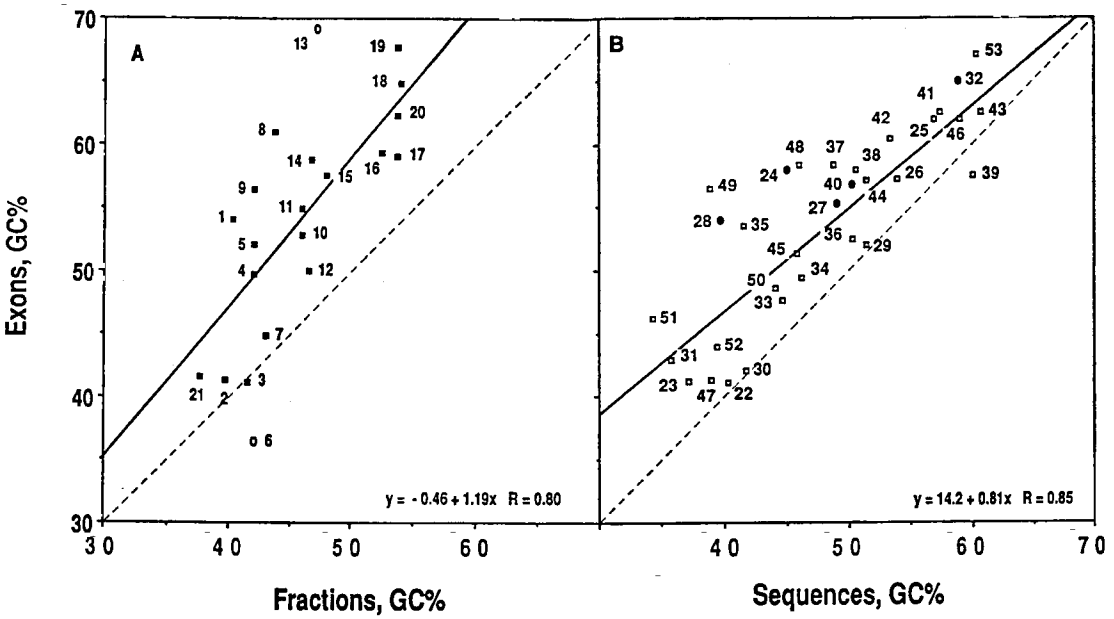

Fig. 1. Plots of GC levels of exons from human genes (see Table 1 for their numbering and their properties) against GC levels (**A**) of the DNA fractions in which the sequences were localized; or (**B**) of the extended DNA sequences in which the genes are present. The least-square lines through the points, their equations, and their correlation coefficients are shown. In **A** points 6 and 13 and in **B** points 24, 27, 28, 32, and 40 are shown as empty and filled circles, respectively, but were neglected in calculating correlation coefficients and least-square lines (see text). Whereas the plot of Fig. 1A is basically a plot of the GC level of exons against the GC level of intergenic sequences, the plot of Fig. 1B corresponds to a plot of the GC level of exons against the GC level of introns (see text).

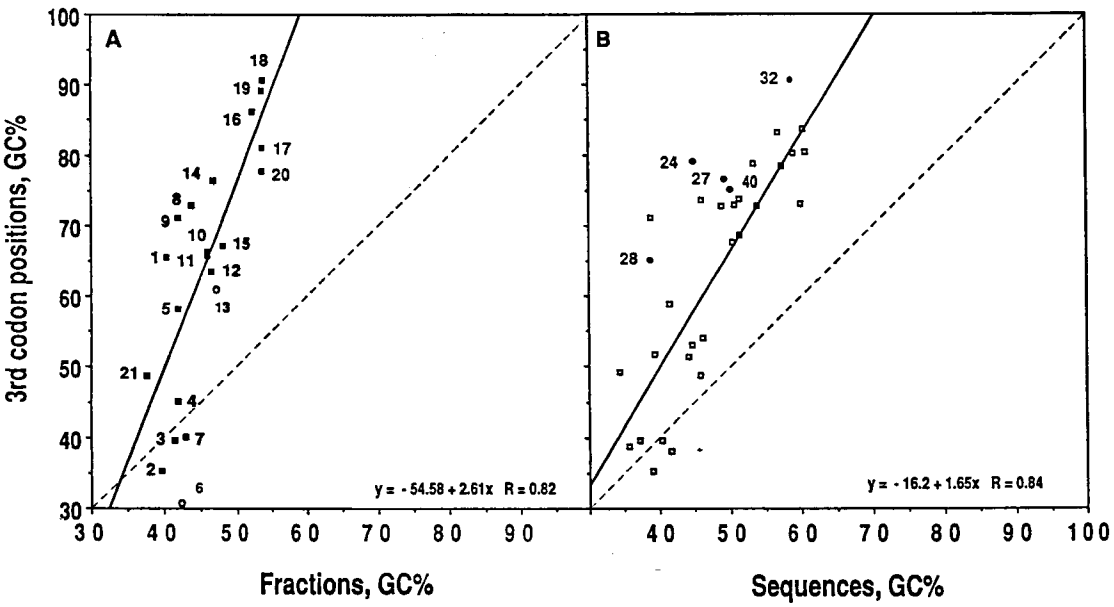

Fig. 2. Plots of GC levels of third codon positions from human genes (see Table 1) against the GC levels (**A**) of DNA fractions in which the corresponding genes had been localized; or (**B**) of the extended DNA sequences in which the genes are present. Other notations are as in Fig. 1. Concerning the meaning of the plots see legend to Fig. 1 and text.

ficient of 0.85 were obtained. The GC levels of 3′ flanking sequences were not studied because the number of available sequences > 1 kb was too small.

Another property of the genes that was investigated here concerned the correlation of GC levels of third codon positions with those of first + second codon positions (Fig. 5). After eliminating the deviant points 6 and 13, a slope of 2.16 and a correlation coefficient of 0.61 were obtained.

Discussion

The Compositional Correlations between Exons and Isochores

The results shown in Fig. 1A define the properties (slope, intercept, and correlation coefficient) of the linear relationship that exists between the GC levels of exons and the GC levels of the large DNA frag-

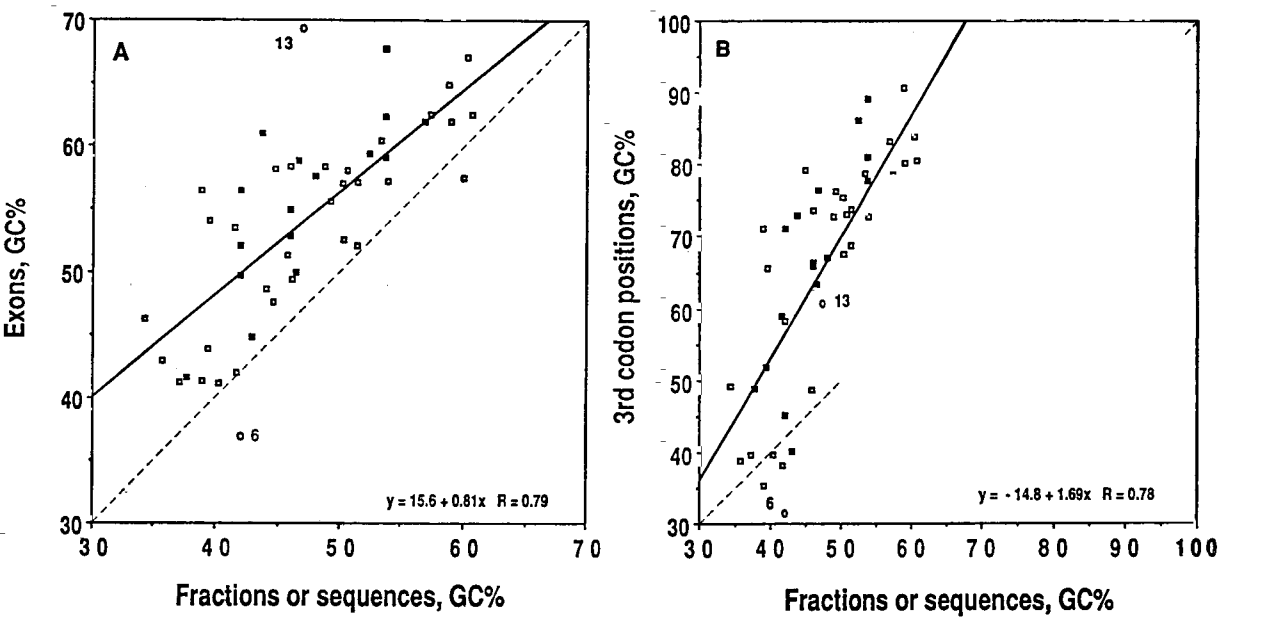

Fig. 3. Cumulative plots (corresponding to Fig. 1A and B and Fig. 2A and B, respectively) of GC levels (**A**) of exons and (**B**) of third codon positions from human genes against GC levels of DNA fractions or extended sequences in which they were located. The least-square lines through the points, their equations, and their correlation coefficients are shown. Points 1, 2, 3, and 18 are not shown and were not taken into account in calculating correlation coefficients and slopes because they were already represented in the second set of genes (they correspond to points 28, 47, 22, and 32, respectively).

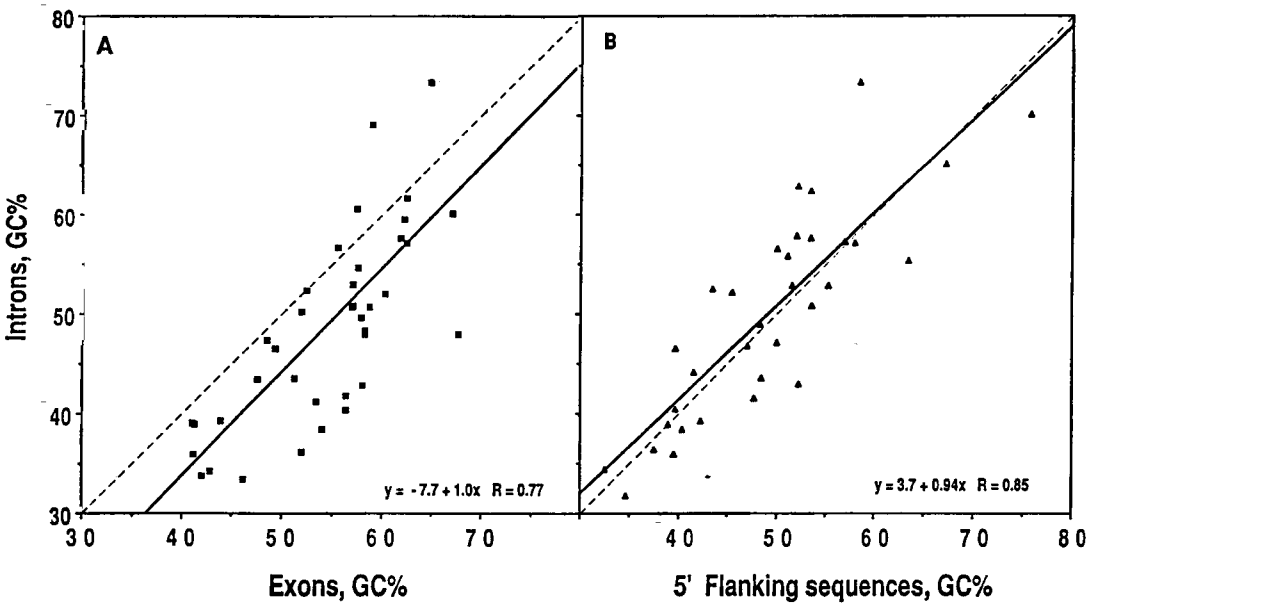

Fig. 4. Plot of GC levels of introns against GC levels (**A**) of exons and (**B**) of 5′ flanking sequences preceding the first exon. The least-square lines through the points, their equations, and their correlation coefficients are shown. In **A**, points 1, 2, 3, and 18 are not shown and were not taken into account in calculating the correlation coefficient and the slope (see legend to Fig. 3). In **B**, data of Table 3 were used.

ments containing the genes under consideration. These results merit several comments.

1) The large DNA fragments containing the genes can be equated with intergenic sequences. Indeed, the plot of Fig. 1A corresponds, for the most part, to a correlation between small coding sequences (average size: 619 ± 204 bp) and large genomic segments (at least 200–300 kb in size, except for points 8, 14, and 17–20, in which case the segment size was at least 70 kb). The large genomic segments correspond essentially to intergenic sequences, as coding sequences represent less than 5% of the human genome (Bernardi 1989). In fact, of the other noncoding sequences, introns represent less than 1–2% of the fragments in all cases where intron size is known and make up 10–15% of the fragments

only in two exceptional cases (coagulation factor IX and HPRTB genes; points 2 and 3), but are highly unlikely to represent as much in the cases where intron size is unknown.

2) Exons located within each gene cluster exhibit very similar GC levels. The six sets of clustered genes investigated were the beta-globin and the alpha-globin genes, the interferon-induced protein genes, the collagen genes, the growth hormone genes, and the gamma-crystallin genes (see Table 2). It had previously been observed that the genes in the two globin gene clusters of human, mouse, and chicken had the same composition within each cluster (Bernardi et al. 1985). This was explained to be due to the fact that the clusters (which are at most ~70 kb in size, in the case of the beta-globin gene cluster) are each contained within a single isochore (Bernardi et al. 1985). Here, this observation is extended to a total of six human gene clusters comprising 19 sequenced genes. It should be noted, however, that some compositional differences were found in the introns of the gamma-globin genes compared to the other genes in the cluster and, moreso, in the introns of the two gamma-crystallin genes, where, in addition, the size of introns in one gene was much larger than in the other one. An analysis of the sequences (not shown) indicated, however, that the differences found in introns were local differences and did not correspond to compositional discontinuities in the DNA segment containing these genes, namely to isochore borders.

3) The collagen and interferon receptor genes correspond to two deviating points. These points were not taken into account in calculating slopes and correlation coefficients of all plots presented because the two alpha (VI) collagen genes (point 13) showed exceedingly high GC levels in their coding sequences due to the very high levels of glycine and proline in collagens (these two amino acids have codons containing only G or C in first and second codon positions), and because the interferon alpha receptor gene (point 6) is one of the GC-poorest human gene sequences.

4) For the intergenic sequences around the alpha- and beta-globin gene clusters GC levels of the large fragments in which the genes were embedded could be compared with the GC levels of long sequences from the bank. In the case of the beta-globin gene (point 1) the fraction in which the exons were located had a GC level of 40.3%. This value compares very well with the value of 39.5%, as obtained for 73,326 bp around the beta-globin gene cluster (point 28), in which only 7% of the sequence was represented by introns. In contrast, in the case of the alpha-globin gene (point 18), the fraction containing the exons was estimated to be 53.7% in GC, whereas the GC level for 12,847 bp around the genes

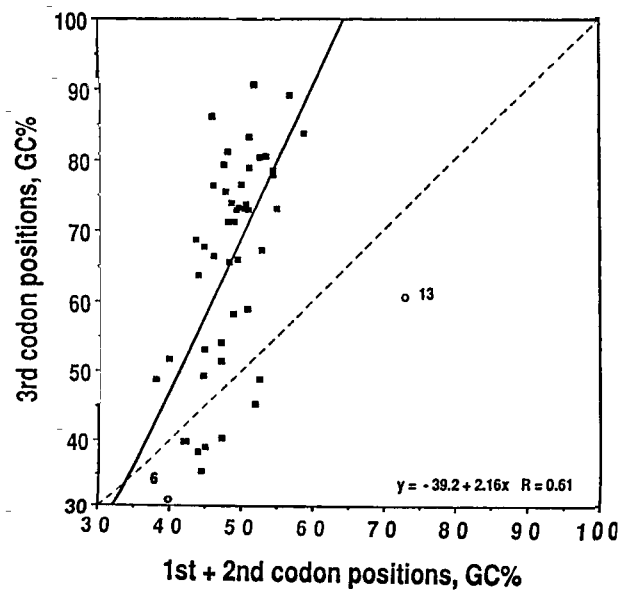

Fig. 5. Plot of GC levels of third codon positions against GC levels of first plus second positions of human genes. The least-square straight-line through the points, its equation, and its correlation coefficient are shown. Points 1, 2, 3, and 18 are not shown and were not taken into account (see legend to Fig. 3). Points 6 and 13 are shown, but they were not taken into account in calculating the correlation coefficient and the slope.

(in which introns represented only 4% of the sequence) is 58.8% (point 32). This discrepancy might be associated (a) with the abundance of CpG islands and Alu sequences in the relatively short sequenced region, as it is known that both such sequences increase in relative amounts in GC-rich isochores (Zerial et al. 1986b; Bernardi 1989); (b) with a higher methylation level in GC-rich isochores (a situation already found in plant genomes; unpublished), because methylation lowers the buoyant densities and leads to underestimation of GC levels of DNA fractions; or (c) with yet another factor, namely the poor resolution of the GC-richest fractions (see below).

5) The strongly heterogeneous GC-richest fractions are poorly resolved. GC levels of DNA fragments were calculated from the modal buoyant densities in CsCl of DNA fractions, as obtained after preparative equilibrium sedimentation in Cs_2SO_4 in the presence of a sequence-specific DNA ligand, BAMD. According to the experimental conditions used (mainly the ligand/nucleotide molar ratio, *rf*), one can increase the resolving power at one end of the base composition spectrum, with the consequence of a loss of resolution at the opposite end (Cortadas et al. 1977; Macaya et al. 1978). Routinely, high *rf* (0.14) values are used (Zerial et al. 1986b). This leads to poor resolution of GC-poor fractions, and therefore, at least potentially, to poor correlation between the modal buoyant densities of DNA fractions that represent large relative amounts of total DNA and the GC levels of the genes contained in them. This effect is, however, limited by

the remarkable compositional homogeneity of GC-poor isochores (see the agreement between fragment GC estimate and sequence data for the beta-globin gene mentioned above). On the other hand, if high *rf* values are used, high GC fractions are better resolved than at low *rf*. The actual results still are less favorable than in the previous case, because compositional heterogeneity is very high in GC-rich fractions, which also contain ribosomal DNA and some satellite DNA (Zerial et al. 1986b). The lack of resolution of GC-rich isochores is likely to account for the fact that four coding sequences (from c-Ha-ras1, alpha globin, proopiomelanocortin, and c-sis genes; points 17–20 of Table 1) ranging from 59 to 68% GC were all found in the same GC-richest DNA fraction (see Figs. 1A and 2A). This fraction, estimated to be 53.7% in GC, is well below the 60% GC level of the long sequences that surround several coding sequences having GC levels in the same range as alpha globins (see points 39, 43, and 53 of Table 1). It should be noted that this possible artifact has the consequence of increasing the slope of Figs. 1A and 2A. An obvious way of overcoming the poor resolution of the GC-richest fractions would be to rerun them in preparative gradients; this is, however, not easily done because of the very small amounts of DNA corresponding to such fractions.

Compositional Correlations between Introns, Exons, and 5′ Flanking Sequences

The compositional correlations between exons and introns are largely defined by the results of Fig. 1B (see below). These correlations were checked by plots of GC levels of introns against GC levels of exons (Fig. 4A), and complemented by the study of the correlation between GC levels of introns and GC levels of 5′ flanking sequences (Fig. 4B).

1) The long sequenced DNA fragments from the bank largely correspond to introns. Indeed, it should be realized that in practically all cases, except for points 24, 27, 28, 32, and 40, introns represent over two-thirds of the sequences in which exons are present (see Table 1; values in parentheses). Therefore, if the exceptional points mentioned above are eliminated, as was done, Fig. 1B essentially corresponds to an exon vs intron plot. If compared with Fig. 1A, the plot of Fig. 1B is characterized by a comparable correlation coefficient, by a comparable higher level above the diagonal line, but by a lower slope (see below).

2) The difference in slope between the plots of Fig. 1A and B is not significant. This difference might be due to the first one being the correlation between coding sequences and intergenic sequences, and the second one between coding sequences and introns. Several points indicate, however, that this difference is not significant. (a) The unity slope of a plot of GC levels of all introns listed in Table 1 against the GC levels of all corresponding exons (Fig. 4A) indicates that, on the average, GC levels increase in parallel in exons and introns from the same genes. (b) The least-square line of Fig. 4A is about 5% lower in GC than the diagonal line, indicating that introns tend to be systematically lower than exons, as also indicated by Fig. 1B. (c) A plot of GC levels of introns against GC levels of corresponding 5′ flanking sequences (see Fig. 4B) shows a unity slope, indicating that the GC level increases in parallel in introns and in the corresponding 5′ flanking sequences; moreover, the line through the points indicates an essential coincidence of GC levels in introns and 5′ flanking sequences. (d) Cumulative plots from both sets of data (including now points 24, 27, 28, 32, and 40, as these sequences, which have low relative amounts of introns, are equivalent to the points of the first set) do not show any significant decrease in correlation coefficient (see Fig. 3). (e) In two cases, data from the first set of genes could be compared with data from the second set. In the case of coagulation factor IX (point 2) and HPRTB (point 3) genes the fractions in which the exons were localized had GC levels of 39.7% and 41.5%, respectively. These values compare very well with values of 39%, as obtained for 38,059 bp around the coagulation factor IX exons (point 47), and 40.3%, as obtained for 56,536 bp around the HPRTB exons (point 22). Because in the two cases under consideration, introns represent 79% and 69%, respectively, of the sequences, this comparison indicates a good match between intergenic sequences (which represent most of the DNA from the fractions) and introns.

The Compositional Correlations between Codon Positions and Isochores

The results of Fig. 2A define the linear relationship between the GC levels of third codon positions and the GC levels of the large DNA fragments containing the genes under consideration. Given the excellent compositional correlation between third codon positions and corresponding exons (the correlation coefficients are 0.95 and 0.97, respectively, for the two sets of genes listed in Table 1), one can expect good compositional correlations between first + second codon positions and large DNA fragments or long DNA sequences around and within the genes. Indeed, correlation coefficients of 0.53 and 0.71, respectively, were found for the two sets of genes (not shown).

The Correlation between the GC Level of Third and First + Second Codon Positions

This correlation (Fig. 5) will be commented upon elsewhere (Bernardi and Bernardi 1991; D'Onofrio

et al. 1991; and unpublished). Suffice it to mention here that the correlation can be used (like the compositional correlation between exons and introns) to establish that the human genes studied here are representative of all human genes [as judged from the 1400 or so coding sequences investigated in D'Onofrio et al. (1991)]; that the correlation fits with the finding (see the preceding section) that both third and first + second codon positions are compositionally correlated with intergenic sequences and introns; and that the correlation is universal (Bernardi and Bernardi 1991).

The Phylogenetic Spread of the Compositional Correlations

If data from Bernardi et al. (1985) are compared with the present results (Figs. 1A and 2A) one finds good agreement in that the exon slope previously found was 0.89 with a correlation coefficient of 0.80, the least-square line being about 10% higher than the diagonal; and the third codon position slope was 2.0 with a correlation coefficient of 0.77. The good agreement between the previous data (Bernardi et al. 1985) concerning genes from several warm-blooded vertebrates and the present results suggests that the correlation is common for all warm-blooded vertebrates. Needless to say, a more extensive investigation of genes from warm-blooded vertebrates other than human would be most interesting at this point to be sure about this conclusion. Recent investigations, which indicate that a correlation very close to those shown in Fig. 2 is found in the genomes of cold-blooded vertebrates (Bernardi and Bernardi 1991), proceed, however, in the direction of a common correlation for the genes of all vertebrates, as was tacitly assumed in the original work (Bernardi et al. 1985). It should be stressed, however, that correlations similar to those investigated here are found in plants (Montero et al. 1990).

Conclusion

The first conclusion is that in the human genome exons are linearly correlated in composition with both intergenic sequences and introns, which are, however, lower by 5–10% GC. In other words, in the human genome, coding and noncoding sequences (whether inter- or intragenic) are compositionally correlated. This conclusion is important for two different reasons. First of all, it is important because the correlations indicate that both coding and noncoding sequences are subject to basically identical compositional constraints (Salinas et al. 1987). Second, if the gene sample is representative of all human genes (as it is; see D'Onofrio et al. 1991), the correlation allows the assignment of an isochore and a chromosomal band (Giemsa-positive, Giemsa-negative, or telomeric) location to genes that have not been localized experimentally (unpublished).

The second conclusion is that compositional correlations exist among codon positions of human genes, a point that will be dealt with in more detail in D'Onofrio et al. (1991).

It should be stressed that although the compositional correlations between exons and introns or exons and intergenic sequences appear to be the same for all vertebrate genomes, those among codon positions are universal (Bernardi and Bernardi 1991). In fact, it has been argued that the compositional correlations between coding and noncoding sequences and those among codon positions amount to a genomic code (Bernardi 1990; and unpublished).

Acknowledgments. We thank the Association Française contre les Myopathies (AFM) for a fellowship to B.A., FEBS and the Di Capua Foundation, Naples, Italy, for fellowships to G.D., and the Association pour la Recherche contre le Cancer (ARC), the Ministry for Research and Technology (MRT), and AFM for financial support.

References

Aebi M, Fäh J, Hurt N, Samuel CE, Thomis D, Bazzigher L, Pavlovic J, Haller O, Staeheli P (1989) cDNA structures and regulation of two interferon-induced human mx proteins. Mol Cell Biol 9:5062–5072

Ambros PF, Sumner AT (1987) Correlation of pachytene chromomeres and metaphase bands of human chromosomes, and distinctive properties of telomeric regions. Cytogenet Cell Genet 44:223–228

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

Bernardi G (1979) Organization and evolution of the eukaryotic genome. In: Morgan J, Whelan WJ (eds) Recombinant DNA and genetic experimentation. Pergamon, London, pp 15–20

Bernardi G (1984) Sequence organization of the vertebrate genome. In: Arber W, Illmensee K, Peacock WJ, Starlinger P (eds) Genetic manipulation: impact on man and society. Cambridge University Press, Cambridge, pp 171–178

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

Bernardi G (1990) Le génome des vertébrés: organisation, fonction et evolution. Biofutur 94:43–46

Bernardi G, Bernardi G (1985) Codon usage and genome composition. J Mol Evol 22:363–365

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

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

Bernardi G, Bernardi G (1990b) Compositional patterns 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 (in press)

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, Mouchiroud D, Gautier C, Bernardi G (1988) Compositional patterns in vertebrate genomes: conservation and change in evolution. J Mol Evol 28:7–18

Corneo G, Ginelli E, Soave C, Bernardi G (1968) Isolation and characterization of mouse and guinea pig satellite DNA's. Biochemistry 7:4373–4379

Cortadas J, Macaya G, Bernardi G (1977) An analysis of the bovine genome by density gradient centrifugation: fractionation in 3,6 bis-(acetato-mercurimethyl) dioxane density gradient. Eur J Biochem 76:13–19

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 111:227–233

D'Onofrio G, Mouchiroud D, Aïssani B, Gautier C, Bernardi G (1991) Correlation between the compositional properties of human genes, codon usage and amino acid composition of proteins. J Mol Evol 32:504–510

Dutrillaux B (1973) Nouveau système de marquage chromosomique: les bandes T. Chromosoma 41:395–402

Filipski J, Thiery JP, Bernardi G (1973) An analysis of the bovine genome by Cs_2SO_4/Ag^+ density gradient centrifugation. J Mol Biol 80:177–197

Gardiner K, Aïssani B, Bernardi G (1990) A compositional map of human chromosome 21. EMBO J 9:1853–1858

Gouy M, Gautier C, Attimonelli N, Lanave C, Di Paola G (1985) ACNUC—a portable retrieval system for nucleic acid sequence database: logical and physical design and usage. Cabios 1:167–172

Ikemura T (1985) Codon usage and tRNA in unicellular and multicellular organisms. Mol Biol Evol 2:13–34

Kettman R, Meunier-Rotival M, Cortadas J, Cuny G, Ghysdael J, Mammerickx M, Burny A, Bernardi G (1979) Integration of bovine leukemia virus DNA in the bovine genome. Proc Natl Acad Sci USA 76:4822–4826

Macaya G, Thiery JP, Bernardi G (1976) An approach to the organization of eukaryotic genomes at a macromolecular level. J Mol Biol 108:237–254

Macaya G, Cortadas J, Bernardi G (1978) An analysis of the bovine genome by density-gradient centrifugation. Eur J Biochem 84:179–188

Meunier-Rotival M, Soriano P, Cuny G, Strauss F, Bernardi G (1982) Sequence organization and genomic distribution of the major family of interspersed repeats of mouse DNA. Proc Natl Acad Sci USA 79:355–359

Montero LM, Salinas J, Matassi G, Bernardi G (1990) Gene distribution and isochore organization in the nuclear genome of plants. Nucleic Acids Res 18:1859–1867

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 human and murids. J Mol Evol 27:311–320

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

Salinas J, Zerial M, Filipski J, Crépin M, Bernardi G (1987) Non-random distribution of MMTV proviral sequences in the mouse genome. Nucleic Acids Res 15:009–3022

Schild D, Brake AJ, Kiefer MC, Young D, Barr PJ (1990) Cloning of three human multifunctional de novo purine biosynthetic genes by functional complementation of yeast mutations. Proc Natl Acad Sci USA 87:2916–2920

Singer MF (1982) SINES and LINES: highly repeated short and long interspersed sequences in mammalian genomes. Cell 28: 433–434

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

Soriano P, Meunier-Rotival M, Bernardi G (1983) The distribution of interspersed repeats is non-uniform and conserved in the mouse and human genomes. Proc Natl Acad Sci USA, 80:1816–1820

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 (1986a) Genomic localization of hepatitis B virus in a human hepatoma cell line. Nucleic Acids Res 14:8373–8386

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

Received December 12, 1990/Revised February 2, 1991