J Mol Evol (1991) 33:57–67

220

Journal of Molecular Evolution

Compositional Properties of Nuclear Genes from Cold-Blooded Vertebrates

Giacomo Bernardi* and Giorgio Bernardi

Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France

Summary. We have investigated the compositional properties of coding sequences from cold-blooded vertebrates and we have compared them with those from warm-blooded vertebrates. Moreover, we have studied the compositional correlations of coding sequences with the genomes in which they are contained, as well as the compositional correlations among the codon positions of the genes analyzed.

The distribution of GC levels of the third codon positions of genes from cold-blooded vertebrates are distinctly different from those of warm-blooded vertebrates in that they do not reach the high values attained by the latter. Moreover, coding sequences from cold-blooded vertebrates are either equal, or, in most cases, lower in GC (not only in third, but also in first and second codon positions) than homologous coding sequences from warm-blooded vertebrates; higher values are exceptional. These results at the gene level are in agreement with the compositional differences between cold-blooded and warm-blooded vertebrates previously found at the whole genome (DNA) level (Bernardi and Bernardi 1990a,b).

Two linear correlations were found: one between the GC levels of coding sequences (or of their third codon positions) and the GC levels of the genomes of cold-blooded vertebrates containing them; and another between the GC levels of third and first + second codon positions of genes from cold-blooded vertebrates. The first correlation applies to the genomes (or genome compartments) of all vertebrates and the second to the genes of all living organisms. These correlations are tantamount to a genomic code.

Key words: Genomes — Genes— Base composition — Vertebrates

Introduction

We have recently investigated the compositional patterns of the nuclear genomes of cold-blooded vertebrates, as far as the compositional distributions of the large DNA fragments that make up these genomes are concerned (Bernardi and Bernardi 1990a,b). We have shown that these patterns share a number of common properties (such as low intermolecular compositional heterogeneities, low CsCl band asymmetries, and, in most cases, small or undetectable amounts of satellite DNAs). These properties set the genomes of cold-blooded vertebrates apart from those of warm-blooded vertebrates. Moreover, the former (in contrast with the latter) cover a relatively wide spectrum of modal buoyant densities, which do not attain, however, the very high levels reached by the GC-richest fragments of DNAs from warm-blooded vertebrates.

In the case of fish genomes, which were more extensively studied, many different orders (and a number of families) were characterized by modal buoyant densities that were different in average values, as well as in their ranges. These differences are indicative of compositional transitions, which were shown (1) to be due essentially to directional base substitutions; (2) to be independent of geological time since the appearance of the order or family; and (3) to be independent of the numbers of species

* *Present address:* Hopkins Marine Station, Stanford University, Department of Biological Sciences, Pacific Grove, CA 93950-3094, USA

Offprint requests to: Giorgio Bernardi

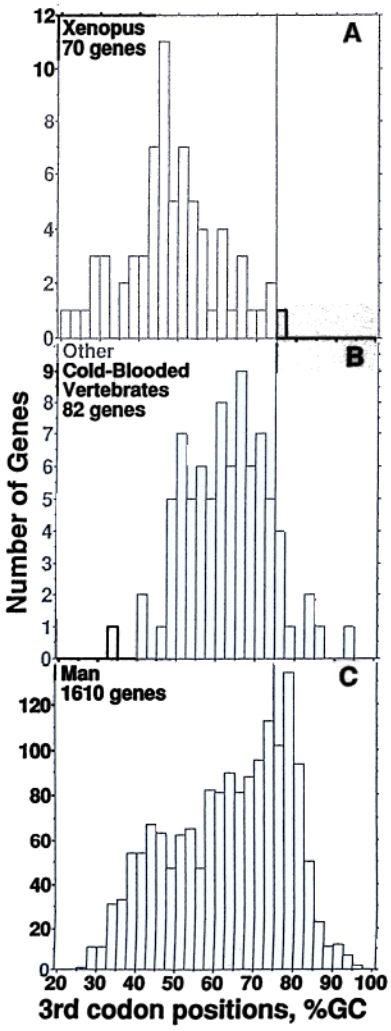

Fig. 1. Histograms of compositional distributions of third codon positions for genes of (A) *Xenopus laevis,* (B) Agnatha, fishes, amphibians other than *Xenopus,* reptiles, and (C) human. The vertical line at 75% GC is shown for the sake of comparison (see text).

within the orders under consideration (Bernardi and Bernardi 1990b).

In the present work, we have investigated (1) the compositional patterns of the genes and the genomes from cold-blooded vertebrates; (2) the compositional differences in homologous coding sequences from cold-blooded and warm-blooded vertebrates; (3) the compositional correlations that hold between coding sequences (or third codon positions) and genomes of cold-blooded vertebrates; and (4) the compositional correlation among different codon positions of genes from cold-blooded vertebrates. Moreover, we have compared the correlations mentioned under (3) and (4) above with those found in other organisms.

The results obtained support, at the gene level, the conclusions previously drawn at the whole genome (DNA) level, namely that the genomes of cold-blooded vertebrates are compositionally different from those of warm-blooded vertebrates; that they exhibit smaller, yet significant compositional differences among themselves; and that all these compositional differences are essentially due to directional base substitutions, as has already been suggested (Bernardi and Bernardi 1990a,b). Finally, we have demonstrated that the compositional correlations between coding sequences and their genomic environments are similar for all vertebrates, and that the compositional correlations among codon positions are the same for the genes of all living organisms.

Materials and Methods

Gene sequences were obtained from Release 65 (September 1990) of GenBank, from Release 24 (August 1990) of EMBL, and, in a few cases, from the literature. GC contents of genomes were taken from Bernardi and Bernardi (1990a).

The algorithm of Kanehisa was used to align nucleotide sequences. Only sequences that could be aligned over at least 70% of their length were analyzed further. Aligned sequences were used to calculate the percent identity, whereas complete sequences were used to calculate compositional differences in first + second and third codon positions.

Results

Table 1 presents the compositional properties of genomes and genes (exons, first, second, and third codon positions) from the cold-blooded vertebrates that were studied in these investigations. These comprised 51 species from 42 genera and 38 families. Data from the *Xenopus* genus (including not only *X. laevis,* but also *X. borealis* and *X. tropicalis*) are not presented, because they do not require specific comments.

Figure 1 shows histograms of the compositional distribution of third codon positions for genes from cold-blooded vertebrates. It should be noted that this is the most meaningful representation because of the wider scatter of GC values in third positions compared to the other codon positions or to exons. Data, shown separately for *X. laevis* (Fig. 1A) and for all other cold-blooded vertebrates of Table 1 (Fig. 1B), are compared with those for human genes (Fig. 1C). The latter are taken here as representing mammalian genes, and also genes of warm-blooded vertebrates in general, even if small differences in the histograms do exist between Myomorpha and other mammals, as well as between birds and mammals (Mouchiroud et al. 1987, 1988; Bernardi et al. 1988). It should also be noted that genes that were present more than once in *Xenopus,* or in the other cold-blooded vertebrates, were averaged and represented only once in order to avoid biasing histograms, which represented only a small number of

Table 1. GC levels of genomes and genes from cold-blooded vertebrates. Continued on pages 60–61

No.	Species	Genome	Gene	Exon	I	II	III
Agnatha							
1	*Petromyzon marinus*	(50.9)	LAL1	62.6	57.5	34.9	**95.3**
2			LAL2	66.1	65.7	42.2	**90.7**
3			A fibrinogen	59.5	56.7	54.1	67.8
4			B fibrinogen	61.9	55.2	45.2	**85.3**
5			G fibrinogen	59.9	56.4	38.8	**84.7**
6	*Myxine glutinosa*		Insulin	52.0	53.9	43.1	58.6
Chondrichthyes							
1	*Heterodontus francisci*		Ig Hc constant	51.7	47.1	45.3	62.4
2	*Scylliorhinus caniculus*	(46.2)	Protamine Z1	44.2	36.5	55.8	40.4
3			Protamine Z2	56.8	47.8	65.2	54.4
4	*Carcharinus plumbeus*	(44.7)	Ig VJC	54.2	49.6	52.5	60.3
5	*Raja erinacea*	45.0	Ig Vh	51.2	50.1	43.2	60.5
6	*Torpedo californica*	(42.1)	Acetylcholinesterase	54.5	73.2	51.9	73.1
7			Acetylcholinesterase receptor A	39.2	47.4	33.4	37.0
8			Acetylcholinesterase receptor G	44.4	48.3	32.5	52.2
9			ATPase A	44.5	51.5	39.7	42.2
10			ATPase B	42.0	48.7	36.2	40.9
11			VAMP 1	53.7	57.8	45.4	57.9
12			Postsynaptic protein	61.6	61.5	37.7	**85.6**
13			Postsynaptic protein	53.5	51.2	46.5	62.8
14			Creatine kinase	57.2	58.7	35.6	**77.3**
15			Intermediate filament (a)	45.9	53.8	32.6	51.1
16	*Torpedo marmorata*	42.1	Creatine kinase	56.5	58.9	35.4	**75.2**
17			Acetylcholinesterase	54.2	52.0	38.5	72.4
Osteichthyes							
18	*Cyprinus carpio*	37.2	Insulin	50.2	58.8	40.3	51.3
19			Urotensin I	55.9	61.0	34.3	72.6
20			Urotensin IIa	52.1	58.7	44.4	53.2
21			Urotensin IIg	51.1	56.4	45.2	51.6
22			Prolactin	55.1	53.6	41.7	57.1
23			Growth hormone	50.7	49.8	35.1	67.3
24			Crystallin γ-m1	50.7	38.0	40.2	73.8
25			Crystallin γ-m2	50.6	35.7	41.9	74.1
26			β-actin	54.9	53.5	41.3	70.0
27	*Carassius auratus*	37.9	Ependym II	48.8	55.8	39.1	51.7
28			Acetylcholine receptor	50.4	51.5	37.9	61.7
29			c-ras	49.6	52.3	32.3	63.8
30			Growth associated protein	57.0	66.3	40.2	64.5
31			Nicotinic receptor	48.6	49.9	37.0	58.7
32			Ig heavy chain V3	55.8	55.5	43.6	68.4
33			Ig heavy chain V5A	54.0	49.2	47.4	65.6
34	*Ctenopharyngobon idella*		β-actin	54.4	53.5	41.3	68.7
35	*Brachydanio rerio*	36.9	Homeobox ZF54	46.5	42.7	41.5	54.9
36			Engrailed	55.0	55.0	41.2	68.8
37			α-tropomyosin	53.8	62.4	26.6	72.1
38	*Astyanax fasciatus*	40.6	Visual pigment (b)	50.3	48.6	45.0	57.3
39	*Salmo salar*	44.4	Growth hormone	51.9	51.6	30.3	73.4
40			Homeobox gene	53.7	46.7	38.7	**76.1**
41	*Salmo gairdneri*	43.5	HSP 70A	57.8	57.5	39.1	**76.7**
42			HSP 70B	50.3	54.2	39.4	57.3
43			Protamine (TP21)	70.6	55.9	88.2	67.6
44			H4	60.5	62.5	46.1	73.1
45			H1	63.5	65.8	51.5	**83.1**
46			H2B	58.9	47.2	45.6	**84.0**
47			H2A	61.0	66.6	44.9	71.3
48			H3	61.6	60.6	48.2	**75.9**
49			MT A	54.6	27.4	74.2	60.6
50			MT B	51.9	26.3	73.8	52.4
51			HMGT 3	55.3	53.2	36.1	**76.6**
52			Growth hormone	53.1	52.1	30.3	**76.5**
53			Growth hormone 2	53.1	52.1	29.8	**77.3**
54			c-myc	51.7	52.8	41.9	60.2

Table 1. Continued

No.	Species	Genome	Gene	Exon	I	II	III
55			Cytochrome P450	54.2	55.7	36.1	70.7
56			Apopolysialoglyco protein	67.4	67.0	69.3	65.9
57			Complement P. C9	53.3	51.2	41.7	66.8
58	*Oncorhynchus keta*	44.7	Growth hormone	52.7	52.1	30.8	75.3
59			Proopiomelanocortin	59.8	55.1	44.8	79.4
60			Protamine	69.6	56.9	88.2	64.7
61			Gonadotropin I α	49.6	39.2	47.9	61.8
62			Gonadotropin I β	54.6	44.2	45.6	73.9
63			Gonadotropin II β	54.1	52.5	44.1	65.8
64			Insulin	53.4	58.5	36.8	65.1
65			Insulin	52.8	57.5	36.8	64.2
66			Calcitonin	52.1	51.1	43.1	59.1
67			Vasotocin	59.3	59.4	58.1	60.6
68			Glyco hormone	49.3	39.2	47.9	60.9
69			Melanin concentrating hormone	56.6	51.8	47.4	70.7
70			Prolactin	54.1	54.1	42.6	65.8
71			Prolactin	54.8	52.9	43.0	68.9
72	*Oncorhynchus kisutch*	44.5	Growth hormone	52.8	52.6	30.8	74.9
73			Insulin-like growth factor	55.5	53.6	52.0	61.0
74	*Oncorhynchus masou*		Apopolysialoglyco protein	65.8	61.1	70.3	66.0
75	*Oncorhynchus tschawytscha*		Gonadotropin	54.6	51.8	44.8	67.2
76			Melanin concentrating hormone	55.4	51.9	42.8	71.4
77			Melanin concentrating hormone	57.4	54.1	46.6	71.4
78	*Pagurus major*		Growth hormone	53.0	53.4	38.7	68.5
79	*Ictalurus punctatus*		Somatostatin 14	63.7	64.3	49.6	77.4
80			Somatostatin 22	56.6	57.6	46.3	66.0
81			Ig μ chain	46.1	46.5	38.8	53.0
82	*Lophius americanus*	40.6	Insulin	58.1	61.6	40.2	62.7
83			Glucagon I	52.6	52.0	38.4	67.2
84			Glucagon II	49.6	48.8	39.8	60.2
85			Preprosomatostatin I	65.9	63.1	46.7	87.7
			Preprosomatostatin II	60.0	62.0	45.2	73.1
87	*Pseudopleuronectus americanus*		Antifreeze protein	68.6	67.4	69.9	76.4
88	*Anarhichas lupus*		Antifreeze protein	54.3	52.8	38.2	71.9
89	*Macrozoarces americanus*		Antifreeze protein	53.8	51.2	37.5	72.8
90	*Paralichthys olivaceus*		Growth hormone	48.2	51.8	34.6	58.1
91			Somatolactin	52.1	52.1	33.6	70.2
92	*Catostomus commersoni*		Isotocin	56.7	60.0	55.5	54.9
93			Vasotocin I	54.4	53.6	58.8	50.9
94			Vasotocin II	59.9	60.9	55.1	63.5
95	*Elops saurus*		Ig H-chain	45.9	46.2	42.9	48.9
96	*Xiphophorus maculatus*	39.5	Oncogene x-egfrB	52.2	61.9	30.9	63.7
97			Melatonin receptor kinase	56.8	55.3	42.2	73.2
98	*Fundulus heteroclitus*	40.4	Lactate dehydrogenase	63.5	59.7	38.8	92.2
99	*Oreochromis niloticus*	41.7	Growth hormone (c)	52.1	51.7	37.1	67.3
100			Prolactin	54.4	52.6	43.7	67.1
101	*Seriola quinqueradiata*		Growth hormone	50.9	52.2	36.6	63.9
102	*Thunnus thynnus*	(37.0)	Growth hormone	52.8	54.2	38.6	65.8
103	*Hemitripterus americanus*		Antifreeze protein	51.6	53.0	53.6	48.2
Amphibians							
1	*Pleurodeles waltlii*	45.6	α actin	52.1	45.2	42.9	68.3
2			Larval α globin	59.6	62.3	43.4	73.5
3			Adult α globin	45.8	34.1	50.3	52.9
4	*Bufo japonicus*	(44.5)	Mesotocin	50.5	42.9	53.2	55.5
5			Vasotocin	58.5	55.1	55.6	69.0
6	*Rana temporaria*	44.2	α tropomyosin	48.8	62.1	26.3	57.9
7			α A crystallin	56.0	54.0	41.4	72.6
8			β 23 crystallin	48.6	50.7	40.7	54.2
9	*Rana pipiens*	(44.2)	Olfactory specific protein	42.1	47.2	31.5	47.7
10			Ranatensin	48.6	38.2	42.2	55.4
11			γ2-crystallin	45.7	47.1	35.3	54.7
12	*Rana catesbeiana*	(44.2)	Growth hormone	44.0	50.5	32.8	48.6
13			Proopiomelanocortin	46.9	47.9	37.3	55.5

Table 1. Continued

No.	Species	Genome	Gene	Exon		II	III
14			Apoferritin L	50.8	56.9	32.8	62.7
15			Apoferritin M	51.6	59.3	27.7	67.8
16			Apoferritin H	52.2	59.9	27.7	68.9
17			β-globin	47.5	62.1	32.5	48.0
18	*Phyllomedusa sauvagei*		Demorphin	37.5	51.0	30.8	30.8
19			Demorphin	37.0	51.5	28.8	31.8
20	*Phyllomedusa bicolor*		Deltorphin	40.1	57.0	25.4	37.7
Reptiles							
1	*Aipysurus laevis*		PLA2	48.7	46.2	47.5	52.4
2			Toxin D	44.7	40.3	45.2	48.8
3			Toxin B	44.7	39.1	46.4	48.8
4	*Laticauda semifasciata*		Erabutoxin a	46.4	39.3	47.6	52.4
5			Erabutoxin b	47.0	41.0	48.2	51.8
6	*Laticauda laticaudata*		PLA2	48.6	48.0	46.5	51.3
7	*Notechis scutatus*		PLA2	49.6	46.5	50.0	52.1
8	*Crotalus durissus*		Crotoxin A	54.9	50.4	46.8	67.6
9			Crotoxin B	53.9	43.2	48.2	70.5
10	*Naja naja*		Acetylcholine receptor A	44.2	42.3	38.4	51.9
11	*Vipera ammodytes*		PLA 2	53.3	43.9	45.3	70.5
12			Ammodytoxin B	54.2	44.2	45.0	73.2
13	*Natrix tessellata*	(42.6)	Acetylcholine receptor	45.2	44.3	38.4	52.9
14	*Bothrops atrox*	(43.8)	Batroxobin	46.1	48.4	43.0	56.9
15	*Caiman crocodylus*		IG H	59.3	55.6	50.4	71.8

GC levels of genomes are from Bernardi and Bernardi (1990a). Value in parentheses concern species from the same genera (or family for sample 1, Agnatha) other than those listed. Values for samples 1 (Agnatha) and 102 (Osteichthyes) are from Vanyushin et al. (1973). Some mnemonics are not yet available; the pertinent references are (a) Frail et al. (1990); (b) Yokoyama and Yokoyama (1990); (c) Rentier-Delrue et al. (1989). Bold figures in third codon positions concern values higher than 75% GC

genes. This was not done for human genes, because doing so would not have led to any significant change in the histogram, due to the very high number of genes represented.

As previously observed (Mouchiroud et al. 1987; Bernardi et al. 1988), the compositional distribution of third codon positions of *Xenopus* genes is remarkably shifted toward lower GC values relative to that of human genes. In fact, only 1.5% of *Xenopus* genes exhibit values higher than 75% GC [which is considered to be the lower limit of genes located in the GC-richest isochore family of the human genome (Mouchiroud et al. 1991)], whereas 27% of human genes do so. This situation found in *Xenopus* is also basically true for other cold-blooded vertebrates. In this heterogenous array of vertebrates, only 11% of the genes are above 75% GC content.

Figure 2A shows a histogram of genome compositions (essentially derived from Bernardi and Bernardi 1990a,b) for cold-blooded vertebrates belonging to 26 species from 21 genera and 18 families. The number of genera was taken into consideration instead of the number of species, in order to avoid a bias due to the fact that species within a genus exhibit the same GC level. A CsCl profile of human DNA is superimposed on the histogram. The histogram is centered on a value that is slightly higher than the modal GC value of human DNA (as ex-

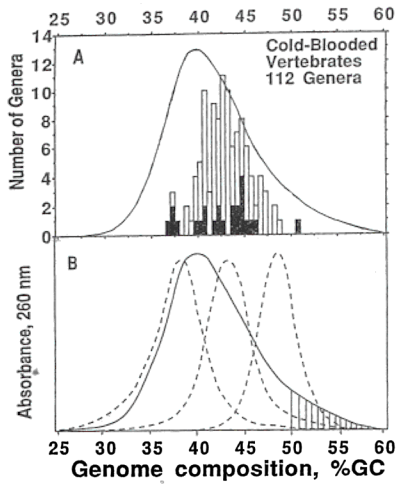

Fig. 2. (**A**) Histogram showing the number of genera of cold-blooded vertebrates characterized by given GC levels [from Bernardi and Bernardi (1990a); except for tuna fish and lamprey, derived from Vanyushin et al. (1973)]; black bars concern genera explored in their genes in the present work. A CsCl profile of human DNA is superimposed on the histogram. (**B**) CsCl profiles of DNAs from cold-blooded vertebrates (broken lines) are compared with the CsCl profile of human DNA (solid line); the hatched region of the profile has the highest gene concentration (Bernardi et al. 1985; Mouchiroud et al. 1991). DNAs from cold-blooded vertebrates were from *Brachydanio rerio* (36.9 %GC), *Sphyraena barracuda* (42.6 %GC), and *Aphyosemion australe* (48.1 %GC).

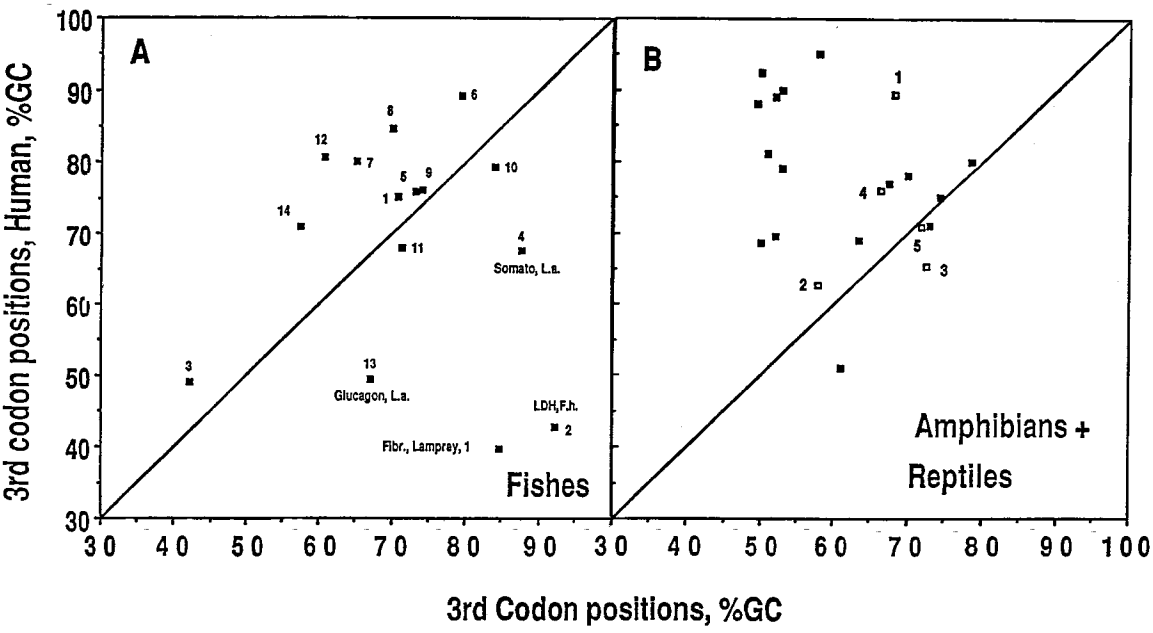

Fig. 3. Plot of GC levels of third codon positions from human genes against GC levels of third codon positions of homologous genes (**A**) from fishes and Agnatha; and (**B**) from *Xenopus laevis* [closed squares; from Bernardi et al. (1988)], other amphibians, and reptiles (open squares; see Table 2 for the numbering).

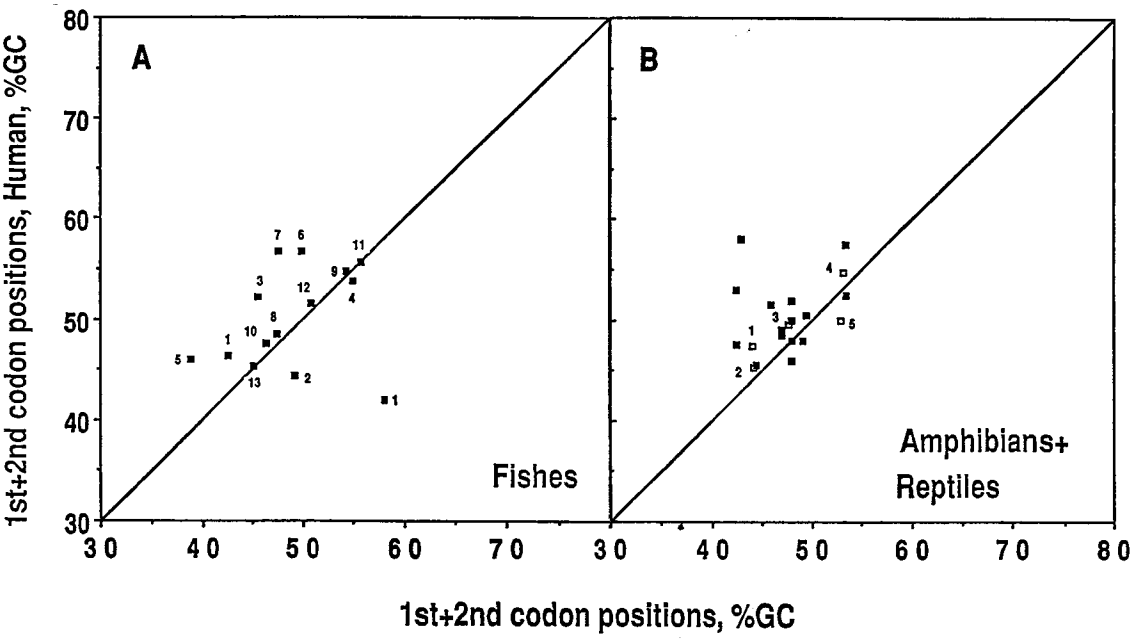

Fig. 4. Plot of GC levels of first + second codon positions from human genes against GC levels of first + second codon positions of homologous genes (**A**) from fishes and Agnatha and (**B**) from *Xenopus laevis* [closed squares; from Bernardi et al. (1988)], other amphibians, and reptiles (open squares; see Table 2 for the numbering).

pected from previous work of Bernardi and Bernardi 1990b). Figure 2A also shows that the compositional distribution of the coding sequences of the genomes investigated here exhibits a satisfactory coverage of the spectrum of genome compositions.

Figure 2B compares the CsCl profiles of DNAs from three fishes (*Brachydanio rerio, Sphyraena barracuda,* and *Aphyosemion australe*) characterized by low, intermediate, and high buoyant densities (corresponding to 36.9% GC, 42.6% GC, and 48.1% GC, respectively) with that obtained for human DNA. All the fish DNA profiles are contained within the broad GC profile of human DNA.

The relatively small number of genes available for cold-blooded vertebrates other than *Xenopus* prompted a rigorous check of the reality of the differences between the compositional patterns of coding sequences from cold- and warm-blooded vertebrates. This was done by comparing GC levels of third codon positions from fishes (Fig. 3A) and from

Table 2. Compositional properties of homologous genes from human and cold-blooded vertebrates

No.	Gene	Human	Cold-blooded vertebrates	% i[a]	% GC M I + II	% GC CBV I + II	% GC M III	% GC CBV III
Agnatha								
1	γ-Fibrinogen	HUMFBRG	FSAFBRG	66.2	42.0	47.6	39.7	84.7
Fishes								
	Growth hormone	HUMGH	FSBGHSAM	59.5				
			FSBGHCOHO	59.6				
			FSBGH1	59.5	46.3	42.6	75.2	70.8
			PAGPREGH	69.4				
			FSBFGH	70.9				
2	Lactate dehydrogenase	HUMLDHB1	FSBLDHBA	76.1	44.3	49.2	42.7	92.2
3	ATPase	HUMATPSY1	FSCATPA	58.6	52.2	45.6	48.9	42.2
4	Somatostatin	HUMSOMI	FSBSOMI	61.7	53.8	54.9	67.5	87.7
5	Crystallin γ	HUMCRYGBC	FSBCRYGM2	64.3	46.0	38.8	76.1	74.1
6	Proopiomelanocortin	HUMPOMC	FSBPOMC	65.6	56.7	49.9	89.2	79.4
7	Insulin	HUMINSO1	FSBINSSAL	70.4	56.7	47.6	80.1	65.1
8	β-actin	HUMACCYBA	FSBACTBA	88.4	48.5	47.4	84.6	70.0
9	Histone H4	HUMHISH4	FSBHIS42B	83.5	54.8	54.3	76.0	73.1
10	Histone H2B	HUMHISH2B	FSBHIS42B	81.9	47.6	46.4	79.4	84.0
11	Histone H2A	HUMHISH2A	FSBHIS2A3	78.7	55.7	55.7	68.0	71.3
12	Metallothionein 2A	HUMMT2A	FSBMETALA	72.2	51.6	50.8	80.6	60.6
13	Glucagon	HUMGG	FSBAFGI	71.0	45.3	45.2	49.4	67.2
14	Visual pigment	HUMCNPGn	[b]				71.0	57.3
Amphibians and reptiles								
1	α-actin	HUMSAACT	SMRAACTR	88.3	47.4	44.1	89.4	68.3
2	α-tropomyosin	HUMTROPA2	RANATROA	84.3	45.3	44.2	62.8	57.9
3	α-crystallin	HUMCRYABA	RANCRYAA	63.6	49.6	47.7	65.3	72.6
4	Histone H4	HUMHISH4	XLHIS4B	80.8	54.8	53.2	76.0	66.3
5	Immunoglobulin	HUMIGHVA	CRCIGHVA	69.0	50.0	52.9	70.9	71.8

[a] The % identity of compared aligned sequences. % GC are given for the whole sequences
[b] Yokoyama and Yokoyama (1990)

amphibians and reptiles (Fig. 3B) with GC levels of third codon positions from homologous genes of warm-blooded vertebrates. Table 2 provides a list of the genes compared, their GC levels in first + second and in third codon positions, and the percent identity of compared sequences; *Xenopus* data used in Fig. 3B are not presented in Table 2 because they were taken from Bernardi et al. (1988).

In the case of fishes, points fell either on the straight line having a unity slope and passing through the origin (indicating identical GC values in the third codon positions of compared genes) or, in the majority of cases, above it. Three points (corresponding to the genes for somatostatin, glucagon, and lactate dehydrogenase) fell below the line, as also did the point for the fibrinogen gene of lamprey, a jawless fish (Agnatha). In the case of amphibians and reptiles, most points fell above the diagonal line, some on the line, and only one slightly below it.

A similar plot comparing first + second codon positions of genes from human and from cold-blooded vertebrates is shown in Fig. 4A and B. In this case too, points either fell on the diagonal line or, in the majority of cases, above it. The lactate dehydrogenase and, much more so, the lamprey fibrinogen points fell, however, below the line.

Figure 5 shows a plot of GC levels of third codon positions of genes from cold-blooded vertebrates against GC levels of the corresponding genomes. GC levels from genes of any given species were averaged. Values from genes of Agnatha, Chondrichthyes, amphibians, and reptiles fell on a line exhibiting a very high correlation coefficient ($R = 0.94$) and a slope ($s = 3.3$) close to that shown when GC levels of third codon positions from human genes are plotted against the GC levels of the large DNA fragments in which the genes were localized (Aïssani et al. 1991). Points from Osteichthyes either fell on a parallel line, which was shifted to the left by about 2% GC relative to that just discussed (points 38, 39, 41, 58, 72), or exhibited a more important shift to the left (points 18, 27, 102, 35, 96, 82, 99). These points concerned DNAs, (from several cyprinids, from tuna fish, and from *Lophius americanus*), characterized by very low GC levels. The lactate dehydrogenase point from *F. heteroclitus* (point 96) also showed a strong deviation.

Figure 6 shows a plot of GC levels of third codon

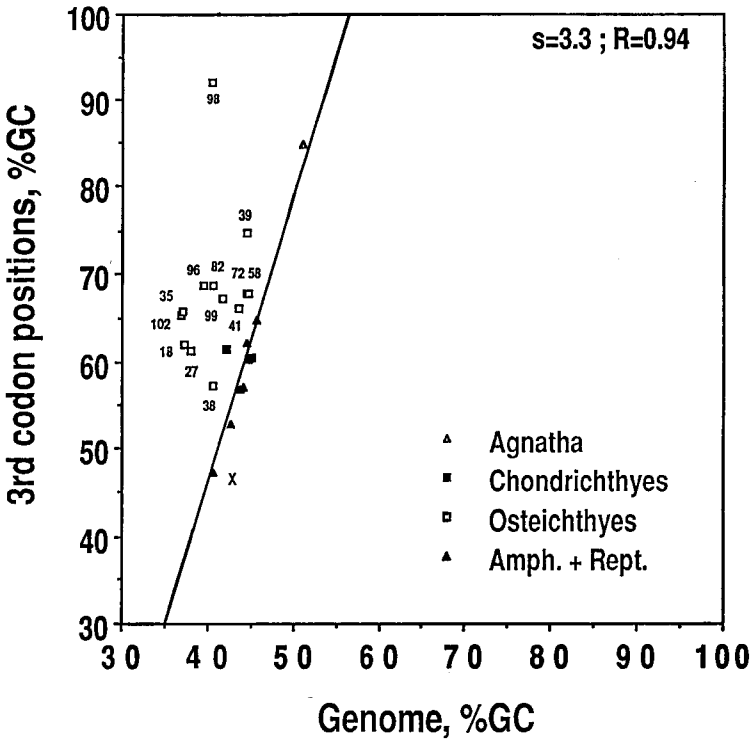

Fig. 5. Plot of GC levels of third codon positions of genes (averaged per genome) against GC levels of genomes from cold-blooded vertebrates. Closed squares refer to genes from Chondrichthyes, open squares to genes from Osteichthyes, closed triangles to genes from amphibians and reptiles, open triangles for Agnatha.The straight line is the correlation found between GC levels of third codon positions of genes from cold-blooded vertebrates (except for Osteichthyes) and GC levels of the corresponding genomes. (s stands for slope and *R* for the correlation coefficient; see Table 1 for the numbering of certain points, X stands for *Xenopus*).

positions against GC levels of first + second codon positions for genes from fishes, amphibians, and reptiles. The linear correlation found for human genes (Aïssani et al 1991) is also shown for the sake of comparison. A group of strongly deviating points (surrounded by a circle) from *Xenopus* was observed for caerulin genes.

Figure 7 shows a plot of GC levels of third codon positions against GC levels of first + second codon positions for genes from a vast array of organisms (data are from Bernardi and Bernardi 1985, 1986). The least-square line through the points is characterized by a correlation coefficient of 0.70 and by a slope of 1.7.

Discussion

Compositional Patterns of Third Codon Positions from Cold-Blooded Vertebrates

These patterns (Fig. 1A, B) are distinctly different from those of warm-blooded vertebrates (Fig. 1C), in that they do not reach the high GC values attained by the latter. Moreover, the relative amounts of genes showing values higher than 75% in third codon positions is much lower. This is not only true for *Xenopus* [in agreement with previous findings (Mouchiroud et al. 1987; Bernardi et al. 1988)], but also for the other cold-blooded vertebrates. It should be noted that, even if the sample of genes from cold-blooded vertebrates studied is small, it is representative, in that those studied cover the whole spectrum of genome compositions (see Fig. 2A).

The compositional patterns of the genomes of cold-blooded vertebrates, previously defined at the whole genome (DNA) level (Bernardi and Bernardi 1990a), are, therefore, paralleled by the compositional patterns of the genes from the same genomes, in that lower GC levels predominate in genes as well. This speaks against the possibility that compositional differences at the whole genome (DNA) level are just due to differences in intergenic sequences and in the repeated interspersed sequences contained in them (because of insertion, deletion, and amplification phenomena). This conclusion, already drawn on the basis of more limited data correlating third codon position GC and genome GC (Bernardi and Bernardi 1990b), is confirmed by the results on homologous genes (see below).

The CsCl profiles of Fig. 2B show that the compositional distributions of DNA fragments from all cold-blooded DNAs are not only narrow, but are also contained within the broad GC profile exhibited by human DNA, shown here as a reference for warm-blooded vertebrates. In fact, even the GC-richest DNAs from cold-blooded vertebrates barely attain the GC levels (above 50% GC) of DNA components from warm-blooded DNAs which have the highest gene concentration (Bernardi et al. 1985, 1988; Mouchiroud et al. 1987, 1991). It is obvious, therefore, that coding sequences being compositionally correlated with DNAs from cold-blooded vertebrates (see below), the genes of the latter will gen-

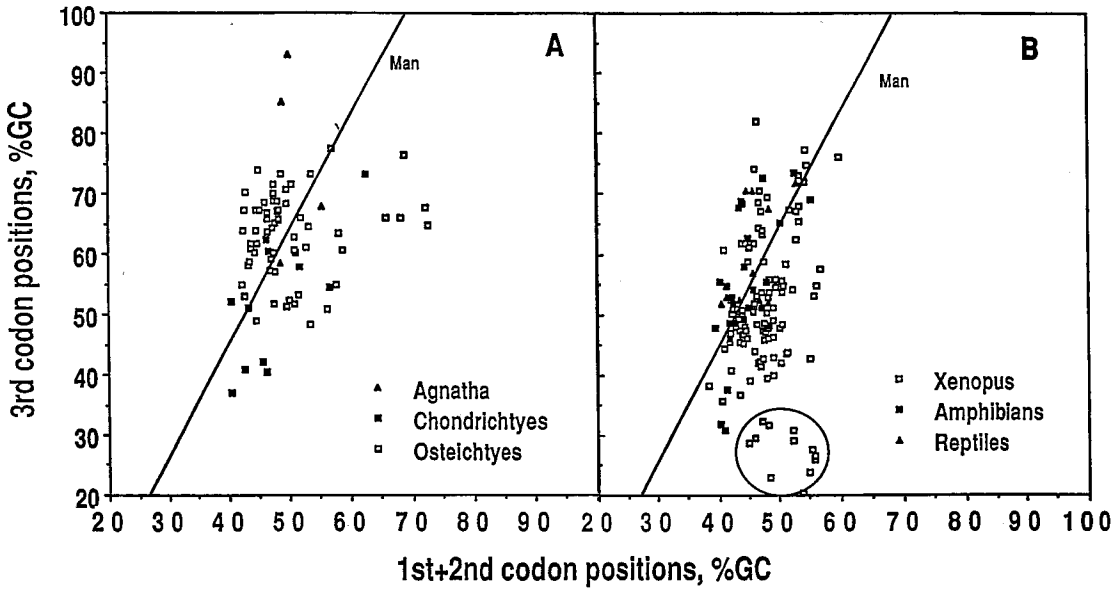

Fig. 6. Plot of GC levels of third codon positions from genes of (A) Agnatha, Chondrichthyes (closed squares), and Osteichthyes (open squares); (B) *Xenopus* (open squares; circled points correspond to caerulin genes), other amphibians (closed squares), and reptiles (closed triangles) against GC levels of first and second codon positions from the same genes. The straight-line corresponds to the linear relationships found in human genes [from D'Onofrio et al. (1991).

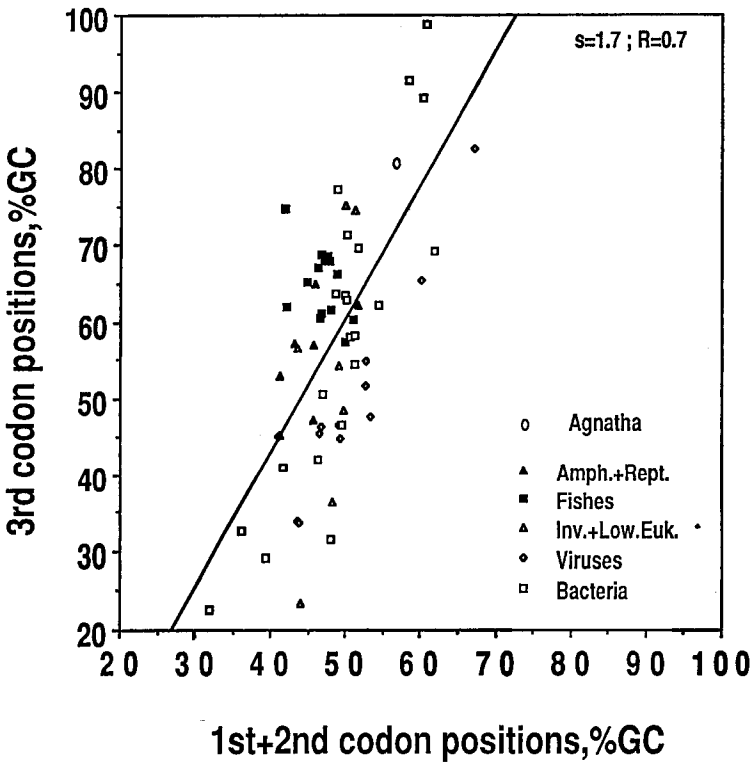

Fig. 7. GC levels of third codon positions from genes of cold-blooded vertebrates, invertebrates, lower eukaryotes, viruses, and bacteria are plotted against GC levels of first + second codon positions (s stands for slope and *R* for the correlation coefficient). Data are from Bernardi and Bernardi (1985, 1986). The least-square line obtained is very close to that found for human genes (D'Onofrio et al. 1991).

erally be lower in GC than the homologous genes of warm-blooded vertebrates. On the other hand, it is understandable that some genes from the GC-richest genomes of cold-blooded vertebrates (like lamprey) may be higher in GC content than genes from the GC-poorer genome compartments of warm-blooded vertebrates (see the following section).

Comparison of Homologous Coding Sequences from Warm- and Cold-Blooded Vertebrates

This comparison shows that most coding sequences from warm-blooded vertebrates exhibit higher GC levels in all codon positions, but especially in third

codon positions, relative to their homologous coding sequences in cold-blooded vertebrates (Figs. 3 and 4). A small number of the latter do not show any compositional difference, and four exceptional cases showing an opposite trend were also found. As far as the latter are concerned. It should be noted that this number has only increased by one (the lactate dehydrogenase point) since our previous survey (Perrin and Bernardi 1987), whereas the number of points above the diagonal has increased much more.

These exceptional cases were interpreted (Perrin and Bernardi 1987) as due to GC increases occurring in part or in the totality of the genomes of cold-blooded vertebrates. An alternative explanation, not exclusive of the former, is that GC increases occurred, before or after the divergence of warm-blooded vertebrates, in evolutionary branches that did not lead to the latter.

These findings do not simply represent an extension of similar results (Bernardi et al. 1988), which concerned coding sequences from *Xenopus* and human. Because the *Xenopus* genome is characterized by a low GC level [40.9% (Thiery et al. 1976)] and a very high compositional homogeneity, it could be argued that the differences found might not concern all cold-blooded vertebrates. On the other hand, previous findings by Perrin and Bernardi (1987) on cold-blooded vertebrates other than *Xenopus* were too limited to be of general use. The present data provide, therefore, the first evidence for the general existence of these differences.

The results of Figs. 3 and 4 also demonstrate that the transition between cold-blooded and warm-blooded was accompanied by directional nucleotide changes leading, in most cases, to increases in GC levels not only in third, but also in first + second codon positions. The consequence of the transition, therefore, is not only a change in codon usage, but also a change in amino acid composition. Such changes concern (D'Onofrio et al. 1991) the amino acids encoded by G and/or C in first and second codon positions (alanine, arginine quartet codons, glycine, and proline) and those encoded by A and/or T in first and second codon positions (asparagine, isoleucine, leucine duet codons, lysine, methionine, phenylalanine, and tyrosine). The general consequences of such changes have already been discussed (Bernardi and Bernardi 1986) and will be further commented upon elsewhere.

The changes under consideration concern genes that are present in what has been called the neogenome of warm-blooded vertebrates (Bernardi 1989), namely in the genome compartments that underwent GC increases (in both genes and intergenic sequences) during the transition from cold- to warm-blooded vertebrates. Other genes that did not change their composition are those present in the paleogenome of warm-blooded vertebrates (Bernardi 1989), the genome compartment that did not undergo a base composition change. Interestingly, the four genes showing a decrease in GC in warm-blooded vertebrates only did so in third codon positions, except for the lamprey fibrinogen gene and, to a much lower extent, the lactate dehydrogenase gene of *F. heteroclitus*. This suggests that, in cold-blooded vertebrates, such genes are located in isochores of higher GC content compared to warm-blooded vertebrates (see Perrin and Bernardi 1987, for further discussion on this point).

The results on homologous coding sequences confirm previous conclusions drawn on the basis of comparisons at the whole genome (DNA) level (Bernardi and Bernardi 1990b), namely that compositional changes occurred during the evolution from cold- to warm-blooded vertebrates, and also that changes essentially consisted of directional point mutations.

Compositional Correlation between GC Levels of Third Codon Positions and GC Level of Genomes from Cold-Blooded Vertebrates

This correlation (Fig. 5) is very close to those previously found for warm-blooded vertebrates (Bernardi et al. 1985; Aïssani et al. 1991). As far as the leftward shift of some data points for fishes is concerned, this might be due to DNA methylation for the points showing a small shift and a parallelism to the general relationship. Indeed, 5-methylcytosine (which lowers buoyant densities and leads, therefore, to underestimating genome GC) tends to be higher in DNAs from fishes than in DNAs from warm-blooded vertebrates (Vanyushin et al. 1970, 1973). This difference cannot, however, be the whole explanation for the most deviant points. In these cases, two possible explanations are that the genes investigated were localized in isochores having a higher GC level than the whole genome, or that the intergenic sequences of these genomes are lower in GC (relative to coding sequences) than those of other fishes. Further investigations should clarify this issue.

It should be stressed that although linear correlations are found in all eukaryotes explored so far, correlations may be characterized by different slopes and intercepts, as exemplified by the case of plant genomes (Montero et al. 1990). In this case, the slope of the compositional relationship between third codon positions and the DNA fractions containing the corresponding genes is higher than that of vertebrates.

Compositional Correlation between Third and First + Second Codon Positions of Genes from Cold-Blooded Vertebrates

Previous investigations (Bernardi and Bernardi 1985) showed that the GC levels of third codon positions of 302 coding sequences from 49 genomes ranging from bacteria to human exhibited the same positive correlation (with a slope of about 2) with the GC levels of the corresponding coding sequences. This indicated a linear relationship between the GC levels of third codon positions and those of first + second codon positions. A linear relationship was also found by Wada and Suyama (1985) for 161 prokaryotic genes, mostly from *Escherichia coli* bacteriophages; this point will be commented upon in detail elsewhere. Subsequent work (Bernardi and Bernardi 1986), using a slightly expanded sample of coding sequences, further stressed the existence of correlations among the GC levels of the three codon positions within and between genomes or genome compartments, as well as the universal nature of the relationships and the effects on the amino acid composition of the encoded proteins.

The present work has shown that the compositional correlation between third and first + second codon positions of genes from cold-blooded vertebrates is the same as that found (D'Onofrio et al. 1991) in human genes (Fig. 6) and that if all the data from Bernardi and Bernardi (1985, 1986) are plotted, one finds that they all fit the same correlation (Fig. 7; the bacterial and human lines are almost coincident with that shown).

From a general viewpoint, this work leads, therefore, to two major conclusions, namely that common compositional correlations exist (1) between third codon positions of genes from all vertebrates and the genomic environments in which they are located [as was tacitly assumed in previous investigations (Bernardi et al. 1985)], and (2) among codon positions of genes from all living organisms. These rules amount to a genomic code (see Bernardi 1990, and paper in preparation, for further discussion on this point).

Acknowledgments. We thank our colleagues Dominique Mouchiroud (Laboratoire de Biométrie, U.A. 243, Université Claude Bernard, Lyon I, 69622 Lyon, France) for the data of Fig. 1A and C, and Brahim Aissani, Giuseppe D'Onofrio, and Gabriel Macaya for very useful discussions.

References

Aïssani B, D'Onofrio G, Mouchiroud D, Gardiner K, Gautier C, Bernardi G (1991) Compositional correlations between genes and isochores in the human genome. J Mol Evol (in press)

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 évolution. 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 patterns in the nuclear genome of cold-blooded vertebrates. J Mol Evol 31: 265–281

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

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

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

Frail DE, Mudd J, and Merliev JP (1990) Nucleotide sequence of an intermediate filament cDNA from *Torpedo californica.* Nucleic Acids Res 18:1910

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

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

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

Rentier-Delrue F, Swennen D, Prunet P, Lion M, Martial JA (1989) Tilapia prolactin: molecular cloning of two cDNAs and expression in *Escherichia coli.* DNA 8:261–270

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

Vanyushin BF, Tkacheva SG, Belozersky AN (1970) Rare bases in animal DNA. Nature 225:948–949

Vanyushin BF, Mazin AL, Vasilyev VK, Belozersky AN (1973) The content of 5-methylcytosine in animal DNA: the species and tissue specificity. Biochim Biophys Acta 299:397–403

Wada A, Suyama A (1985) Third letters in codons counterbalance the (G+C)-content of their first and second letters. FEBS Letters 188:291–294

Yokoyama R, Yokoyama S (1990) Isolation, DNA sequence and evolution of a color visual pigment gene of the blind cave fish *Astyanax fasciatus.* Vision Res 30:807–816