Gene, 106 (1991) 185–195

GENE 06030

CpG islands, genes and isochores in the genomes of vertebrates

(CpG doublets; exons; introns; flanking sequences)

Brahim 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 shown that human genes associated with CpG islands increase in number as they increase in % of guanine + cytosine (GC) levels, and that most genes associated with CpG islands are located in the GC-richest compartment of the human genome. This is an independent confirmation of the concentration gradient of CpG islands (detected as *Hpa*II tiny fragments, or *HTF*) which was demonstrated in the genome of warm-blooded vertebrates [Aïssani and Bernardi, Gene 106 (1991) 173–183]. We then reassessed the location of CpG islands using the data currently available and confirmed that CpG islands are most frequently located in the 5′-flanking sequences of genes and that they overlap genes to variable extents. We have shown that such extents increase with the increasing GC levels of genes, the GC-richest genes being completely included in CpG islands. Under such circumstances, we have investigated the properties of the 'extragenic' CpG islands located in the 5′-flanking segments of homologous genes from both warm- and cold-blooded vertebrates. We have confirmed that, in cold-blooded vertebrates, CpG islands are often absent; when present, they have lower GC and CpG levels; the latter attain, however, statistically expected values. Finally, we have shown that CpG doublets increase with the increasing GC of exons, introns and intergenic sequences (including 'extragenic' CpG islands) in the genomes from both warm- and cold-blooded vertebrates. The correlations found are the same for both classes of vertebrates, and are similar for exons, introns and intergenic sequences (including 'extragenic' CpG islands). The findings just outlined indicate that the origin and evolution of CpG islands in the vertebrate genome are associated with compositional transitions (GC increases) in genes and isochores.

INTRODUCTION

In the preceding paper (Aïssani and Bernardi, 1991) we have demonstrated that (*i*) CpG islands (detected as *HTF*) exhibit a concentration gradient in the genome of warm-blooded vertebrates, their frequency increasing with increasing GC levels of the associated genes or of the isochores hosting them; (*ii*) CpG islands are less abundant, less rich in *Hpa*II sites, rare-cutter sites and G/C boxes (GGGGCGGGGC and related sequences) in murids compared to man, the latter showing features common to most mammals; (*iii*) CpG islands are scarce in the genomes of cold-blooded vertebrates; when they exist, they are characterized by GC poverty and absence of *Hpa*II sites, rare-cutter sites, G/C boxes; CpG doublets are scarce, but approach statistical frequency.

One should now consider that murids are characterized by genomes which do not attain the highest GC levels reached by other mammalian genomes (Salinas et al., 1986; Zerial et al., 1987; Mouchiroud et al., 1987; 1988; Bernardi

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: bp, base pair(s); GC, % of guanine + cytosine; H1, H2 and H3, isochore families (see Fig. 1 legend); *HTF*, *Hpa*II tiny fragment(s); Ig, immunoglobulin; L1 and L2, isochore families (see Fig. 1 legend); ORF, open reading frame; *T.*, *Torpedo*; *X.*, *Xenopus*.

et al., 1988) and that genomes from cold-blooded vertebrates are even much farther below those levels (Bernardi and Bernardi, 1990a,b; 1991). Then, the conclusions mentioned above point to the existence of a general correlation between CpG islands and GC levels of the associated genes, or of the isochores in which they are embedded. In turn, this general correlation might help in understanding the origin and evolution of CpG islands. With this in mind, we have investigated here (*i*) the compositional distribution of the coding sequences associated with CpG islands in the human genome; (*ii*) the location of CpG islands relative to human genes characterized by different GC levels; (*iii*) the properties of 'extragenic' CpG islands in warm- and cold-blooded vertebrates; (*iv*) the correlations between CpG levels of exons, introns, intergenic sequences and 'extragenic' CpG islands, with the corresponding GC levels; and (*v*) the origin and evolution of CpG islands in the vertebrate genome.

RESULTS AND DISCUSSION

(a) The distribution of CpG islands associated with human genes

The correlation between the presence of CpG islands associated with genes and the GC levels of the latter was studied on the human genes investigated by Gardiner-Garden and Frommer (1987) and by Kusuda et al. (1991). Fig. 1 indicates that 55% of the genes showing CpG islands are very GC-rich and are located in the isochore family H3, which only represents 3% of the human genome

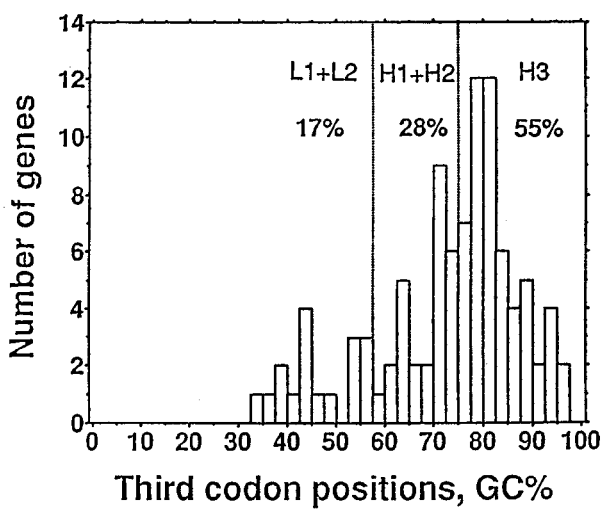

Fig. 1. Numbers of human genes associated with CpG islands (data from Gardiner-Garden and Frommer, 1987; Kusuda et al., 1990) are plotted against GC levels of third codon positions of the genes. Vertical dotted lines separate genes belonging to the isochore families indicated (L1 + L2,H1 + H2,H3; see Bernardi et al. (1985) and pertinent references cited in this report.

(Mouchiroud et al., 1991). In contrast, the CpG islands associated with genes located in the other isochore families H1 + H2 and L1 + L2, which represent 31% and 62% of the genome, only correspond to 28% and 17% of the genes, respectively.

These values should be compared with the relative amounts of genes in the isochore families H3, H1 + H2, and L1 + L2, 28%, 38% and 34%, respectively (Mouchiroud et al., 1991). This comparison shows that there are four times more genes associated with CpG islands in H3 compared with L1 + L2 (and almost three times more genes compared with H1 + H2). However, the relative numbers of CpG islands associated with genes present in H3 may be underestimated for two reasons. First, the border between isochore families H1 + H2 and H3 is still uncertain (Mouchiroud et al., 1991); if the border is placed at 70% GC in third codon position (as suggested by Fig. 1) instead of 75% (following Mouchiroud et al., 1991), almost 80% of CpG islands would be in H3; second, housekeeping genes have been found regularly associated with CpG islands (Gardiner-Garden and Frommer, 1987) and they are, at present, strongly under-represented in the bank; it is likely that, under conditions of fair representation, they would increase the number of CpG islands associated with genes present in H3.

These considerations indicate that the vast majority of CpG islands are located in the GC-richest isochores of the human genome. Interestingly, it has been estimated that 30% of human genes are associated with CpG islands (Kusuda et al., 1991). This figure matches the percentage of genes, 28%, estimated to be present in the GC-richest isochores (Mouchiroud et al., 1991). Taken at face value, this would indicate that all genes associated with CpG islands are located in isochore family H3, which clearly is not the case; yet, the estimate of Kusuda et al. (1991) supports the idea that such a situation may be approached. Along the same line, no labeled high-M_r DNA fragments remain after *Hpa*II degradation of the GC-richest fractions, as if all genes and intergenic regions of the H3 isochore family basically corresponded to CpG islands (Aïssani and Bernardi, 1991).

(b) The location of CpG islands associated with human genes

A compilation of the location of CpG islands relative to human genes using the same set of data analyzed in Fig. 1 indicates that the vast majority, over 80%, of the islands occupy the 5′-flanking sequences and, in most cases, extend into the genes. This extension may cover the 5′ end of the genes, but it may also involve the totality of the genes and even reach into the 3′-flanking sequences. Less frequently, the islands occupy sequences within the genes (8%), or sequences extending from inside the gene into the 3′-flank-

ing sequences (4%), or, most rarely, 3′-flanking sequences only (2%). In some cases (9%), it was impossible to classify the CpG islands, because no sequences flanking the islands were available; these CpG islands are, however, likely to be distributed as the CpG islands that can be classified. These results modify somewhat the generally accepted picture of the location of CpG islands, in that they stress that the overwhelming majority of them occupy 5′-flanking sequences of genes and, in most cases, extend into genes. Whether CpG islands exist in intergenic sequences far from genes is not known.

The variable extension of CpG islands (examples of which are given in Fig. 5 of Aïssani and Bernardi, 1991) prompted an investigation on the possible correlation between the degree to which CpG islands overlap gene sequences and the GC level of the corresponding genes. As shown in Fig. 2, such correlation exists, the genes located in the GC-poor isochores being only overlapped in their 5′ ends and those located in GC-rich isochores being included in the associated CpG islands. This suggests that the overlapping of CpG islands over increasing extents of gene sequences is due to the increase in GC in genes located in GC-rich isochores. Overlapping of a whole gene by a CpG island is not, as suggested, the result of a gene being short, but of a gene being GC-rich.

(c) An analysis of 'extragenic' CpG islands

The results and considerations presented in section **b** suggest that CpG islands should be considered on their own, namely as sequences located on the 5′ flanks of genes, and not necessarily in conjunction with the associated genes. An additional reason for this are the very different levels of *Hpa*II sites, rare-cutter sites and G/C boxes in the 'extragenic' and 'genic' parts of CpG islands (Aïssani and Bernardi, 1991). This prompted an analysis of 'extragenic' CpG islands, namely of CpG islands, as they exist in the 5′-flanking sequences of genes (in fact, before the AUG start codon).

This analysis (Table I) modifies the values of Table I in Aïssani and Bernardi (1991), and changes a number of CpG islands (particularly from cold-blooded vertebrates) into non-islands and vice versa. However, the basic conclusions remain the same, namely (*i*) that CpG islands from cold-blooded vertebrates are characterized by low GC levels and low CpG levels; these range from zero to about half those found in CpG islands from warm-blooded vertebrates (with one exception; see below); CpG levels attain, however, statistical frequencies; and (*ii*) that many genes from cold-blooded vertebrates simply do not contain anything resembling a CpG island in their 5′-flanking sequences, unlike their homologues from warm-blooded vertebrates.

It should be noticed that the data of Table I concerning cold-blooded vertebrates mainly comprise genes from *Xenopus*, but only a few genes from Osteichthyes, and that a gene from a Chondrychthyan, *Torpedo californica*, exhibits a CpG island similar to those of warm-blooded vertebrates. One may wonder, therefore, whether the data of Table I are

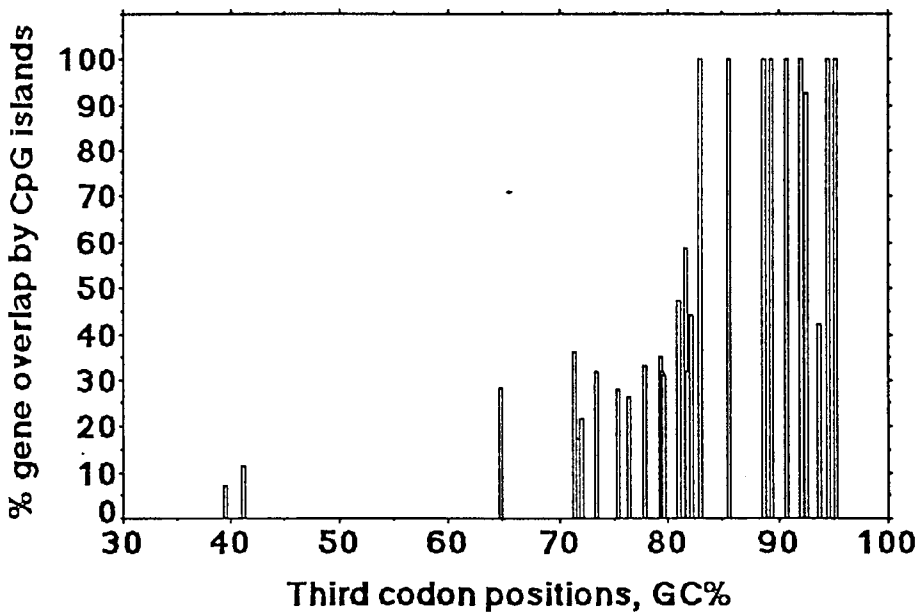

Fig. 2. Histogram of % of human coding sequences overlapped by CpG islands as a function of GC levels of third codon positions. Data from Kusuda et al. (1990) were used. Only genes whose entire coding sequences are known were used in the histogram. Five genes were not taken into consideration for the following reasons. (*a*) The CpG island of the parathymosin-encoding gene is located in the 5′-flanking sequence and does not extend into the gene, as reported; (*b*, *c*) the GC levels of the CpG islands of the ferritin L and homeobox-encoding genes are below 60%; (*d*) the intragenic part of the CpG island of the β-actin encoding gene corresponds to the first nonexpressed exon; (*e*) the intragenic CpG island of the oncogene c-*abl* was smaller in size than the moving window.

TABLE I

5'-Extragenic CpG-islands associated with homologous genes of vertebrates[a]

Gene (Name of gene product)	Species[a]	Size[b] bp	GC %	CpG[c]		GC[d] >60%	CpG[d] >0.6	*Hpa*II sites
				%	o/e			
Metallothionein II	**Human**	**478**	68.9	10.7	0.91	+	+	7
	Mouse	439	64.9	9.10	0.86	+	+	4
	Salmon	320	38.1	5.00	1.38	−	+	1
β-Actin	**Human**	1157	72.7	13.7	1.04	+	+	18
	Chicken	1157	75.3	14.0	1.02	+	+	13
	Rat	1157	65.9	10.1	0.74	+	+	13
	Carp	1157	42.8	3.40	0.75	−	+	6
Enkephalin A	**Bovine**	**110**	71.8	12.8	1.03	+	+	2
	Rat	328	58.9	3.60	0.36	(−)	−	1
Retinol binding protein	**Human**	**52**	71.2	9.80	0.77	+	+	1
	X. laevis	**41**	39.1	0.00	0.00	−	−	
α-Tubulin	**Human**	675	70.5	8.90	0.72	+	+	10
	X. laevis	675	49.2	4.70	0.80	−	+	2
Myc oncogene	**Human**	559	64.4	8.80	0.85	+	+	6
	Chicken	2268	65.6	10.9	1.02	+	+	37
	Rat	2268	57.6	5.90	0.71	(−)	+	21
	Mouse	2268	58.1	6.30	0.75	(−)	+	15
	X. laevis	142	50.0	6.40	1.14	−	+	1
Heat-shock protein	**Human**	351	62.2	8.50	0.88	+	+	4
	Chicken	351	60.5	9.10	1.05	+	+	1
	Mouse	351	61.7	6.00	0.78	+	+	5
	X. laevis	351	48.8	5.70	0.98	−	+	1
Histone H4	**Human**	163	53.1	10.8	1.56	−	+	1
	Duck	119	74.0	11.9	0.87	+	+	1
	Mouse	163	46.3	3.70	0.70	−	+	1
	X. laevis	163	45.1	2.50	0.49	−	−	
	Salmon	163	51.2	1.80	0.97	−	+	
PDGF	**Human**	338	73.1	11.6	0.90	+	+	7
	Mouse	**63**	80.9	16.1	1.02	+	+	1
	X. laevis	338	56.6	6.40	0.80	(−)	+	2
α-Globin	**Human**	941	68.4	13.4	0.78	+	+	12
	Mouse	405	52.3	0.70	0.10	−	−	1
	X. tropicalis	941	32.5	0.60	0.15	−	−	
Int-1 oncogene	**Human**	464	73.3	10.8	0.81	+	+	5
	X. laevis	158	55.7	3.20	0.43	(−)	−	
Yes protein	**Human**	208	73.5	13.0	0.98	+	+	9
	X. laevis	203	52.7	3.00	0.44	−	−	
Na/K-ATPase	**Human**	127	65.3	8.70	0.82	+	+	2
	T. californica	329	61.3	8.20	0.96	+	+	

[a] All the sequences in this and the following tables were obtained from GenBank Release 66 (December 1990) using the ACNUC retrieval system (Gouy et al., 1984). Warm-blooded vertebrates are in bold type. PDGF, platelet-derived growth factor.

[b] Bold numbers correspond to sequences shorter than 100 bp.

[c] CpG data are presented as both % and observed/expected (o/e) ratios.

[d] Values higher than 60% (for GC) or 0.6 (for CpG o/e ratio) are indicated by +, lower values by −. Parentheses indicate bordeline values. These columns relate to the two preceding columns. See also Table I in Aïssani and Bernardi (1991).

not biased by the over-representation of *Xenopus* data and whether the CpG island of *T. californica* is not a relatively common occurrence.

To answer these questions, the conclusion drawn above concerning CpG islands from cold-blooded vertebrates were checked by analyzing all available 5′-flanking sequences of genes from cold-blooded vertebrates. These data (Table II) comprised over 30 genes from fishes and twelve genes from *Xenopus*. In contrast with Table I, no comparison of 5′-flanking sequences could be made, in this case, with corresponding sequences associated with homologous genes from warm-blooded vertebrates, because either these homologous genes were not available in gene banks, or, more frequently, because their 5′-flanking sequences were not available.

The data of Table II show that (*i*) only one sequence attains a GC levels of 60% (the creatine kinase gene of *T. marmorata*), the vast majority being lower than 50%; (*ii*) CpG levels attain at most half the levels reached in CpG islands from warm-blooded vertebrates; (*iii*) less than half of the sequences reach an observed/expected CpG ratio of 0.60. Since it is likely that many genes of Table II are associated with CpG islands in warm-blooded vertebrates, the data of Table II confirm the conclusions drawn from Table I. Moreover, the data concerning genes from Chondrichthyes show that the case of the CpG islands from *T. californica* mentioned above (Table I) is exceptional.

(d) Correlations between CpG and GC levels in exons, introns and intergenic sequences of vertebrates

A correlation between CpG levels and GC levels was already reported for coding sequences of vertebrates (Bernardi et al., 1985; Bernardi, 1985), for coding sequences of viruses of warm-blooded vertebrates (Bernardi and Bernardi, 1986) and for coding sequences of plants (Montero et al., 1990). Incidentally, these results ruled out the suggestions that CpG shortage is associated with constraints exerted by the translation machinery (Subak-Sharpe et al., 1966), and that the bulk of vertebrate DNA derives from, and maintains the CpG shortage of, polypeptide-specifying DNA (Russell et al., 1976; a more detailed discussion on this point can be found in Bernardi, 1985).

Here we have re-examined this point by using the set of human coding sequences investigated by Aïssani et al. (1991) and a random set of genes from cold-blooded vertebrates (Table III). In the case of human exons, taken here as representative of exons from warm-blooded vertebrates, CpG concentration increases linearly with GC level of exons; the slope is 0.25 and the correlation coefficient is high (Fig. 3A). The same plot of the GpC doublet is just barely shifted to the left (Fig. 3A). In the case of exons from cold-blooded vertebrates, the CpG plot is practically identical to that of human exons, and the GpC plot is slightly more shifted to the left (Fig. 3B). Plots of observed/statistically expected CpG for exons from both man and cold-blooded vertebrates are again practically identical in their features, whereas observed/expected ratios for GpC oscillate around unity values, as expected (not shown). Basically, the correlations concerning CpG levels in exons of man and cold-blooded vertebrates with increasing GC are the same. In both cases, the shortage of the doublet decreases with increasing GC, approaching statistical values.

In the case of human introns (from Aïssani et al., 1991) CpG concentration increased linearly with GC level of introns (Fig. 4A); the slope was slightly lower than in the case of exons, and the correlation coefficient was almost the same. The same plot for the GpC doublet was barely shifted to the left. In the case of introns from cold-blooded vertebrates, the data were very similar (Fig. 4B) to those for human introns. The observed/expected CpG ratios vs. intron GC exhibited lower and higher slopes for man and cold-blooded vertebrates, respectively, compared to the same plots as obtained with exons, whereas in both cases, GpC plots exhibited values around unity (not shown). Overall, the correlations obtained with introns were similar in the case of genes from man and cold-blooded vertebrates. They were also basically similar to those obtained with exons, except that lower slopes were found.

Fig. 5A shows that CpG increases with increasing GC of human 5′ flanking sequences; the slope, 0.37, is higher than in the case of exons; the correlation coefficient is 0.92. The most interesting feature of the plot is that points from CpG islands are aligned with 5′ flanking sequences deprived of islands and with intergenic sequences flanking 5′ CpG islands. A similar plot for GpC shows a slight shift to the left. In the case of cold-blooded vertebrates, 5′-flanking sequences (from Table II) larger than 650 bp were taken into consideration. The slope of the CpG plot was lower than in the case of human sequences (0.23; Fig. 5B), and CpG values attained were one fourth those reached by human 5′-flanking sequences, while GC values barely reached 55% against 85% in human sequences. If 5′-flanking sequences less than 650 bp in size were used, the scatter of points increased but the slope and the intercept did not vary much. Finally, the GpC plot was again shifted to the left. A plot of observed/statistically expected CpG showed a good correlation coefficient (Fig. 6A), which was also the case for cold-blooded vertebrates (Fig. 6B). Once more, the slopes of the GpC plot were not significant, because of the very low correlation coefficients (not shown). To sum up, the results concerning CpG in 5′-flanking sequences show an increase with increasing GC of the sequences. This increase is characterized by a slope which is not very differ-

TABLE II

5'-Flanking sequences from cold-blooded vertebrate genes[a]

Species	Gene (Name of gene product)	Size (bp)	GC %	CpG[b]		GC[c] >60%	CpG[c] >0.6
				%	o/e		
Chondrichthyes							
Heterodontus francisci	Ig heavy chain	193	38.0	0.50	0.14	–	–
	IgV region	173	52.9	4.70	0.67	–	+
	IgVD region	212	49.3	3.80	0.63	–	+
	Ig light chain	442	46.9	1.40	0.27	–	–
	λ122 Ig light chain	431	46.0	1.60	0.31	–	–
Raja erinacea	Ig heavy chain	517	52.4	4.30	0.80	–	+
Torpedo californica	Acetylcholine receptor	187	41.4	1.10	0.27	–	–
	Acetylcholinesterase	130	55.1	3.10	0.41	–	–
Torpedo marmorata	Creatine kinase	84	60.7	4.90	0.54	+	–
Osteichthyes							
Cyprinus carpio	Growth hormone	273	39.7	1.80	0.48	–	–
Electrophorus electricus	Na^+ channel protein	385	37.3	2.10	0.61	–	+
Carassius auratus	Ras protein	205	30.0	0.00	0.00	–	–
	Ig heavy chain V5A region	227	55.7	2.20	0.29	–	–
	Ig heavy chain	244	52.2	1.70	0.25	–	–
	Acetylcholine receptor	93	51.6	4.30	0.81	–	+
	Nicotinic receptor	145	42.3	4.90	1.10	–	+
Brachydanio rerio	Homeobox 2.2	551	29.3	1.80	0.84	–	+
Astyanax fasciatus	Colour pigment	550	42.5	1.80	0.39	–	–
Salmo salar	Growth hormone	671	38.1	2.20	0.61	–	+
Salmo gairdneri	Ig heavy chain	130	41.1	0.80	0.21	–	–
	Protamine component c-II	355	36.2	1.40	0.43	–	–
	Histone H4	838	45.1	2.60	0.52	–	–
	Growth hormone	397	40.3	2.30	0.57	–	–
	Protamine	269	36.9	1.90	0.63	–	+
Onchorhynchus keta	Protamine	2068	37.0	0.90	0.26	–	–
	Insulin	754	42.5	2.00	0.44	–	–
	Vasotocin 1	66	43.9	3.10	0.63	–	+
Ictalurus punctatus	Somatostatin	105	42.3	2.90	0.72	–	+
Pseudopleuronectus americanus	Antifreeze protein	157	44.5	1.30	0.25	–	–
Anarhichas lupus	Antifreeze protein	400	40.6	1.00	0.24	–	–
Macrozoarces americanus	Antifreeze protein	270	37.2	1.50	0.44	–	–
Elops saurus	Ig heavy chain	357	49.2	0.60	0.10	–	–
Xiphophorus maculatus	Promoter region	506	49.4	3.60	0.60	–	+
Seriola quinqueradiata	Growth hormone	97	46.3	0.00	0.00	–	–
Hemitripterus americanus	Antifreeze protein	430	39.7	1.20	0.31	–	–
Amphibians							
Xenopus laevis	Keratin B1	821	37.5	0.40	0.11	–	–
	Keratin B2	975	41.2	0.80	0.19	–	–
	Transposon containing ORF	758	39.4	0.80	0.21	–	–
	Vitellogenin A2	1042	32.3	0.60	0.23	–	–
	Serum albumin	689	34.9	0.70	0.23	–	–
	Vitellogenin A1	1034	37.6	0.80	0.23	–	–
	Sarcomeric actin	2462	41.8	1.20	0.28	–	–
	Keratin A1	1357	37.5	1.50	0.42	–	–
	Ribosomal protein L14	909	47.2	3.00	0.55	–	–
	U1 sRNA	1138	56.1	5.00	0.64	–	+
	Elongation factor	1772	46.9	4.40	0.81	–	+
Xenopus tropicalis	Larval α-globin	636	33.1	0.90	0.33	–	–

[a] See footnote a in Table I.

[b] CpG data are presented as both % and observed/expected (o/e) ratios.

[c] Values higher than 60% (for GC) or 0.6 (for CpG o/e ratio) are indicated by plus symbols, lower values by minus symbols.

See also footnote d in Table I and Aïssani and Bernardi (1991), Table I.

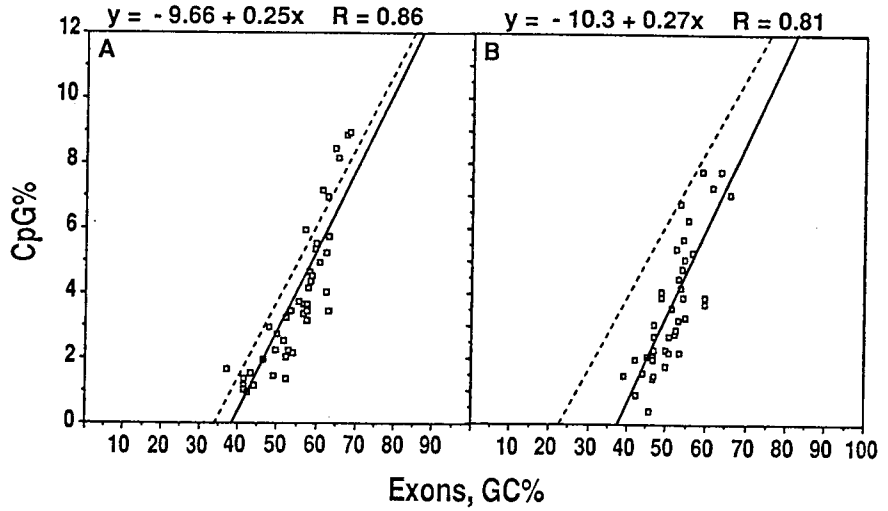

Fig. 3. Relationship between GC of exons and the % of CpG. (A) CpG levels of human exons (data from Table I of Aïssani et al., 1991) are plotted against GC levels of exons. The least-square line through the points, their equation, and their correlation coefficient, R, are shown. The dashed line is the corresponding plot for CpG (y = −7.1 + 0.27x, and R = 0.84). (B) CpG levels of exons from cold-blooded vertebrates (data of Table III) are plotted against GC levels of exons. Other indications as in Fig. 3A. The dashed line is the corresponding plot for CpG (y = −5.33 + 0.23x, and R = 0.74).

ent from those found for exons. CpG doublets from CpG islands are on the same line as CpG doublets from 5′-flanking sequences that do not correspond to CpG islands.

(e) The relationship of CpG islands to the isochores and the evolutionary origin of CpG islands

The data of Tables I and II indicate that, in contrast with warm-blooded vertebrates, cold-blooded vertebrates have very scarce CpG islands in their genomes, in agreement with the results of Aïssani and Bernardi (1991). When present, CpG islands from cold-blooded vertebrates have a much lower CpG level than CpG islands from warm-blooded vertebrates and only share with the latter a CpG frequency higher than 60% of the statistical expectation.

If the features of CpG islands in cold-blooded vertebrates are the ancestral ones, as is certain, two questions can be raised, namely how did the primitive CpG islands from cold-blooded vertebrates acquire their other features present in warm-blooded vertebrates, and how did the primitive CpG islands arise in the first place.

As far as the first question is concerned, the appearance of *Hpa*II sites, rare-cutter sites and G/C boxes in the CpG islands from the genomes of warm-blooded vertebrates should all be seen as consequences of the compositional

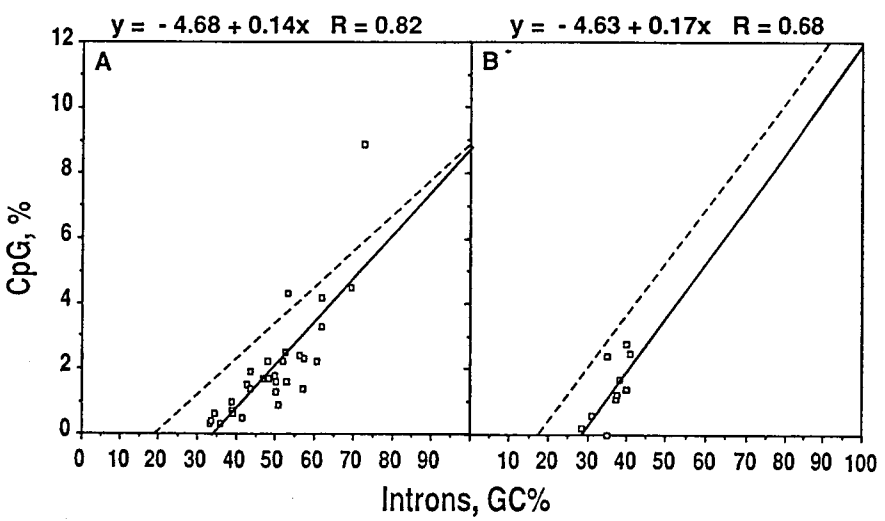

Fig. 4. Relationship between GC of introns and the % of CpG. (A) CpG levels of human introns (data from Table I of Aïssani et al., 1991) are plotted against GC levels of introns. The least-square line through the points, their equation, and their correlation coefficient, R, are shown. The dashed line is the corresponding plot for CpG (y = −3.26 + 0.18x, and R = 0.98). (B) CpG levels of introns from cold-blooded vertebrates (data of Table III) are plotted against GC levels of introns. Other indications as in Fig. 3A. The dashed line is the corresponding plot for CpG (y = −2.58 + 0.16x, and R = 0.66).

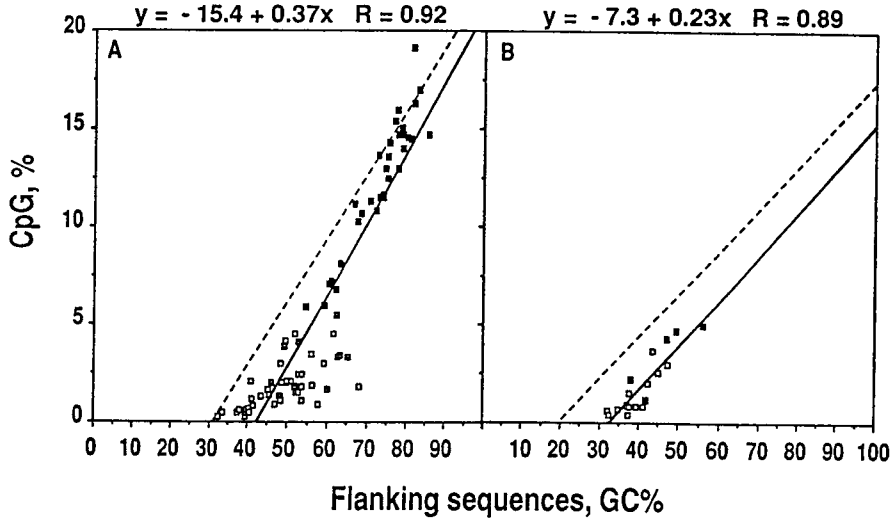

Fig. 5. Relationship between GC of flanking sequences and the % of CpG. (A) CpG levels of 5′-flanking sequences from human genome (data from Kusuda et al., 1990; Aïssani et al., 1991) are plotted against GC levels of sequences. Other indications as in Fig. 3A. Blackened squares correspond to CpG islands, barred squares to sequences flanking CpG islands. The dashed line is the corresponding plot for CpG ($y = -10.42 + 0.33x$, and $R = 0.94$). (B) CpG levels of 5′-flanking sequences (larger than 650 bp in size) from cold-blooded vertebrates (data from Table II) are plotted against GC levels of sequences. Other indications as in Fig. 3A. Blackened squares correspond to CpG islands. The dashed line is the corresponding plot for GpC ($y = -4.4 + 0.22x$, and $R = 0.84$).

transitions that led to the formation of GC-rich isochores in the latter (Bernardi et al., 1988). This proposal is most strongly supported by the comparison of CpG islands associated with homologous genes from mouse/rat or cold-blooded vertebrates and man, because all changes found in human CpG islands are correlated with GC increases. The continuity in the plot of CpG vs. GC found when studying both 5′-flanking sequences and CpG islands from the human genome (Fig. 5A) is in agreement with this conclusion, because it shows that CpG levels increase with increasing GC in 5′-flanking sequences.

Concerning the second question, the same basic explanation applies. Indeed, it is very likely that CpG islands arose in cold-blooded vertebrates through local GC increases. Data from Bernardi and Bernardi (1990a,b; 1991) indicate that there is some degree of compositional compartmentalization in the genome of cold-blooded vertebrates and that, at least in a number of cases (those showing the largest degree of compositional compartmentalization), it can be shown that genes appear to be concentrated in the GC-richest segments. Interestingly, this seems to be the case for many Osteichthyes (see Fig. 5 in Bernardi and Bernardi,

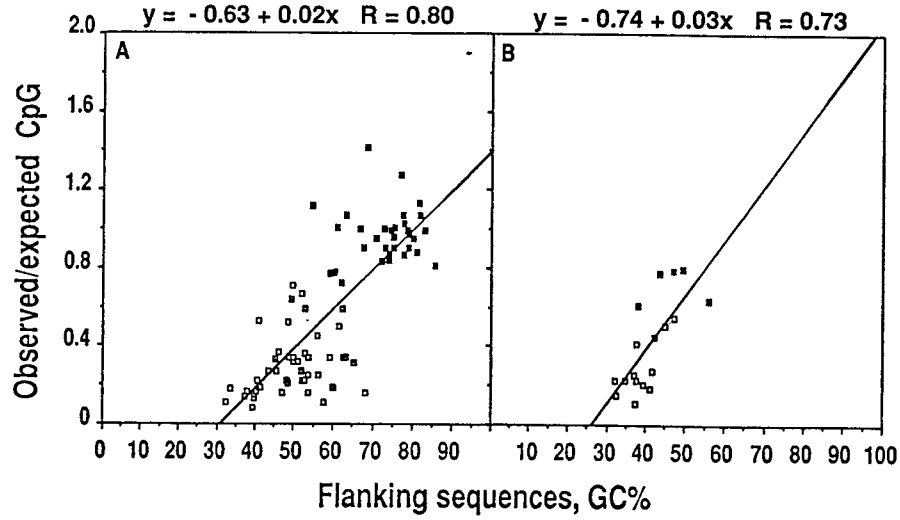

Fig. 6. Relationship between GC of flanking sequences and the o/e of CpG ratios. (A) Observed/statistically (o/s) expected CpG ratio of human flanking sequences is plotted against GC levels of flanking sequences. Other indications as in Fig. 3A. Blackened squares correspond to CpG islands. (B) Same plot as in A for 5′-flanking sequences of cold-blooded vertebrates. Other symbols as in Fig. 3A. Blackened squares correspond to CpG islands, open squares to flanking sequences that do not contain CpG islands.

TABLE III

List of cold-blooded vertebrate genes

Species	Genes*	Mnemonics
Chondrichthyes		
Heterodontus francisci	IgH constant region*	Hfighvc
Raja erinacea	IgH V-D-G region*	Reihcc
Torpedo californica	Na/K-ATPase	Fscatpbr
	Acetylcholine receptor	Fsachra
Torpedo marmorata	Creatine kinase	Fscck
Osteichthyes		
Cyprinus carpio	β-Actin*	Fsbactba
	γ-Crystallin	Fsbcrygm2
	α-Globin	Fsbaglob
Electrophorus electricus	Na^+ channel protein	Eescp01
Carassius auratus	Acetylcholine receptor	Caa3nnar
	Ras related protein*	Caras1
	Nicotinic receptor	Cagfna3
Salmo gairdneri	Growth hormone 2*	Fsbghrtb
	Metallothionein-B*	Fsbmetb
	Opiomelanocortin	Fsbpomc
	Histone H4	Fsbhis42b
	Protamine component II	Fsprc2a
	α-Tubulin	Fsbtubats
	Histone H2b	Fsbhis42b
	Vasotocin	Okvt1np
Onchorhynchus keta	Insulin	Fsbinssal
Ictalurus punctatus	Somatostatin-14	Ipsom1
Lophius americanus	Glucagon	Fsbafgi
	Somatostatin-i	Fsbsomi
Anarhichas lupus	Antifreeze III protein*	Alafp3a
Elops saurus	IgH V region	Esighvf
Fundulus heteroclitus	Lactate dehydrogenase-B	Fsbldhba
Amphibians		
Xenopus laevis	Myc oncogene	Xelmyc
	α-Tubulin*	Xeltuba
	Heat-shock protein	Xelhsp70a
	PDGF	Xelpdgfa
	Int-1 oncogene	Xelint1
	Yes protein	Xelyes
	HMG-coA reductase	Xelhmgcoa
	Opiomelanocortin	Xelpomc
	Ras protein	Xlrasx
	Protein S6 kinase	Xls6kiia
	Thyroid hormone receptor	Xlthya
Xenopus tropicalis	α-Globin*	Xethbba
Rana catesbeiana	Opiomelanocortin	Rcpomc
Reptiles		
Crotalus durissus	Crotoxin A	Cdcrota

* The listed genes are those used to construct Figs. 3B and 4B. Genes marked with asterisks are those used to construct Fig. 4B. Genes are specified by their products.

1991, and the relevant discussion), in which CpG islands have been most frequently observed. In agreement with this suggestion, we have shown that CpG increases in the 5′-flanking sequences of genes from cold-blooded vertebrates

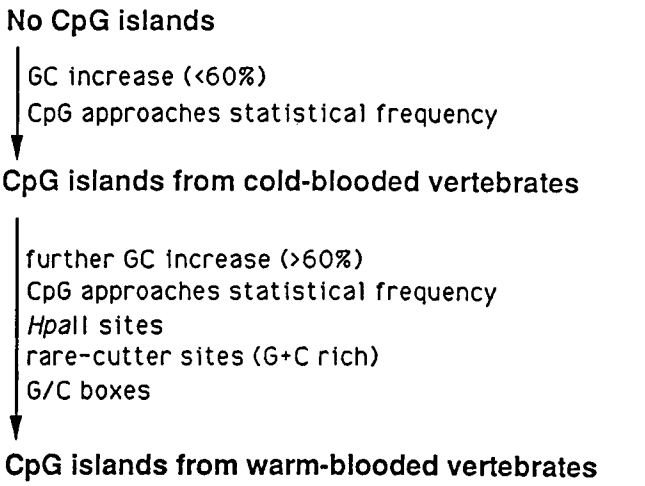

Fig. 7. Scheme for the formation and evolution of CpG islands in the genome of vertebrates.

as the sequences increase in GC, and that CpG points from CpG islands are aligned with those from 5′-flanking sequences that do not correspond to CpG islands (Fig. 5B). This indicates a direct correlation between GC increase and CpG increase, which is, in fact, the same as found for CpGs from human 5′-flanking sequences and CpG islands (see above). This correlation appears to be a very general one, since it is basically followed by exons and introns, as well.

To sum up, what is proposed here (Fig. 7) is that the genomes of cold-blooded vertebrates have scarce CpG islands which arose in association with local (regional) GC increases; these islands are still characterized by relatively low GC and CpG levels, and yet CpG frequency in the islands already approaches statistical values. This explanation, associating the origin of the primitive CpG islands of cold-blooded vertebrates with relatively high local GC levels, has the additional interest that it can account for the appearance of CpG islands in plants (Antequera and Bird, 1988), which have no common ancestor (carrying CpG islands) with vertebrates. When the formation of GC-rich isochores took place, this was accompanied by further GC increases and by the consequent increases in *Hpa*II sites, rare-cutter sites and GC/boxes in CpG islands. As expected, these CpG islands mainly were associated with the genes located in the GC-rich isochores.

In agreement with the above proposals, the unmethylated state of CpG doublets in CpG islands appears to be due to the fact that the increase in GC occurs through directional fixation of mutations (Bernardi and Bernardi, 1986; Bernardi et al., 1988), and that unmethylated CpG is stable in contrast to methylated CpG. The alternative proposals 'that CpG islands represent the last remnants of the long tracts of nonmethylated DNA that make up most of the invertebrate genomes', or that the α- or β-globin genes from chicken or mouse 'have lost ancestral islands during evolu-

Subak-Sharpe, J.H., Burk, R.R., Crawford, L.V., Morrison, J.M., Hay, J. and Keir, A.M.: An approach to evolutionary relationships of mammalian DNA viruses through analysis of the pattern of nearest neighbour base sequences. Cold Spring Harbor Symp. Quant. Biol. 31 (1966) 737–748.

Tazi, J. and Bird, A.P.: Alternative chromatin structure at CpG islands. Cell 60 (1990) 909–920.

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.