Gene, 106 (1991) 173–183
© 1991 Elsevier Science Publishers B.V. All rights reserved. 0378-1119/91/$03.50

GENE 06029

CpG islands: features and distribution in the genomes of vertebrates

(Isochores; CpG doublets; DNA methylation; CsCl density gradients)

Aïssani and Giorgio Bernardi

Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 75005 Paris (France)

Received by G. Roizès: 17 May 1991
Revised/Accepted: 21 June/24 June 1991
Received at publishers: 16 July 1991

SUMMARY

We have investigated the distribution of unmethylated CpG islands in vertebrate genomes fractionated according to their base composition. Genomes from warm-blooded vertebrates (man, mouse and chicken) are characterized by abundant CpG islands, whose frequency increases in DNA fractions of increasing % of guanine + cytosine; % G + C (GC), in parallel with the distribution of genes and CpG doublets. Small, yet significant, differences in the distribution of CpG islands were found in the three genomes. In contrast, genomes from cold-blooded vertebrates (two reptiles, one amphibian, and two fishes) were characterized by an extreme scarcity or absence of CpG islands (detected in these experiments as *Hpa*II tiny fragments or *HTF*). CpG islands associated with homologous genes from cold- and warm-blooded vertebrates were then compared by analyzing CpG frequencies, GC levels, *Hpa*II sites, rare-cutter sites and G/C boxes (GGGGCGGGGC and closely related motifs) in sequences available in gene banks. Small, yet significant, differences were again detected among the CpG islands associated with homologous genes from warm-blooded vertebrates, in that CpG islands associated with mouse or rat genes often showed low CpG and/or GC levels, as well as low numbers of *Hpa*II sites, rare-cutter sites and G/C boxes, compared to homologous human genes; more rarely, CpG islands were just absent. As far as cold-blooded vertebrates were concerned, a number of genes showed CpG islands, which exhibited a much lower frequency of CpG doublets than that found in CpG islands of warm-blooded vertebrates, but still approached the statistically expected frequency; none of the other features of CpG islands associated with genes from warm-blooded vertebrates were present. Other genes did not show any associated CpG islands, unlike their homologues from warm-blooded vertebrates.

INTRODUCTION

Unmethylated CpG islands (which will be referred to as CpG islands henceforth) are sequences 0.5–2 kb in size which have been mainly investigated in the mouse genome (Bird, 1986; see also Gardiner-Garden and Frommer, 1987). CpG islands are characterized (*i*) by high GC levels (over 60% G + C); (*ii*) by clustered, unmethylated CpG doublets which occur at frequencies approaching those statistically expected in vertebrates; CpG doublets are the only potential sites of DNA methylation and the only doublets that are strongly under-represented relative to statistical expectations (this is the so-called CpG shortage); and (*iii*) by G/C boxes, namely GGGGCGGGGC (the consensus recognition sequence for transcription factor Sp1; Dynan, 1986; 1989) and closely related sequences. Because of their nt sequences, CpG islands contain frequent *Hpa*II sites, CCGG, and can, therefore, be detected as *Hpa*II tiny fragments (or *HTF*); moreover, they contain clustered sites

Correspondence to: Dr. G. Bernardi, Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris (France)
Tel. (33-1)43 29 58 24; Fax (33-1)44 27 79 77

Abbreviations: BAMD, 3,6 bis-(acetatomercurimethyl)dioxane; EtdBr, ethidium bromide; GC, % of guanine + cytosine (% G + C); *HTF*, *Hpa*II tiny fragment(s); nt, nucleotide(s); *O.*, *Odontophrinus*; r, ribosomal; ρ, buoyant density (g/ml); *tsp*, transcription start point(s); UWGCG, University of Wisconsin Genetic Computer Group.

for rare-cutting restriction enzymes, which recognize GC-rich sequences comprising one or two unmethylated CpG doublets (Smith et al., 1987). In fact, over 80% of *Not*I sites are associated with CpG-rich islands in sequenced human DNA (Kusuda et al., 1991). CpG islands were reported in the genomes of mammals, birds, reptiles, amphibians and fishes (Cooper et al., 1983), and were found to be associated with the 5′-flanking sequences, 5′ exons and 5′ introns of all housekeeping genes tested so far, and of some tissue-specific genes, as well as with the 3′ exons of some tissue-specific genes (Bird, 1986; Gardiner-Garden and Frommer, 1987).

Along a different research line, work from our laboratory (Bernardi, 1985; 1989; Bernardi et al., 1985; Bernardi and Bernardi, 1986; Mouchiroud et al., 1991) showed that the distribution of genes and CpG doublets is strikingly nonuniform in the genome of warm-blooded vertebrates. Indeed, the concentration of genes increased with increasing GC of the isochores harboring them. (Isochores are the long, compositionally homogeneous segments, which make up the mosaic genome of vertebrates; isochores belong in several families characterized by different GC levels.) On the other hand, CpG doublets, which are very strongly avoided in the GC-poor genes located in GC-poor isochores, are increasingly less avoided in genes located in isochores of increasing GC.

The increasing frequency of genes and CpG doublets in isochores of increasing GC levels in warm-blooded vertebrates led us to investigate the distribution of CpG islands in these genomes and to compare it with the distribution in the genomes of cold-blooded vertebrates, which are characterized by the absence or scarcity of very GC-rich isochores and genes (Bernardi and Bernardi, 1990a,b; 1991).

These investigations showed that, in warm-blooded vertebrates, CpG islands are abundant and increase in frequency in isochores of increasing GC levels. Moreover, at variance with the previously claimed presence of CpG islands in all vertebrate genomes (Cooper et al., 1983), CpG islands (detected as *HTF*) are extremely scarce or absent in the genomes of cold-blooded vertebrates. These observations, already mentioned elsewhere (Bernardi et al., 1988; Bernardi, 1989), will be presented here in detail, along with the additional finding that quantitative differences in the distribution of CpG islands appear to exist among the genomes of the warm-blooded vertebrates investigated.

This paper will also report a comparison of CpG islands associated with homologous sequenced genes from both cold- and warm-blooded vertebrates. This analysis showed that (*i*) CpG islands of mouse and rat genomes tend to be lower in GC, to be less rich in *Hpa*II sites, rare-cutter sites, and G/C boxes compared to CpG islands from the human genome; more rarely, they are just absent; (*ii*) when present, CpG islands from cold-blooded vertebrates are different from those found in warm-blooded vertebrates because they exhibit much lower GC levels (less than 60% and, in most cases, less than 50% GC), a scarcity or an absence of *Hpa*II sites and rare-cutter sites, and no G/C boxes; the only characteristic feature of CpG islands from warm-blooded vertebrates found in the rare CpG islands of cold-blooded vertebrates is the presence of CpG doublets at levels approaching those predicted by base composition; CpG levels are, however, much lower than in CpG islands of warm-blooded vertebrates; very often CpG islands are just absent, whereas they are present next to homologous genes from warm-blooded vertebrates.

RESULTS AND DISCUSSION

(a) The distribution of CpG islands in the genomes of warm-blooded vertebrates

Fig. 1a shows the analytical CsCl profiles, the modal buoyant densities and the relative amounts of human DNA fractions. Fig. 1b shows similar data, as obtained with human DNA depleted of rDNA. In both cases, a satellite DNA centered at ρ of about 1.700 g/ml can be seen in the GC-rich fractions. In Fig. 1b, the CsCl profiles are characterized, as expected, by much lower levels of rDNA (which corresponds to the peak centered at ρ of about 1.712 g/ml).

Fig. 1,c and d display the gel electrophoresis of *Hpa*II fragments from the experiment of Fig. 1a after terminal labelling; gels were loaded with equal amounts of either DNA (Fig. 1c) or radioactivity (Fig. 1d; the latter procedure was used in all other experiments shown). Both series of results indicate that CpG islands, as detected in the form of *HTF*, increase in relative amounts as human DNA fractions increase in GC. It is noteworthy that in fractions 5–9 from the experiment of Fig. 1a (or in the corresponding fractions from the experiment of Fig. 1b), practically no long labelled fragments can be seen. This indicates, on the one hand, a very high frequency of unmethylated *Hpa*II sites in those fractions, and on the other hand that the GC-poor satellite DNA (ρ = 1.700 g/ml) present in those fractions could not be labelled because of the absence or extreme scarcity of *Hpa*II sites. In agreement with the latter point, large unlabelled fragments, presumably derived from the satellite, could be detected at the top of the gel by EtdBr staining (not shown).

The results of Fig. 1,c and d are not due to *Hpa*II degradation of GC-rich rDNA, which accounts (Bird et al., 1985) for one fifth of total CpG islands in the mouse genome, because (*i*) rDNA was found, by hybridization experiments with a rDNA probe (Fig. 1e), to be centered on fraction 8, whereas CpG islands show a steady increase over all fractions, but especially over fractions 7, 8 and 9, the latter being the richest in *HTF*; and (*ii*) human DNA,

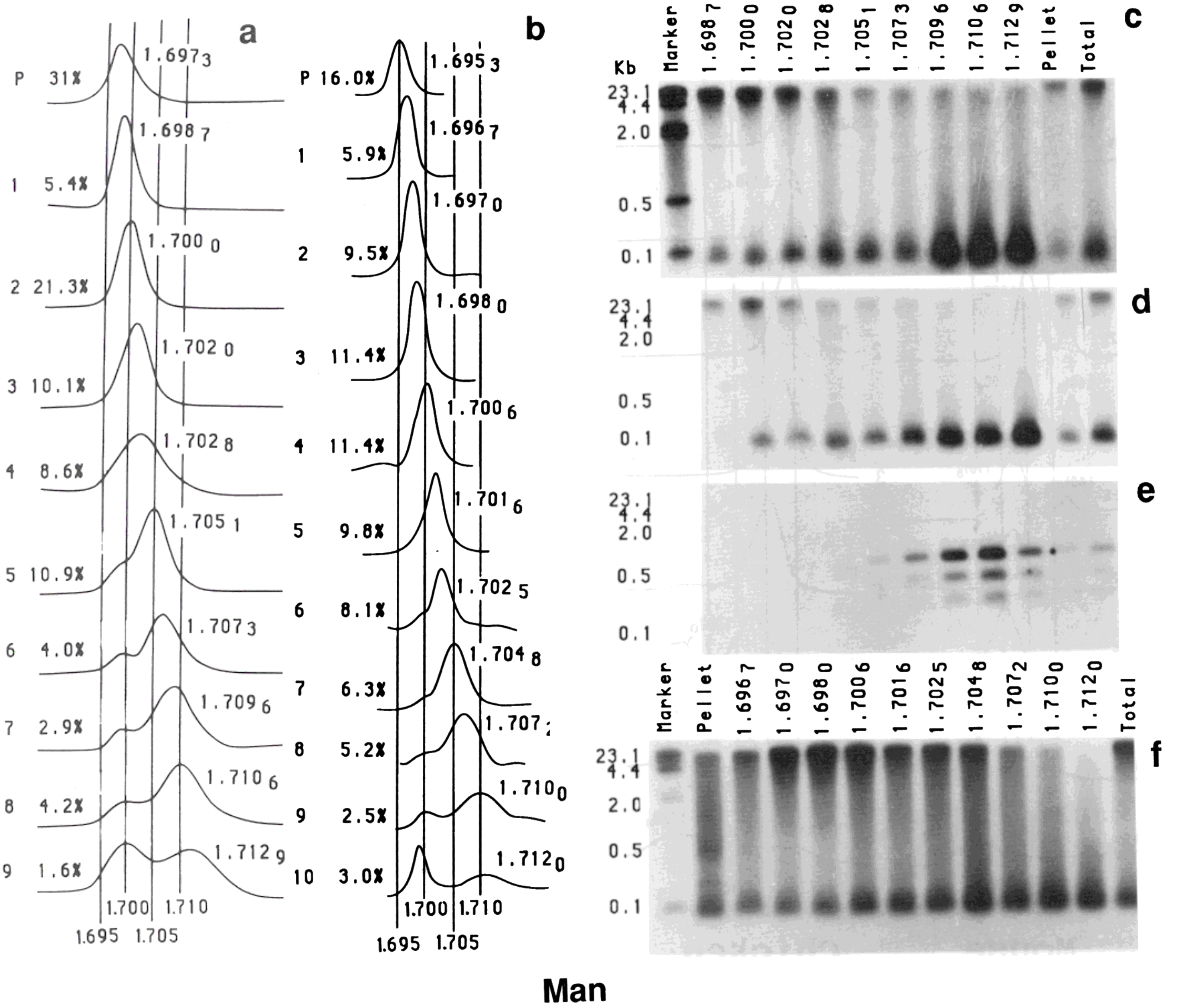

Fig. 1. CsCl profiles and *Hpa*II fragments of compositional fractions from human DNA. (**a** and **b**) CsCl profiles of fractions from human placenta DNA (in panel **b** after depletion of rDNA as described by Hemleben et al., 1977), as obtained from preparative centrifugation in Cs_2SO_4/BAMD density gradient (Zerial et al., 1986). An *rf* (ligand/nt molar ratio) of 0.14 was used. Modal buoyant densities and relative amounts are indicated. P indicates the pellet. Notice the satellite peak (centered at about 1.700 g/ml) in the last fractions. (Panels **c** and **d**) Autoradiograms of terminally labelled (Cooper et al., 1983; 1.2% agarose gels were used in all experiments) *Hpa*II fragments from DNA fractions of experiment **a**. Equal DNA amounts were loaded in panel **c** (in which case the pellet was under-labelled), whereas equal amounts of radioactivity were loaded in panel **d** (this procedure was used in all other experiments). (Panel **e**) Hybridization of a rDNA probe (pX1r14; Botchan et al., 1977) on a *Hpa*II digest of fractions from experiment **a**. (Panel **f**) Autoradiogram of terminally labelled *Hpa*II fragments from DNA fractions of experiment **b**.

depleted of rDNA (Fig. 1b), showed the expected decrease of *HTF* in GC-rich fractions, but no overall change in *HTF* distribution (Fig. 1f).

The increasing level of CpG islands in DNA fractions with increasing GC was also found in mouse and chicken DNAs (Fig. 2,a and b). Differences were found, however, among the distributions of CpG islands associated with the genes from the three warm-blooded vertebrates tested. CpG islands covered a narrower and a wider compositional range in mouse (Fig. 2c) and chicken (Fig. 2d), respectively, compared to man. Moreover, while practically no labelled large fragments could be detected in the GC-richest fractions of human and chicken DNA, non-negligible amounts of them were found in the GC-rich fractions of the mouse genome.

The three genomes examined were chosen on the basis of the fact that they are characterized by different compositional patterns at the level of both DNA fragments and coding sequences (Bernardi et al., 1985; 1988; Salinas et al., 1986; Zerial et al., 1986; Mouchiroud et al., 1987;

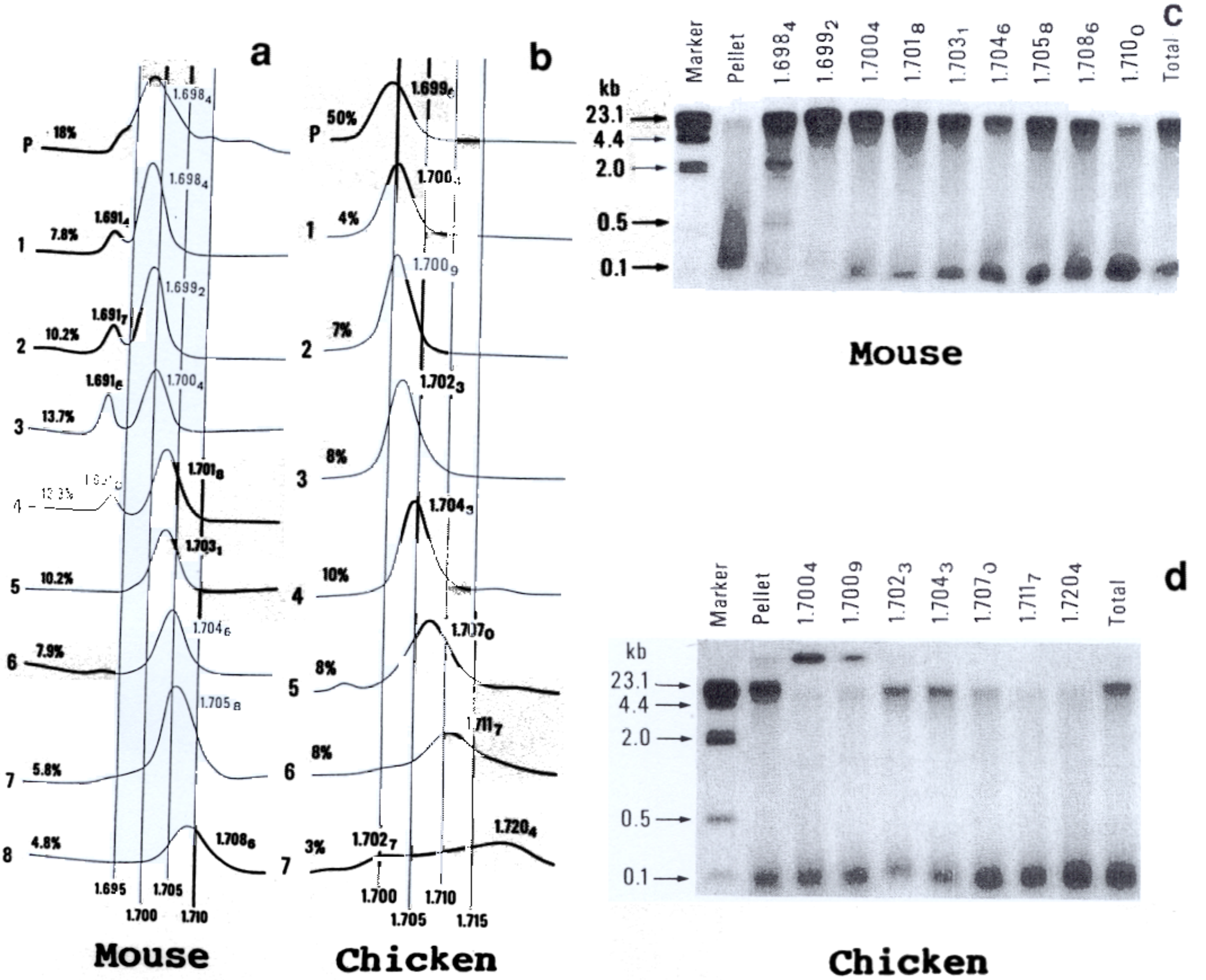

Fig. 2. CsCl profiles and *Hpa*II fragments of compositional fractions from mouse and chicken DNAs. (**a** and **b**) CsCl profiles of mouse liver and chicken erythrocyte (Salinas et al., 1986; Cortadas et al., 1979) DNA fractions as obtained after preparative centrifugation in Cs_2SO_4/BAMD density gradient. The *rf* values used in the fractionation of mouse and chicken DNAs were 0.12 and 0.14, respectively. Modal buoyant densities and relative amounts are indicated. P, pellet. Note the satellite peak (1.691–1.692 g/ml) in mouse DNA fractions 1–4. (Panels **c** and **d**) Autoradiograms of terminally labelled *Hpa*II fragments from DNA fractions of experiments **a** and **b**, respectively. See also Fig. 1 legend.

1988). Indeed, murids (as well as cricetids and spalacids) show a narrower distribution of both DNA fragments and coding sequences relative to the other mammalian genomes explored so far, and, in particular, relative to the human genome. On the other hand, the chicken genome is characterized by a distribution of DNA fragments and coding sequences which reach higher GC levels than in the human genome. The compositional patterns just described are paralleled by the distribution of CpG islands, which are more narrowly or more widely distributed in murids and birds, respectively, than in man.

The gradient of concentration of CpG islands in the genomes of warm-blooded vertebrates, as seen by end-labelling of *Hpa*II fragments (Figs. 1 and 2), parallels those of genes and of CpG doublets. Indeed, the genomes of warm-blooded vertebrates are characterized by a gene concentration which increases with the GC levels of the isochores in which the genes are located (Bernardi et al., 1985; 1988; Mouchiroud et al., 1987; 1988; Bernardi, 1989). In the best-studied case, that of the human genome, the gene concentration is low and constant in GC-poor isochores, which represent about two thirds of the genome, increases in the GC-rich isochores, and reaches a maximum (over 16 times higher than the value of the low, constant region) in the GC-richest isochores (Mouchiroud et al., 1991). On the other hand, a correlation between CpG levels and GC levels was already reported for coding sequences of vertebrates (Bernardi, 1985; 1989; Bernardi et al., 1985), for coding sequences of viruses of warm-blooded vertebrates (Bernardi and Bernardi, 1986), and for coding sequences of plants (Montero et al., 1990).

As CpG islands appear to be less frequent in the GC-rich fractions from the murine than in those from the human genome, the overall amount of CpG islands (estimated to represent 1% of the mouse genome; Bird, 1986) may be higher in the human genome, and may represent a very sizable part of the GC-richest H3 compartment, which only corresponds to 3% of the human genome. This is also indicated by the absence of labelled, large fragments at the top of the gel after *Hpa*II digestion of the GC-richest fractions from the human genome (Fig. 1).

(b) The CpG islands in the genomes of cold-blooded vertebrates

0 The genomes of five cold-blooded vertebrates (two reptiles, a crocodile and a turtle; one amphibian, *Xenopus laevis*; and two fishes, a trout and a carp) were explored in respect to their *HTF*.

The analytical CsCl profiles and the relative amounts of reptilian DNA fractions are shown in Fig. 3,a and b. In the case of crocodile (a reptile characterized by a relatively GC-rich DNA, 46% GC; Bernardi and Bernardi, 1990a), *HTF* were detected in total DNA (Fig. 3c). An analysis of the fractions indicated that this finding could, however, not be taken at face value, namely as a demonstration of CpG

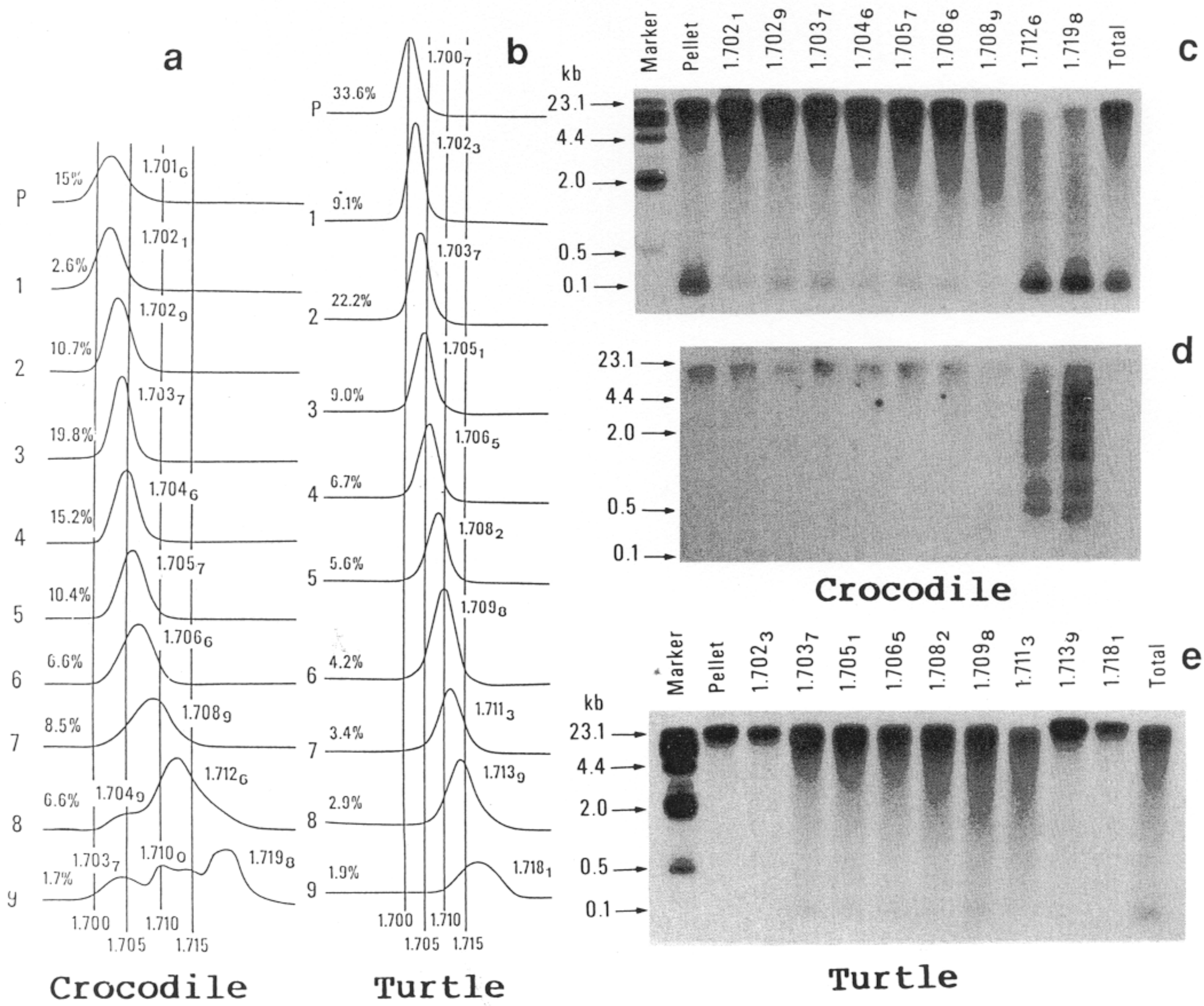

Fig. 3. CsCl profiles and *Hpa*II fragments of compositional fractions from crocodile and turtle DNAs. (**a** and **b**) CsCl profiles of crocodile (*Crocodilus niloticus*) and turtle (*Testudo graeca*) erythrocyte DNA fractions as obtained after preparative centrifugation in Cs_2SO_4/BAMD density gradient. The *rf* values used in the fractionations of crocodile and turtle DNA were 0.14 and 0.16, respectively. Modal buoyant densities and relative amounts are indicated. (Panels **c** and **e**) Autoradiograms of terminally labelled *Hpa*II fragments from DNA fractions of experiments **a** and **b**, respectively. (Panel **d**) Hybridization of an rDNA probe on an *Hpa*II digest of the crocodile DNA fractions. Small fragments (less than 0.1 kb) were not seen because they were lost in the transfer. The lowest hybridization band is at 0.5 kb, as expected from the fact that the probe used was a coding sequence and not a spacer probe. See also Fig. 1 legend.

islands. Indeed, *HTF* were only found (Fig. 3c) in the two GC-richest fractions, 8 and 9 (especially in fraction 9) and in the pellet (most probably as a consequence of contamination from the last fraction). Now, the two GC-richest fractions (*i*) only corresponded to 6.6% and 1.7%, respectively, of the genome; (*ii*) were characterized by extremely high buoyant densities (1.712_6 and 1.710–1.719_8 g/ml), by the presence of a satellite peak (centered at about 1.704 g/ml) and, in the case of fraction 9, by a strikingly multimodal profile; (*iii*) contained rDNA (which was especially abundant in the last fraction), as shown by hybridization experiments (Fig. 3d); and (*iv*) did not exhibit large, unlabelled DNA fragments at the top of the gel, as seen by EtdBr staining (not shown). The latter result indicates a complete degradation of the DNA present in the last two fractions, whereas the former ones suggest that the *HTF* may, at least very largely, correspond to fragments derived from satellite and rDNAs. In any case, the results were completely different from those found in the case of DNAs from warm-blooded vertebrates, in that they did not display the typical increase with increasing GC, and in that *HTF* were only obtained from a very small fraction of the genome. No *HTF* were detected in the DNA fractions from another reptile, the turtle (Fig. 3e; turtle DNA is also GC-rich, 45.1 % GC; Bernardi and Bernardi, 1990a).

Fig. 4,a, b and c show the analytical CsCl profiles of DNA fractions from *Xenopus*, trout and carp, respectively. *HTF* could barely be detected in total *Xenopus* DNA (Fig. 4d), but they were seen in the pellet and in fraction 1 ($\rho = 1.6986$ g/ml), which only corresponded to 1% of total

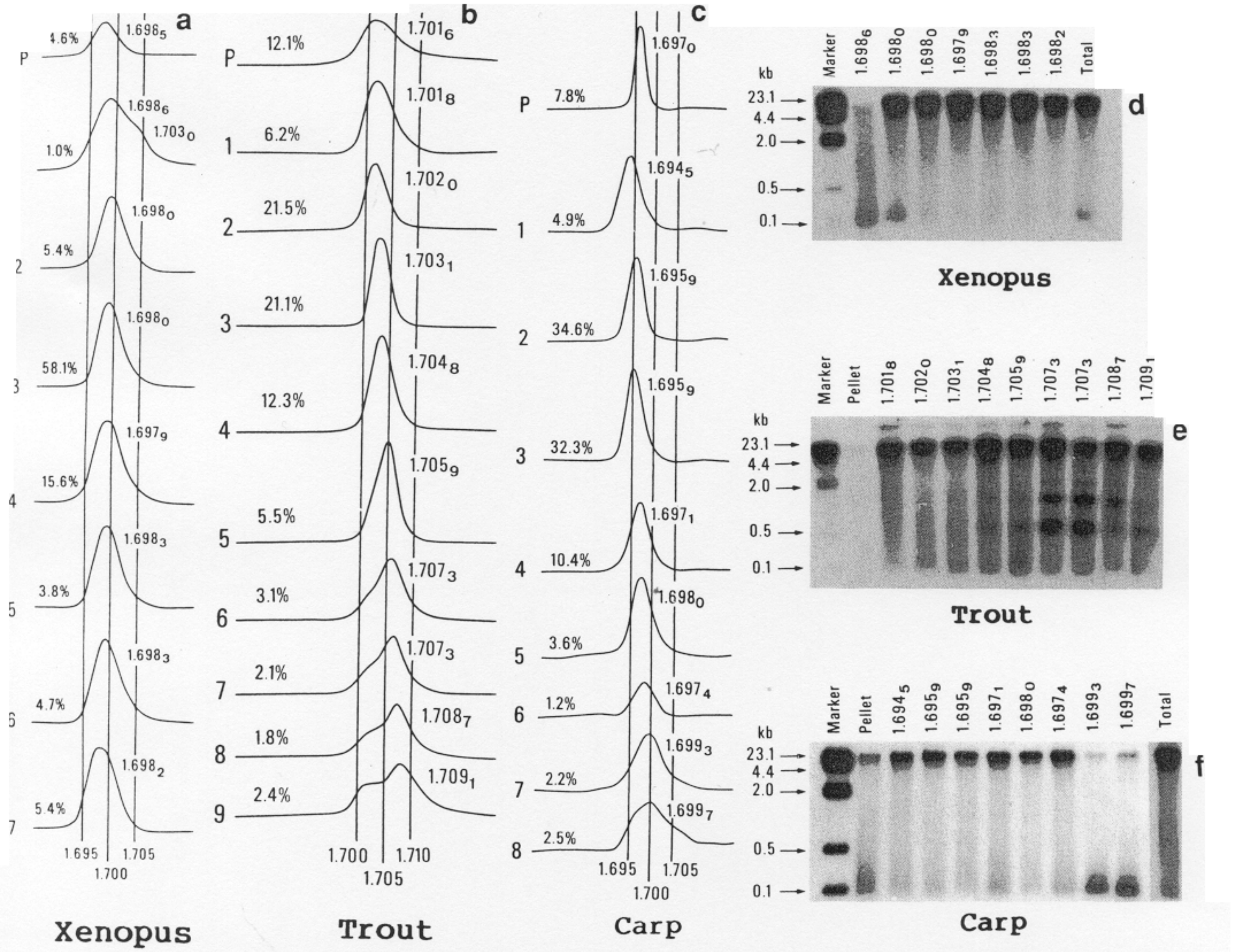

Fig. 4. CsCl profiles and *Hpa*II fragments of compositional fractions from *Xenopus*, trout and carp DNAs. (**a**, **b** and **c**) CsCl profiles of *Xenopus laevis*, trout (*Salmo gairdneri*) and carp (*Cyprinus carpio*) liver DNA fractions as obtained after preparative centrifugation in Cs_2SO_4/BAMD density gradient (Bernardi and Bernardi, 1990a); *rf* values of 0.14 were used. Modal buoyant densities and relative amounts are indicated. P, pellet. Satellite peaks or shoulders are visible in the CsCl profiles of fraction 1 of *Xenopus* DNA, fractions 6–9 of trout DNA and fractions 7 and 8 of carp DNA. (Panels **d**, **e** and **f**) Autoradiograms of *Hpa*II fragments from DNA fractions from experiments **a**, **b** and **c**, respectively. In panel **e**, the pellet was very poorly labelled. See also Fig. 1 legend.

DNA and exhibited a 1.703 g/ml satellite peak, that might be their source. These results are at variance with those published by Cooper et al. (1983).

As far as fish genomes are concerned, trout DNA did not show any *HTF*, again at variance with previous results by Cooper et al. (1983), but in agreement with those of Cross et al. (1991); a ladder of bands were present in several fractions which showed satellite peaks (Fig. 4f). Carp DNA (Fig. 4g) only showed *HTF* in the two GC-richest fractions, each of which only represented 2–2.5% of the genome. These fractions were separated, however, as a distinct peak in the preparative Cs_2SO_4/BAMD centrifugation (not shown), suggesting that they might correspond to a satellite DNA; shoulders were also seen in the CsCl profile of the last fraction (Fig. 4c). As in the case of crocodile, the results obtained with carp were strikingly different from those obtained with warm-blooded vertebrates (no concentration gradient of CpG islands, and *HTF* only obtained from a very small fraction of the genome).

To sum up, the results obtained with the genomes of cold-blooded vertebrates are profoundly different from those obtained with the genomes of warm-blooded vertebrates, because *HTF* were either absent or very scarce (turtle, *Xenopus* and trout), or mostly or totally originated from satellite and/or rDNAs (crocodile and carp).

Our conclusion for cold-blooded vertebrates is at variance with a previous report indicating a situation identical to that found in the genomes of warm-blooded vertebrates (Cooper et al., 1983). Recent data (Cross et al., 1991) confirm, however, the absence or great scarcity of CpG islands in fishes, which had been previously reported (Bernardi et al., 1988; Bernardi, 1989) for fishes and amphibians. Cross et al. (1991) interpret the *HTF* obtained with total DNAs from crocodile and from the amphibian *O. americanus*, as indicative of CpG islands. We have already presented evidence against this interpretation in the case of crocodile DNA. The positive results reported by Cross et al. (1991) for *O. americanus* are also probably due to degradation of its rDNA [previous studies by Cortadas and Ruiz (1988) and by Ruiz and Brison (1989) indicate the existence of small fragments in *Hpa*II digests].

On the basis of their interpretation of crocodile and *O. americanus* data, Cross et al. (1991) were led (*i*) to propose that the scarcity or absence of *HTF* was a property of fish genomes, and not of cold-blooded vertebrates in general; and (*ii*) to deny any 'simple correlation between low body temperature and absence or reduction of the *HTF*'. In contrast, our results based on the analysis of *HTF* from compositional fractions, and not only from total DNA, indicate that DNAs from cold-blooded vertebrates are characterized by the absence or by a severe scarcity of *HTF*; when present, the latter are different in their origin from those found in warm-blooded vertebrates (see sections **b** and **d**).

(c) An analysis of CpG islands associated with homologous genes from warm-blooded vertebrates

Comparisons of CpG islands associated with homologous genes from warm-blooded vertebrates are shown in Table I. CpG islands were delimited in the first gene of each homologous series on the basis of GC level (>60%) and observed/expected ratio (>0.6) of the CpG doublet. For the other genes in the homologous series, equivalent positions, relative to the AUG codon, were examined. The data of Table I are schematically presented in Fig. 5, in order of increasing overlap of the genes by the CpG islands. Fig. 6 displays some examples of the actual GC levels and CpG profiles as well as of the distribution of several restriction sites across the CpG islands and the associated genes. The case shown is that of a CpG island which is present in the human α-globin gene but absent in the homologous genes from mouse and *Xenopus*.

CpG islands associated with human genes were always characterized by GC levels higher than 60% (with the only exception of the genes encoding histone H4) and by a ratio of observed over statistically expected CpG higher than 0.60. Similar properties were shared by the CpG islands of other warm-blooded vertebrates, except for the chicken heat-shock protein encoding gene. CpG islands from rat or mouse were, however, characterized in several cases by lower levels of both GC and CpG or of GC only and, at least in two cases, that of the mouse α-globin gene and that of the rat enkephalin A gene, by the absence of CpG islands (for the case of enkephalin A see also Table I of Aïssani and Bernardi, 1991). The different properties of CpG islands from murids appear to be confirmed by differences in the number of *Hpa*II sites and rare-cutter sites and G/C boxes, which tended to be systematically lower than in human CpG islands (Table I and Fig. 6).

An analysis of the data of Fig. 6 indicates that the 'extragenic' part of CpG islands, or rather the part preceding the AUG start codon, is endowed with different properties relative to the intragenic part (see also Table I). Indeed, *Hpa*II sites are much less frequent, rare-cutter sites are practically nonexistent, and G/C boxes are absent in the intragenic part. This justifies considering CpG islands on their own, independently of the CpG island-like features exhibited by the genic sequences. Further observations and comments on this point will be reported in the following paper (Aïssani and Bernardi, 1991).

(d) A comparison of CpG islands associated with homologous genes from cold-blooded vertebrates

A common situation found in the sequences associated with genes from cold-blooded vertebrates which corre-

TABLE I

List and properties of homologous vertebrate genes investigated

Genes						CpG islands								
Names of gene products[a]	Mnemonics[b]	%GC[c]		Translation[d]		Positions[e]	GC[f]	CpG[g]		%GC	CpG[h]	*Hpa*II[i]		G/C[j]
							%			>60	>0.6	a	b	boxes
		Exons	3rd	Start	Stop			%	o/e					
Metallothionein II	**Hummet2**	62.4	82.2	838	1527	360–869	68.6	10.6	0.91	+	+	8	1	1
	MusmetII	58.6	74.2	439	1018	1–470	64.7	8.70	0.83	+	+	5	1	/
	Fsbmetb	50.8	52.5	320	1247	1–351	37.8	4.60	1.29	−	+	1	1	/
β-Actin	**Humaccyba**	60.6	84.1	1157	2810	1–1254	69.4	13.9	1.15	+	+	21	3	3
	Chkacb	53.5	64.9	1543	4175	386–1640	73.9	13.6	0.98	+	+	16	3	16
	Rataccyb	56.8	73.7	1242	3132	85–1339	66.1	10.5	0.96	+	+	15	2	3
	Fsbactba	54.9	70.0	1568	3053	411–1665	43.7	3.70	0.78	−	+	8	2	/
Enkephalin A	**Bovenkeph***	56.6	67.5	110	901	1–275	69.8	10.2	0.84	+	+	3	1	/
	Ratenkar*	53.4	66.7	328	1137	218–493	59.8	4.70	0.54	−	(−	2	1	/
Retinol-binding protein	**Humrbp***	59.0	81.0	52	651	1–337	63.5	8.60	0.86		+	2	1	/
	Xelsrbp*	45.5	50.5	41	634	1–337	43.3	3.00	0.77		+	1	1	/
α-Tubulin	**Humha44g**	56.9	71.1	1794	5299	994–2892	66.0	7.30	0.67			17	7	/
	Xeltuba	49.8	51.1	675	6733	1–1764	46.3	2.60	0.49			2	2	/
Myc oncogene	**Hummycm***	58.8	76.6	559	1878	1–1253	63.8	8.70	0.86	+	+	9	3	/
	Xelmyc*	53.9	67.1	149	1401	1– 836	54.4	4.00	0.54	−	−	2	1	/
	Chkmyc	61.0	81.7	2268	4489	1–3442	65.5	10.5	0.98	+	+	44	7	2
	Ratmyc	57.3	72.7	4163	6481	1895–5337	57.4	5.60	0.68	(−)	+	23	2	/
	Muscmyc1	58.1	76.1	2539	>3295	271–3306	58.9	6.50	0.87	(−)	+	17	2	/
Heat-shock protein	**Humhsp70d**	62.1	92.0	489	2411	138–2247	63.0	8.20	0.83	+	+	12	8	/
	Chkhsp	52.2	62.5	392	2296	41–2150	53.6	3.80	0.53	−	−	4	3	/
	Mushsp70a	58.1	79.1	631	2532	280–2389	58.9	10.1	0.88	(−	+	12	7	/
	Xelhsp70b	47.0	48.5	488	2431	137–2247	47.1	2.20	0.40	−	−	2	1	/
Histone H4	**HumH4**	69.0	73.1	241	552	78– 727	55.8	8.50	1.10	−	+	2	1	/
	Dukhist34	68.9	97.2	119	430	1– 611	63.8	10.3	1.00	+	+	4	3	/
	Mushist4	62.9	79.8	258	559	95– 750	56.7	6.70	0.84	−	+	2	1	/
	Xelhis4	60.0	73.0	408	719	245– 815	52.7	4.40	0.63	−	+	1	1	/
	Fsbhis42b	60.5	73.1	838	1149	675–1330	56.0	8.10	0.96	−	+	1	1	/
Platelet-derived growth factor	**Humpdgfar***	58.8	75.2	338	1023	1–850	68.0	9.50	0.83	+	+	9	2	/
	Muspdgfa*	58.2	71.1	63	653	1–525	62.5	6.30	0.65	+	+	3	2	/
	Xelpdgfa*	47.0	45.8	416	1096	28–878	53.9	5.10	0.70	−	+	2	2	/
α-Globin	**Humhba4**	64.8	88.8	10514	11208	10000–11208	72.0	10.6	0.82	+	+	17	5	1
	Mushba	58.1	67.9	405	1089	1– 1089	55.0	2.10	0.24	−	−	2	1	/
	Xethbba	46.8	51.1	941	1834	427– 1635	37.3	0.70	0.23	−	−	1	1	/
Proopiomelanocortin	**Humpomc**	67.7	89.2	4485	8184	7356–8184(3′)	69.4	10.1	0.84	+	+	8	8	/
						200–1750(5′f)	60.6	7.50	0.82	+	+	19	0	/
	Ratpomc1			exon1	partial	1– 872(5′f)	56.1	3.90	0.50	−	−	4	0	/
	Muspomc3	63.2	83.9	36	611	1– 712(3′)	62.2	7.00	0.72	+	+	5	5	/
	Pigpomcr*	69.8	90.3	78	881	197– 844	70.0	12.5	0.91	+	+	11		/
	Rcpomc*	46.9	55.5	58	846	177– 809	46.4	2.10	0.39	−	−	0	0	/
	Xelpomc*	44.4	48.0	<1	612	1– 575	44.7	1.70	0.35	−	−	0	0	/
	Fsbpomc*	59.8	76.4	<1	408	1– 371	60.6	3.50	0.38	+	−	0	0	/

[a] All the nt sequences were extracted from GenBank Release 66 (December 1990) using the ACNUC retrieval system (Gouy et al., 1984).

[b] The mnemonics of the mammalian genes were used to search for homologous genes in the genomes of cold-blooded vertebrates. Sequences were then aligned and assessed for homology which was found to be 60–80% between genes from cold- and warm-blooded vertebrates and 75–92% among genes from warm-blooded vertebrates. Warm-blooded vertebrate sequences are in bold type. The order of genes is that of Fig. 5 (see figure legend). Asterisks indicate cDNA sequences.

[c] GC of (*i*) exons and (*ii*) third codon positions.

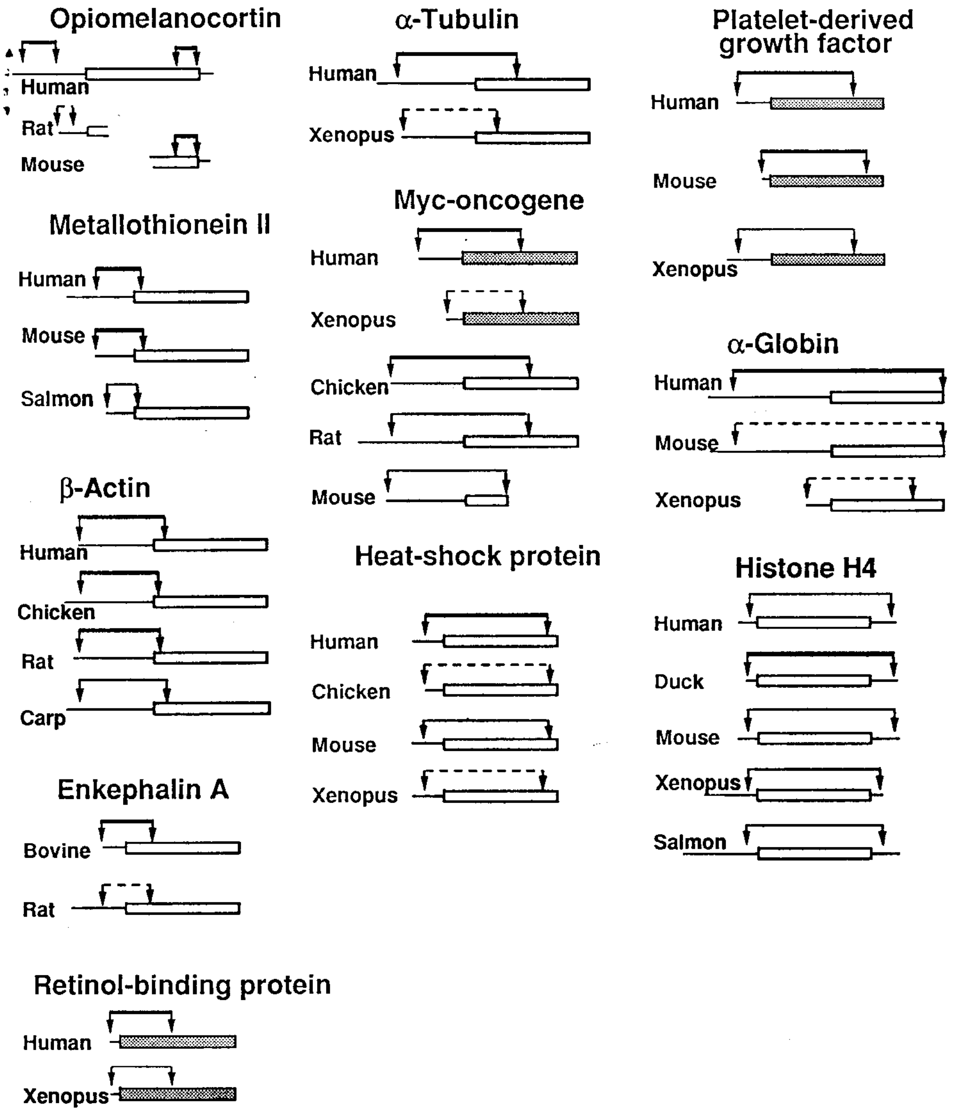

Fig. 5. Schematic representation of CpG islands as found in association with homologous genes from warm- and cold-blooded vertebrates. CpG islands are presented in an order of increasing overlap with gene sequences. Vertical arrows indicate the start and end of CpG islands, as defined in the first gene of each set, and equivalent positions for other sequences; thick and thin lines connecting the arrows indicate whether GC levels are above or below 60%; broken lines indicate that GC is lower than 60% and CpG is lower than 60% of the expected statistical value. Boxes indicate genes, shaded boxes coding sequences. Gene boxes are not proportional to gene size, but CpG islands are.

sponded to CpG islands from warm-blooded vertebrates was that of a lower GC level and of a CpG level exhibiting an observed/expected ratio lower than 0.6 (Table I). Indeed, this was the case for genes encoding the α-globin, enkephalin A, α-tubulin, heat-shock protein and for c-*myc* gene from *Xenopus*; an identical case was found for the opiomelanocortin-encoding gene of *Rana catesbeiana*. Since these two key features of CpG islands, as found in

[d] Positions of translation start and stop codons.

[e] Positions (nt) of CpG islands as determined on the first gene from warm-blooded vertebrates (nt numbers refer to the GenBank data). For other genes in each set, positions which are equivalent relative to the AUG codon, are indicated. In the case of proopiomelanocortin genes of man and rat, positions on 3′ end (3′) of gene or 5′ flanking (5′f) sequence are relative to the *tsp*, because the CpG island was located in the 5′-flanking sequence of the gene, over 2 kb in size away from the gene.

[f] %GC of CpG islands.

[g] CpG is presented both as (*i*) % of dinucleotides and as (*ii*) observed/expected (o/e) ratio.

[h] Values higher than 60% (for GC) or 0.6 (for CpG o/e ratio) are indicated by a plus symbol, lower values by a minus symbol. Parentheses indicate borderline values. These columns relate to the two preceding columns.

[i] Numbers of *Hpa*II sites; columns a and b concern the whole (a) CpG island and its section (b) after the AUG start codon, respectively.

[j] Sequences identical with the core sequence of the *Sp*1 site GGGCG. For columns a and b, see footnote i. Slashes denote absence of GGGCG.

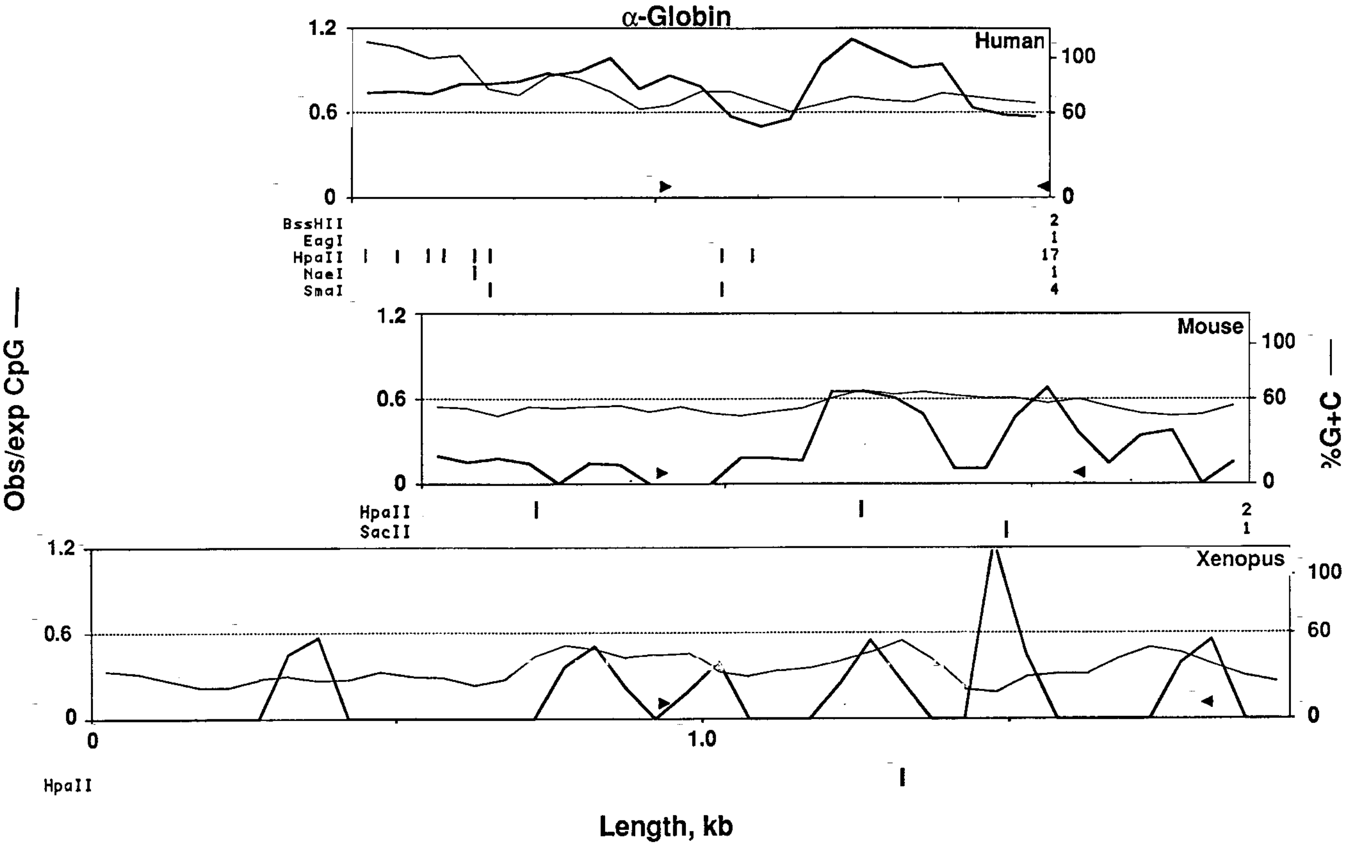

Fig. 6. Observed CpG/expected CpG (thick line) and GC (thin line) values are given for the α-globin genes of man, mouse and *Xenopus*. A moving window of 50 nt was used. Positions and numbers of certain restriction sites (*Hpa*II, *Eag*I, *Nae*I, *Sac*II, *Nru*I, *Sma*I, *Mlu*I, *Not*I, *Bss*HII) are also indicated. Solid arrowheads indicate the positions of the AUG start codon and the stop codons. Data were plotted using Statplot and Mapplot software packages of the UWGCG program (Devereux et al., 1984).

warm-blooded vertebrates, were missing, this situation corresponded to the absence in cold-blooded vertebrates of CpG islands associated with genes, whose homologues do show CpG islands in warm-blooded vertebrates.

A second situation was that of sequences from cold-blooded vertebrates corresponding to CpG islands basically characterized by lower GC and CpG levels, compared to the CpG from homologous genes of warm-blooded vertebrates, CpG doublets approaching, however, statistical frequency (Table I; Fig. 6). This was found in genes encoding the retinol-binding protein, histone H4 and platelet-derived growth factor of *Xenopus*, as well as in genes encoding *β*-actin and histone H4 of fishes. Not only were GC levels lower than 60% (in most cases even lower than 50%), but rare-cutter sites and G/C boxes were also absent. Moreover, *Hpa*II sites were either absent or much more rare, so explaining the severe scarcity or absence of *HTF*.

(e) Conclusions

The main conclusions of the present work are (*i*) that CpG islands exhibit a concentration gradient in the genomes of warm-blooded vertebrates; the increase in CpG islands accompanies the GC increase in isochores and parallels the increases in concentration of CpG doublets and of genes; (*ii*) that CpG islands, as present in the genome of murids, are lower in amounts, in GC levels, *Hpa*II sites, rare-cutter sites and G/C boxes compared to CpG islands from the human genome; these differences parallel the lower GC levels attained by genes and isochores from murids compared to man; (*iii*) that CpG islands are very scarce in the genomes of cold-blooded vertebrates; when present, they only share with the CpG islands from warm-blooded vertebrates the property that CpG levels, although much lower than in warm-blooded vertebrates, approach statistical expectation.

These conclusions indicate a strong correlation between CpG levels and GC levels in the genome of vertebrates. This point and its implications for the origin and evolution of CpG islands are further investigated in the following paper (Aïssani and Bernardi, 1991).

ACKNOWLEDGEMENTS

We wish to thank our colleagues Giacomo Bernardi, Giuseppe D'Onofrio and Jun Kusuda for very useful information and discussions, and the Association Française contre les Myopathies (AFM) for a Fellowship to B.A. and for financial support, and the Ministère de la Recherche et de la Technologie for a grant.

REFERENCES

Aïssani, B. and Bernardi, G.: CpG islands, genes and isochores in the genome of vertebrates. Gene 106 (1991) 185–195.

Bernardi, G.: The organization of the vertebrate genome and the problem of the CpG shortage. In: Cantoni, G.L. and Razin, A. (Eds.), Chemistry, Biochemistry and Biology of DNA Methylation. Alan Liss, New York, 1985, pp. 3–10.

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

Bernardi, G. and Bernardi, G.: Compositional constraints and genome evolution. J. Mol. Evol. 22 (1986) 363–365.

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

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

Bernardi, G. and Bernardi, G.: Compositional properties of nuclear genes from cold-blooded vertebrates. J. Mol. Evol. (1991) in press.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F.: The mosaic genome of warm-blooded vertebrates. Science 228 (1985) 953–956.

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

Bird, A.P.: CpG-rich islands and the function of DNA methylation. Nature 321 (1986) 209–213.

Bird, A., Taggart, M., Frommer, M., Miller, O.J. and Macleod, D.: A fraction of the mouse genome that is derived from islands of non-methylated, CpG-rich DNA. Cell 40 (1985) 91–99.

Botchan, P., Reeder, H.D. and Dawid, I.B.: Restriction analysis of the nontranscribed spacers of *Xenopus laevis* rDNA. Nucleic Acids Res. 11 (1977) 599–607.

Cooper, D.N., Taggart, M.H. and Bird, A.P.: Unmethylated domains in vertebrate DNA. Nucleic Acids Res. 11 (1983) 647–658.

Cortadas, J. and Ruiz, I.R.G.: The organization of ribosomal genes in diploid and tetraploid species of the genus *Odontophrynus* (Amphibia, Anura). Chromosoma 96 (1988) 437–442.

Cortadas, J., Olofsson, B., Meunier-Rotival, M., Macaya, G. and Bernardi, G.: The DNA components of the chicken genome. Eur. J. Biochem. 99 (1979) 179–186.

Cross, S., Kovarik, P., Schmidtke, J. and Bird, A.: Non-methylated islands in fish genomes are GC-poor. Nucleic Acids Res. 19 (1991) 1469–1474.

Devereux, J., Haeberli, P. and Smithies, O.: A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. 12 (1984) 387–395.

Dynan, W.S.: Promoters for housekeeping genes. Trends Genet. 2 (1986) 196–197.

Dynan, W.S.: Understanding the molecular mechanism by which methylation influences gene expression. Trends Genet. 5 (1989) 35–36.

Gardiner-Garden, M. and Frommer, M.: CpG islands in vertebrate genomes. J. Mol. Biol. 196 (1987) 261–282.

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

Hemleben, V., Grierson, D. and Dertmann, H.: The use of equilibrium centrifugation in actinomycin-caesium chloride for the purification of ribosomal DNA. Plant Sci. Lett. 9 (1977) 129–135.

Kusuda, J., Hirata, M., Yoshizaki, N., Kameoka, Y., Takahashi, I. and Hashimoto, K.: Over 80% of *Not*I sites are associated with CpG rich islands in the sequenced human DNA. Japan J. Hum. Genet. 35 (1990) 277–282.

Montero, L.M., Salinas, J., Matassi, G. and Bernardi, G.: Gene distribution and isochore organization in the nuclear genome of plants. Nucleic Acids Res. 18 (1990) 859–1867.

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

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

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

Ruiz, I.R.G. and Brison, O.: Methylation of ribosomal cistrons in diploid and tetraploid *Odontophrynus americanus* (Amphibia, Anura). Chromosoma 98 (1989) 86–92.

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

Smith, D.I., Golembieski, W., Gilbert, J.D., Kizyma, L., Miller, O.J.: Overabundance of rare-cutting restriction endonuclease sites in the human genome. Nucleic Acids Res. 15 (1987) 1173–1184.

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

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