ELSEVIER

Gene 205 (1997) 109–118

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

Evolutionary changes in CpG and methylation levels in the genome of vertebrates

Kamel Jabbari [a], Simone Cacciò [a,1], Jean Paul Païs de Barros [b], Jean Desgrès [b], Giorgio Bernardi [a,*]

[a] *Laboratoire de Génétique Moléculaire, Institut Jacques Monod, 2 Place Jussieu, 75005 Paris, France*
[b] *Laboratoire de Biochimie Médicale, Faculté de Médecine et Centre Hospitalier Universitaire, 7 Bd Jeanne d'Arc, 21033 Dijon, France*

Accepted 31 July 1997

Abstract

We have analysed the levels of 5-methylcytosine (5mC) in DNAs from 42 vertebrates, and compiled, including data from literature, a table of genomic 5mC and GC levels (as well as the available *c*-values, i.e., the haploid genome sizes) of 87 species from all vertebrate classes. An analysis of the data indicates that (i) two positive correlations hold between the 5mC and GC levels of the genomes of fishes/amphibians and mammals/birds, respectively; (ii) the genomes of fishes and amphibians are, on average, about twice as methylated as those of mammals, birds and reptiles, this difference being unrelated to the amounts of repetitive DNA sequences; (iii) the 5mC and CpG observed/expected values show no overlap between the two groups of vertebrates and suggest the existence of two equilibria. The transition separating the two equilibria appears to have taken place at the time of appearance of reptiles. Its possible cause(s) and its implications are discussed. © 1997 Elsevier Science B.V.

Keywords: CpG dinucleotide; Isochore; Vertebrate evolution

1. Introduction

It is well known that vertebrate DNAs show a characteristic deficiency in the frequency of CpG doublets, whereas all other doublets are present at about the frequency expected from a random distribution of nucleotides (Josse et al., 1961; Swartz et al., 1962). Two explanations were proposed to account for this outstanding feature of the vertebrate genome, the so-called CpG shortage.

The first explanation (Subak-Sharpe et al., 1966) was based on the observation that the CpG shortage is also exhibited by the genomes of small vertebrate viruses (like polyoma and SV40), which use essentially all their genetic information for directing protein synthesis. It was proposed (Subak-Sharpe, 1967, Subak-Sharpe et al., 1974) that the genomes of these viruses derived from polypeptide-specifying DNA of the host cells and, therefore, exhibited the same CpG shortage, which was visualized as reflecting constraints from the translation apparatus. In agreement with this proposal, CpG shortage was absent in tRNA, 5sRNA and rRNA genes. The lack of CpG shortage in the genome of intermediate and large viruses (such as adenovirus and herpes simplex virus, respectively) was attributed to their capacity to modify the host translation apparatus. Since only a very small proportion of vertebrate DNA is actively involved in protein coding and since the CpG shortage was found in all compositional fractions of guinea-pig DNA as obtained by gradient centrifugation, by differential renaturation or from dispersed and condensed chromatin, it was further concluded that 'the bulk of vertebrate DNA derives from, and maintains the gross sequence characteristics of polypeptide-specifying DNA' (Russell et al., 1976).

While the logic of the hypothesis was compromised by the discovery that rabbit (Salser, 1977) and human (Forget et al., 1979) α-globin mRNAs carry a large number of CpG doublets, and by the uniform levels of CpG in both coding and non-coding sequences from the human α and β gene clusters (Lennon and Fraser, 1983), subsequent investigations (Bernardi et al., 1985; Bernardi, 1985) definitely demonstrated that CpG short-

* Corresponding author. Tel.: +33 1 43295824; Fax: +33 1 44277977.
[1] Present address: Istituto Superiore di Sanità, Viale Regina Elena 299, 00161 Roma, Italy

Abbreviations: Br84, 8-bromoguanosine; 5mC, 5-methylcytosine; GC, molar fraction of guanine+cytosine in DNA; HPLC, high-performance liquid chromatography..

PII S0378-1119(97)00475-7

age is not related to polypeptide coding. In fact, CpG shortage (i) is strong in GC-poor genes, but becomes increasingly weaker in GC-rich genes from warm-blooded vertebrates; and, moreover (ii) decreases with increasing GC in non-coding (intronic and intergenic) sequences. In other words, CpG levels in both coding and non-coding sequences from warm-blooded vertebrates are correlated with the GC levels of the corresponding isochores. These findings not only contradicted the explanation proposed by Subak-Sharpe and co-workers for the CpG shortage, but also demonstrated that CpG shortage is not uniform throughout the genome, at least in the case of warm-blooded vertebrates, in contrast with the conclusions of Russell et al. (1976). Finally, as far as the different CpG levels exhibited by the genomes of small and large vertebrate viruses are concerned, it was shown that these differences were not related to genome size, but simply to the GC levels of the corresponding genome. Indeed, a plot of CpG vs GC for viral genomes exhibited a positive linear relationship (Bernardi and Bernardi, 1986) which was very similar to that previously found for vertebrate genes (Bernardi et al., 1985). Unfortunately, the explanation of Bernardi and Bernardi (1986) escaped the attention of Karlin et al. (1994), who proposed that CpG suppression in the genomes of virtually all small eukaryotic viruses, but not those of large eukaryotic viruses, "might relate to methylation effects and modes of viral integration and excision. Other contributing factors relate to dinucleotide stacking energies, special mutation mechanisms, and evolutionary events".

An alternative explanation for the CpG shortage in vertebrates was based on the consideration that the C residues in CpG doublets are highly methylated in vertebrates (Sinsheimer, 1955; Doskocil and Sorm, 1962; Grippo et al., 1968; Van der Ploeg and Flavell, 1980; Gruenbaum et al., 1981; Kunnath and Locker, 1982), to the extent of 50–90% in different mammalian DNAs (Gruenbaum et al., 1981). It was proposed that mCpG represents a hot-spot for mutation (Salser, 1977; Coulondre et al., 1978) since it can be deaminated to TpG. Even in the presence of a repair system that specifically corrects mispaired T/G back to C/G (Kramer et al., 1984), the evolutionary effect will be a progressive loss of CpG dinucleotides, leading to a CpG shortage, and to a concomitant increase in TpG (and of its complementary doublet CpA on the other strand). Indeed, a scenario which was put forward for the evolution of DNA methylation in vertebrates (Cooper and Krawczak, 1989) is that the vertebrate genome originally was strongly methylated (and GC-rich) and that the methylated CpG doublets (as well as the GC levels) subsequently underwent a monotonous decay. Along a similar line, unmethylated CpG islands had been previously visualized as remnants of the less methylated genomes of invertebrates ancestral to vertebrates (Bird, 1987).

Investigations aimed at understanding the biological significance of DNA methylation in vertebrates have been carried out along two lines. Studies on the correlation between promoter methylation and gene expression have led to a number of interesting insights. Siegfried and Cedar (1997) have recognized three different strategies by which methylation could affect gene expression: (i) methylation of CpG may interfere directly with the binding of specific transcription factors to DNA; this does not appear to affect gene-specific transcription factors, but rather ubiquitous factors; (ii) the direct binding of specific factors to methylated DNA may lead to gene repression; (iii) methylation may cause repression by altering chromatin structure; indeed, DNA methylation affects positioning of nucleosomes and influences the sensitivity of DNA to DNAses.

In contrast, analyses of DNA methylation levels in different vertebrates and different tissues (Vanyushin et al., 1970, 1973; Pollock et al., 1978; Ehrlich et al., 1982; Gama-Sosa et al., 1983a,b; Serrano et al., 1993) have not led to any general conclusions. Our knowledge of the organization and evolution of the vertebrate genome has, however, made considerable progress in recent years (for a recent review, see Bernardi, 1995). We decided, therefore, to re-examine the problem of DNA methylation in vertebrates. Here, we report investigations on 5mC levels in 42 vertebrate genomes analyzed by us and in 45 additional vertebrate genomes studied by others. We show the existence of two different sets of values, a higher one in fishes and amphibians and a lower one in mammals, birds and reptiles, and we discuss the time in evolution of the transition separating these two levels, its possible cause(s) and its implications.

2. Materials and methods

2.1. DNA sources

DNAs were prepared as previously described (Cuny et al., 1981).

2.2. Analytical procedure for quantitative analysis of nucleosides in DNA

Deoxyribonucleosides were separated by reversed-phase high-performance liquid chromatography (RP-HPLC). Enzymatic hydrolysis of DNA samples (5–10 mg) was carried out using a procedure adapted from methods previously described for tRNAs (Desgrès et al., 1989).

The enzymatic hydrolysates containing major and modified deoxyribonucleosides were analyzed by HPLC using a liquid chromatograph (Kontron Instruments) equipped with a Supelcosil LC-18S column, 250 × 4.6 mm (Supelco), a temperature controlled column oven (26°C ± 0.2°C), a forward scanning diode array

detector (Kontron, Model 440), and a data system DS450-MT2/DAD software. Liquid chromatography was carried out using experimental conditions adapted from procedures previously developed (Gehrke et al., 1984). The deoxyribonucleosides were quantified using an internal standard method in which two wavelengths were used to obtain an optimal sensitivity and accuracy: 255 nm for deoxythymidine (dT), deoxyadenosine (dA) and deoxyguanosine (dG), and 280 nm for deoxycytidine (dC) and 5-methyl-deoxycytidine (m5dC). The average molecular response factors (RMR) using Br8G as internal standard were established from the RP-HPLC of several equimolar solutions of reference nucleosides. The standard deviation was less than 1% for the major deoxyribonucleosides.

3. Results

3.1. Two different methylation levels in vertebrate genomes

Table 1 presents the GC and 5mC levels of the 42 vertebrate genomes studied in the present work, as well as those reported in the literature for 45 additional species. *c*-Values (haploid genome sizes) and the tissues from which DNAs were extracted are also indicated. Even in the sample of Table 1, the largest studied so far, vertebrate classes are not represented in a balanced way, fishes and mammals being, by far, predominant.

A summary of the results of Table 1 indicates (Table 2) that the average methylation values in fishes (1.70%) and amphibians (1.98%) are roughly twice as large as those found in mammals (0.88%), birds (1.02%) and reptiles (1.00%). Since average GC levels are close to 42.5% for all five vertebrate classes, this means that approx. 9% of all cytosines are methylated in fishes and amphibians compared with approx. 4.5% in mammals, birds and reptiles.

This conclusion about two different DNA methylation levels in vertebrates deserves several comments.

(1) It is well known that differences in methylation levels exist among different tissues of a given organism (see Table 1 for examples). This raises a concern about the possible biases introduced by comparing methylation levels of DNAs from different tissues. However, the same 2-fold difference is observed when data from a single tissue (such as sperm or liver) from fishes and mammals are compared.

(2) 5mC levels were originally determined using paper or thin layer chromatographic analysis of bases as obtained by chemical hydrolysis of DNA (Vanyushin et al., 1970, 1973), whereas recent studies, including the present one, make use of the more precise high-performance liquid chromatography (HPLC) of nucleosides obtained by enzymatic hydrolysis. This raises another concern about the validity of comparisons of data obtained with different methods. However, differences related to the methods used are of the order of 10% (Gama-Sosa et al., 1983a; Serrano et al., 1993), a value which is much smaller than the differences under consideration here.

(3) Although previous work on a small sample of fishes and birds (Vanyushin et al., 1970) also showed a difference between these two classes, the difference in methylation levels between the genomes of fishes/amphibians and those of reptiles/mammals/birds reported here rests on a much larger sample also covering other vertebrate classes.

(4) At least part of the variation of 5mC levels within each vertebrate class, order or family is related to the presence of highly methylated satellite DNAs. A typical case is that of the bovine genome, where GC-rich satellites represent 25% of the genome (Filipski et al., 1973; Thiery et al., 1976; Cortadas et al., 1977; Macaya et al., 1978; Kopecka et al., 1978) and are highly methylated in somatic tissues, leading to a 5mC level of 1.30–1.40%, almost twice as high as that, 0.75%, of germ cells, where these satellites are undermethylated (see Table 1). This remark mainly applies to mammals, where abundant GC-rich satellites are frequent (see Sabeur et al., 1993).

(5) Results reported in the following paper show that compositional DNA fractions from *Xenopus* have approx. twice as high a methylation level compared with the compositionally similar DNA fractions from three warm-blooded vertebrates (chicken, mouse and man), ruling out the possibility that the difference is essentially due to satellite or interspersed repetitive DNA.

Fig. 1 shows that two statistically significant correlations ($P<0.0001$) hold between 5mC and genomic GC levels, one for fishes and amphibians and another one for mammals and birds. It should be noted that (i) a few mammalian values deviate from the correlation shown by mammals/birds; deviating points concern, beside the bovine somatic cells mentioned above, *Erinaceus europeus, Meriones unguiculatus, Cricetus norvegicus, Manis* sp. (lower values) and *Pteropus poliocephalus* (higher value); these deviations might be due to under- or over-methylated satellites; (ii) reptilian values show a larger scatter compared with values from mammals/birds (see Table 2) with most of them being intermediate between cold- and warm-blooded vertebrates; yet closer to the latter than to the former; (iii) one of the fishes, the only chondrychthyan studied, showed an especially low value compared with other fishes, 1.15%.

The correlation between 5mC and genomic GC levels is understandable if one recalls that CpG frequency is linearly correlated with the GC content of the genome (Bernardi, 1985; Aïssani and Bernardi, 1991a,b) and

Table 1
5mC, GC and *c*-values in vertebrate DNAs[a]

Order	Family	Species	Tissue	*c*-value	5mC (mol%)	GC (mol%)	5mC ref.
Fishes							
Cyclostomata							
Petromyzontiformes	Petromyzontidae	*Lampetra fluviatilis*	blood		2.09	50.9	1
Chondrichthyes							
		Petrolamiops longimanus	sperm		1.16	45.4	1
Osteichthyes							
Acipenseriformes	Acipenseridae	*Acipenser guldenstädtii*	sperm		1.79	43.8	1
		Huso huso	sperm	1.8	1.81	43.4	1
		Acipenser ruthenus	sperm		1.87	43.0	1
		Acipenser stellatus	sperm		1.87	42.9	1
Antheriniformes	Orizatidae	*Oryzias latipes*	liver	1.1	1.85	40.8	5
Clupeiformes	Clupeidae	*Clupea harengus*	sperm	0.77	1.87	45.2	1
Cypriniformes	Cyprinidae	*Cyprinus carpio*	sperm	1.7	1.56	38.9	1
			blood		1.38	38.9	1
		Brachydanio rerio	liver	1.8	1.27	37.3	1
Siluriformes	Siluridae	*Silurus glanis*	blood		1.64	41.1	1
	Ictaluridae	*Ictalurus nebulosus*	liver	1.2	1.92	41.3	5
Salmoniformes	Esocidae	*Esox lucius*	sperm	0.85	1.87	43.5	1
			liver		1.98	44.2	1
			blood		1.47	43.5	1
	Salmonidae	*Osmerus eperlanus*	sperm		2.01	45.4	1
		Oncorhynchus sp.	sperm	3	1.64	44.0	1
			liver		2.13	44.1	1
		Oncorhynchus tschawitscha	sperm	3.3	1.73	41.9	1
		Oncorhynchus mykiss	liver	2.8	1.45	42.9	5
		Salmo gairdneri	liver	3.2	1.45	42.1	5
Gadiformes	Gadidae	*Gadus morhua*	blood	0.4	2.24	50.0	1
		Eleginus navaga	blood		2.15	48.2	1
Scorpaeniformes	Cottidae	*Cottus* sp.	sperm	1	1.70	41.0	1
Perciformes	Scombridae	*Trichiurus japonicus*	sperm		1.50	41.3	1
	Thynnidae	*Thunnus thynnus*	sperm	1	1.54	37.0	1
	Cichlidae	*Oreochromis grahami*	liver		1.29	41.5	6
		Oreochromis niloticus	liver	0.95	1.26	41.0	6
		Tilapia buttikofferi	liver		1.46	40.3	6
	Mugilidae	*Mugil cephalus*	sperm	0.99	1.87	39.5	2
			liver		1.80	39.5	2
Pleuronectiformes	Pleuronectidae	*Pleuronectes flesus*	blood		1.70	44.9	1
		Liopsetta glacialis	blood	0.73	1.84	45.2	1
Tetraodontiformes	Tetraodontidae	*Sphaeroides* sp.	sperm	0.5	1.50	41.3	1
Amphibia							
Anura	Leptodactylidae	*Odontophrynus americanus*	blood		1.63	41.9	6
	Pipidae	*Xenopus laevis*	liver	3.1	1.40	42.2	1
	Ranidae	*Rana temporaria*	liver	4.2	2.60	45.1	1
			blood		2.61	45.1	1
		Rana pipiens	liver	5.8	2.30	47.3	1
Reptilia							
Crocodylia	Crocodylidae	*Crocodilus niloticus*	blood		1.06	44.4	6
Chelonia	Testudinidae	*Testudo horsfieldi*	blood		1.30	45.9	1
			liver		1.50	44.8	1
		Testudo graeca	blood	5.4	0.77	45.7	6
Squamata	Varanidae	*Varanus griseus*	blood	2.4	1.15	43.7	1
Ophidia	Crotalidae	*Bothrops neuwiedi*	blood		1.01	40.0	6
		Bothrops jararaca	blood		0.97	39.7	6
	Boinae	*Boa constrictor amarali*	blood	3.15	0.61	40.4	6
Aves							
Rheiformes	Rheidae	*Rhea americana*	blood		0.98	44.0	6
Sphenisciformes	Spheniscidae	*Spheniscus demersus*	blood	1.6	1.01	43.3	6
Anseriformes	Anatidae	*Cairina moschata*	blood	1.35	0.64	41.5	6
Galliformes	Phasianidae	*Gallus gallus*	blood	1.25	1.04	44.2	1
Columbiformes	Columbidae	*Columba livia*	blood	1.3	1.02	44.0	6
Passeriformes	Corvidae	*Corvus corone*	blood	1.7	1.15	45.4	1
Ralliformes	Rallidae	*Fulica atra*	blood	1.7	1.33	49.3	1

Table 1 (*continued*)

Order	Family	Species	Tissue	*c*-value	5mC (mol%)	GC (mol%)	5mC ref.
Mammalia							
Monotremata							
Monotremata	Ornithorhynchidae	*Ornitorhyncus anatinus*	blood	3.78	1.32	45.6	6
	Tachyglossidae	*Tachyglossus aculeatus*	blood	3.57	0.99	45.9	6
Euteria							
Insectivora	Erinaceidae	*Erinaceus europeus*	thymus		0.53	45.5	6
	Soricidae	*Crocidura russula*	liver		0.66	41.4	6
Dermoptera	Cynocephalidae	*Cynocephalus variegatus*	liver		0.89	40.6	6
Megachiroptera	Pteropodidae	*Pteropus* sp.	liver		0.96	40.5	6
	Pteropodidae	*Pteropus poliocephalus*	liver		1.39	41.5	6
Microchiroptera	Hipposideridae	*Hipposideros galeritus*	whole animal		0.86	41.4	6
	Rhinolophidae	*Rhinolophus creaghi*	whole animal		0.81	41.5	6
	Vespertilionidae	*Myotis lucifugus*	whole animal	3.2	0.97	43.5	6
	Phyllostomidae	*Chiroderma salvinii*	liver		0.86	41.4	6
	Nycteriidae	*Nycteris thebaica*	whole animal		0.88	42.9	6
	Molossidae	*Chaerephon pumila*	liver	2.9	0.86	41.4	6
Primates	Hominidae	*Homo sapiens*	liver	3.5	0.88	42.6	3
			sperm		0.84	42.6	3
			blood		0.96	42.6	3
	Cercopithecidae	*Macaca mulatta*	liver	3.15	0.86	41.3	4
		Macaca fascicularis	liver		0.96	41.3	4
		Cercopithecus aethiops	liver		0.98	41.3	4
	Cebidae	*Saimiri sciureus*	liver		0.85	41.3	4
	Tupaiidae	*Tupaia montana*	liver		1.08	41.9	6
	Lemuridae	*Hapalemur griseus*	liver		0.93	41.4	6
Pholidota	Manidae	*Manis* sp.	liver		0.59	42.4	6
Lagomorpha	Leporidae	*Oryctolagus cuniculus*	liver	3.2	0.86	44.3	6
Rodentia	Sciuridae	*Sciurus vulgaris*	liver	5.1	0.61	39.5	6
	Cricetidae	*Cricetus norvegicus*	liver	4	0.37	40.7	6
	Spalacidae	*Spalax* sp.	liver		0.58	38.4	6
	Caviidae	*Cavia porcellus*	liver	2.95	0.74	39.7	6
	Muridae	*Rattus norvegicus*	liver	3.8	0.94	43.9	6
		Mus musculus	liver	3.25	0.95	42.2	4
			sperm		0.83	42.2	4
		Meriones unguiculatus	nd		0.70	45.2	2
	Procaviidae	*Procavia capensis*	liver	3.64	0.70	41.0	6
Cetacea	Balenopteridae	*Balenoptera physalus*	liver		0.94	41.3	6
	Physiteridae	*Physeter macrocephalus*	liver		1.06	41.9	6
	Phocoenidae	*Phocena phocena*	liver		0.97	41.4	6
Carnivora	Canidae	*Canis familiaris*	liver	3.3	0.67	41.1	6
	Felidae	*Panthera uncia*	liver		0.94	41.5	6
Perissodactyla	Equidae	*Equus caballus*	liver	2.95	1.04	42.8	6
Artiodactyla	Bovidae	*Bos taurus*	liver	3.35	1.40	44.0	4
			kidney		1.30	44.0	4
			sperm		0.75	44.2	4
	Suidae	*Sus scrofa*	kidney	2.77	1.20	44.0	4
			sperm		0.77	42.1	4
	Ovidae	*Ovis aries*	liver	2.9	1.13	42.6	4
			sperm		0.76	42.0	4

[a] 5mC values are from (1) Vanyushin et al. (1970, 1973); (2) Pollock et al. (1978), Gama-Sosa et al. (1983b); (3) Ehrlich et al. (1982); (4) Gama-Sosa et al. (1983a); (5) Serrano et al. (1993) and (6) present work. GC values are from Bernardi and Bernardi (1990) for fishes, amphibians and reptiles, Kadi et al. (1993) for birds and Sabeur et al. (1993) for mammals. *c*-Values are from literature quoted in the above papers and from Bachmann (1972) and Venturini et al. (1986).

that CpG is the major site of methylation, methylation on other dinucleotides (Woodcock et al., 1987) and on CpNpG trinucleotides (Clark et al., 1995) being very limited. A similar correlation was also found in plant genomes (Matassi et al., 1992; Montero et al., 1992) and in isochore families from individual vertebrate genomes (human, mouse, chicken and Xenopus; see Cacciò et al., 1997).

3.2. The possible reasons for two methylation levels

The reasons for the two different methylation levels in vertebrate genomes might be related to the different

Table 2
5mC and GC levels in different classes of vertebrates

Class	Species	GC %	SD[a]	5mC	SD[a]
Fishes[b]	28	42.44	(2.87)	1.70	(0.26)
Amphibians	4	44.13	(2.56)	1.98	(0.56)
Reptiles	7	42.75	(2.63)	1.00	(0.26)
Birds	7	44.53	(2.41)	1.02	(0.21)
Mammals	39	42.07	(1.68)	0.88	(0.20)
Vertebrates	85	42.55			

[a]SD, standard deviation.
[b] Only Osteichthyes were taken into consideration.

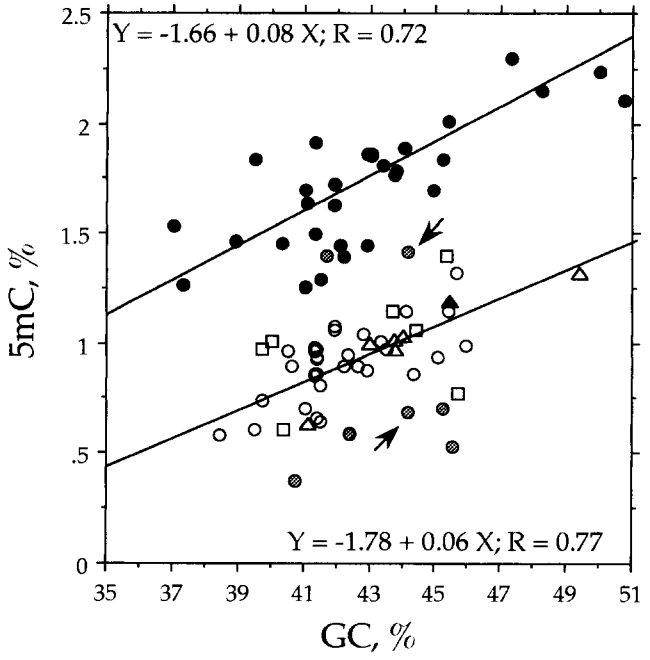

Fig. 1. Plots of 5mC levels against GC levels for the genomes of fishes/amphibians (solid circles), reptiles (open squares), mammals (open circles) and birds (open triangles). When data from different tissues were available, mean values were used. Deviating points of mammals (see text) are in grey (two of them, indicated by arrows, corresponding to DNAs from bovine somatic and germ cells, respectively) and were not taken into account in drawing the regression lines of mammals and birds; the single deviating point from fishes is shown as a solid triangle. Correlation coefficients, slopes and intercepts are given for fishes/amphibians and mammals/birds, respectively.

relative amounts of repetitive DNA sequences since (i) repetitive DNA sequences, as isolated by reassociation experiments, are more methylated than single-copy DNA sequences, at least in mammals (Romanov and Vanyushin, 1981; Gama-Sosa et al., 1983b); and (ii) the proportion of repetitive DNA sequences is larger in fishes/amphibians (59–77%) than in reptiles/mammals/birds (27–41%) (see Olmo et al., 1989, for a review). As expected from these premises, when the relative amount of repetitive DNA sequences (from Ginatulin, 1984) was plotted against the methylation level of the corresponding genomes, a positive, statistically significant correlation was found (see Fig. 2).

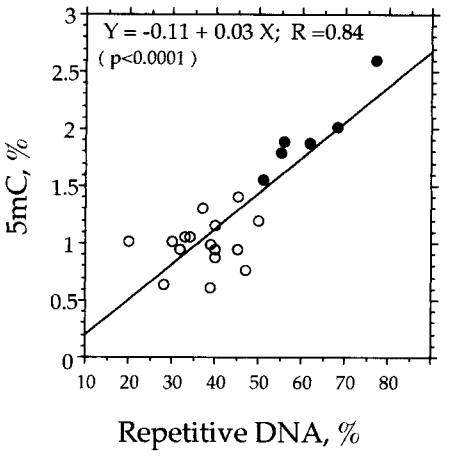

Fig. 2. Plot of 5mC against the percentage of repetitive DNA sequences of the corresponding genomes (from Ginatulin, 1984). Solid circles correspond to fishes/amphibians, open circles to mammals/birds. The correlation coefficient, slope, intercept and *p*-value are shown.

There are, however, some problems with Fig. 2. (i) While the correlation persists when plotting only the results from five fishes and one amphibian, no significant correlation is found when plotting only data from birds and mammals. (ii) At least one of the points of Fig. 2 is very doubtful. Indeed, *Clupea harengus*, with a *c*-value of only 0.77 pg (see Table 1), was estimated to comprise 62% repetitive sequences, leaving only 0.29 pg for non-repetitive DNA. Now, the work of Pizon et al. (1984) has shown that in *Arothron diadematus*, one of the Tetraodontid fishes which have the smallest genome size of all vertebrates, 0.45 pg, repetitive sequences amount to 13%, leaving 0.39 pg for non repetitive DNA. This result casts a serious doubt on the *Clupea harengus* results, and decreases the points supporting the correlation to only five. More importantly, the results of Pizon et al. (1984) led to the conclusion that the reduction in genome size, such as occurred in *Tetraodontiformes*, took place at the expense of repetitive sequences and, as shown by recent work (Brenner et al., 1993), also of introns. Conversely, this conclusion confirms the suggestion (Olmo et al., 1989) that genome size increases largely take place by an increase of repetitive sequences. (iii) The genomes of fishes characterized by the lowest *c*-values and by the lowest amount of repetitive sequences (*Tetraodontiformes* and *Gadidae*) show a methylation level that is as large as, or even larger than, those of fishes with a genome size at least twice as large. Along the same lines, there is no substantial difference in methylation level between the two classes of warm-blooded vertebrates, in spite of the fact that birds have a genome size close to one-half that of mammals and a much larger amount of slow-reassociating DNA (84% in chicken DNA vs. 62% in human DNA; Olofsson and Bernardi, 1983; Soriano et al., 1981). It should, therefore, be concluded that the difference in methylation

level between fishes/amphibians and reptiles/mammals/birds is not correlated with the amounts of repetitive sequences within either one of the two sets of genomes.

This conclusion is confirmed by investigations concerning the correlation between methylation and *c*-values. As discussed above, *c*-values are related to the amounts of repetitive sequences. In fact, they have the advantage of providing a large data set compared with reassociation data. As shown in Fig. 3a, a negative correlation holds between 5mC levels and *c*-values. This negative correlation can be understood on the basis of the following considerations. A negative correlation was previously found between GC and genome size in the case of fish genomes (Bernardi and Bernardi, 1990; see Fig. 3b). This indicates that genome expansion (in a *c*-value range where polyploidization is absent) takes place by an increase in the size of intergenic regions, which, as a general rule (Clay et al., 1996), are GC-poorer than coding sequences and, as a consequence, have fewer methylatable sites. Therefore, increasing the genome size means increasing GC-poorer regions of the genome which leads to a relative decrease of methylation, so accounting for the results of Fig. 3a.

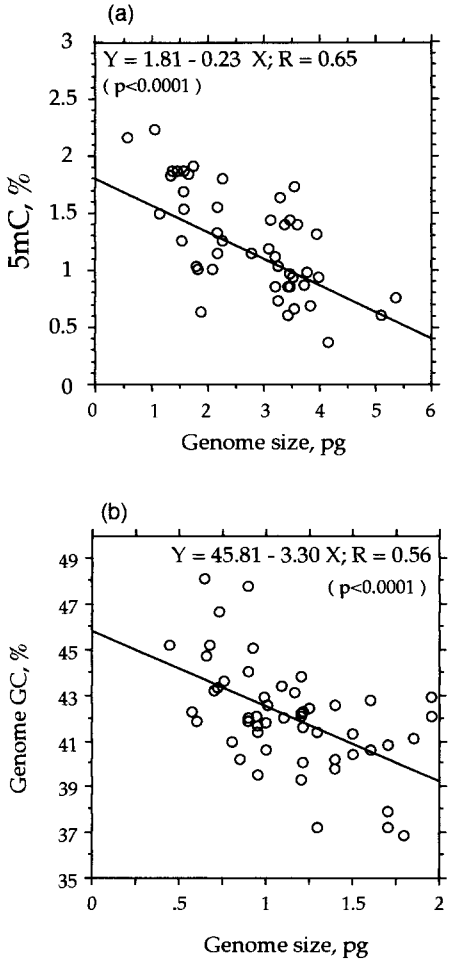

Fig. 3. (a) Plot of 5mC levels against the corresponding haploid genome sizes of vertebrates. Data from Table 1 were used to construct the plot. (b) GC levels of fish DNAs are plotted against the corresponding haploid genome sizes. Data are from Table 3 of Bernardi and Bernardi (1990). *c*-Values were used only up to 2 pg, higher values probably being due to polyploidization. Other indications as in Fig. 2.

3.3. CpG shortage in vertebrate genomes

Table 3 displays the CpG frequency and the observed/expected ratios for the CpG dinucleotides, as well as the genomic GC levels. The data show that the genomes of fishes and amphibians ranging from Chondrychthyes to Anura (*Squalus acanthius*, *Salmo salar*, *Latimeria* sp., and *Rana catesbeiana*) have a similar CpG o/e ratio, 0.37, which is significantly higher than that, 0.26, found for nine genomes, from eight mammals and one bird. Unfortunately, no data are available for reptiles. Fig. 4 shows that CpG levels and methylation levels are correlated with each other in fishes and amphibians, on the one hand, and in mammals and birds, on the other.

4. Discussion

The results obtained in the present work, summarized in Fig. 4, indicate two different levels of methylation and CpG shortage, one in fishes and amphibians, and the other one in mammals and birds. As already pointed out, the possible biases introduced by tissue-specific variations and by the use of different techniques to measure 5mC do not affect the validity of the conclusion

Table 3
GC and CpG levels of some vertebrate genomes

	GC	CpG	CpG o/e	Reference
Squalus acanthius	45.3	1.8	0.35	4
Salmon salar	43.1	1.7	0.37	1
Latimeria sp.	41.7	1.8	0.41	4
Rana catesbeiana	45	1.7	0.33	5
average	43.78	1.75	0.37	
Oryctolagus cuniculus	44.3	1.3	0.25	1
Rattus norvegicus	43.9	1.2	0.25	2
Mus musculus	42.2	1	0.23	1
Bos taurus	44	1.4	0.29	1
Phocoena phocoena	41.4	1.3	0.33	3
Homo sapiens	42	1	0.25	1
Halichoerus sp.	39	0.9	0.24	3
Cavia porcellus	40	1	0.26	3
Gallus gallus	44.2	1.1	0.23	1
average	42.33	1.13	0.26	*

References: (1) Swartz et al. (1962); (2) Skalka et al. (1966); (3) Russell (1974); (4) Russell and Subak-Sharpe (1977); (5) McGeoch (1970).

**Bos taurus* was excluded from the average because of large amounts (25%) of GC-rich satellites (Macaya et al., 1978).

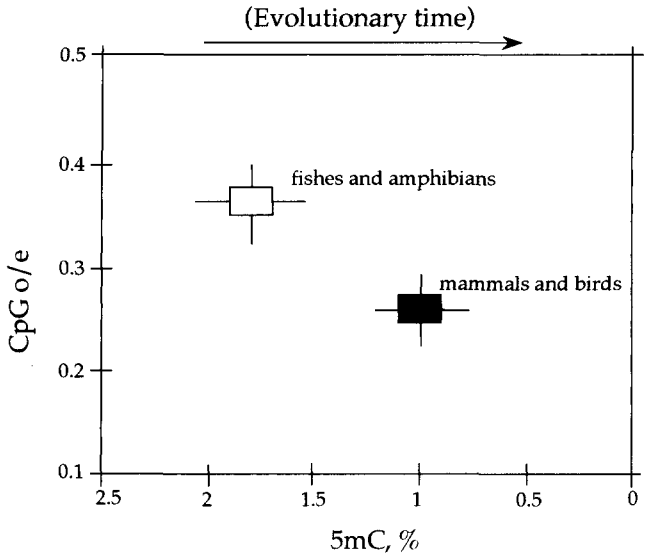

Fig. 4. Plot of CpG observed/expected ratio (CpG o/e) against 5mC level in the genomes of cold-blooded and warm-blooded vertebrates. In the case of CpG o/e values, data were taken from the available literature for four cold-blooded vertebrates (three fishes and one amphibian) and nine warm-blooded vertebrates (eight mammals and one bird). In the case of methylation, all available data were used, namely 32 cold-blooded vertebrate and 46 warm-blooded vertebrate species, respectively. Horizontal and vertical lines crossing the boxes correspond to standard deviations.

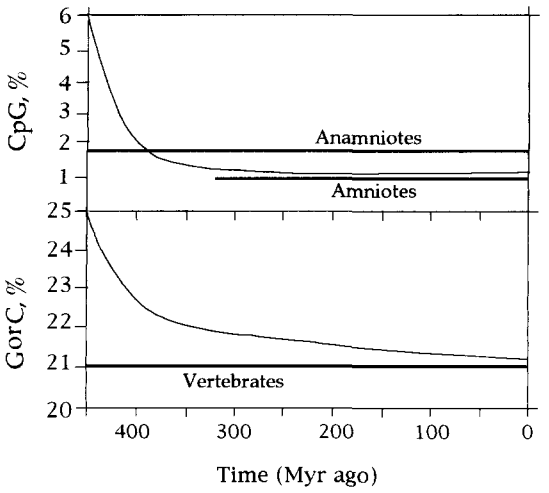

Fig. 5. The results of the present work on DNA methylation of fishes/amphibians (upper horizontal line) and of reptiles/mammals/birds (lower horizontal line) and the average GC level (horizontal line) are compared with the monotonous decrease of CpG and GC hypothesized by Cooper and Krawczak (1996).

about the different methylation levels in the two groups of vertebrates.

Several important questions then arise concerning (i) the equilibria in methylation and CpG shortage; (ii) the time in evolution at which the transition between the two equilibria occurred; and (iii) the causes of the transition.

4.1. Two equilibria in DNA methylation and CpG shortage of vertebrates

The first conclusion of this work is that two equilibria of DNA methylation and CpG shortage exist in vertebrate genomes. This is stressed by the almost complete lack of overlap of the two sets of 5mC and CpG o/e data for fishes/amphibians and mammals/birds, respectively (Tables 1 and 3; Figs. 1 and 4) and by the very similar methylation levels of mammals belonging to orders separated from each other by more than 65 million years.

The existence of two different methylation levels separated by a transition has an important implication, namely that it contradicts the hypothesis (Cooper and Krawczak, 1989) of a monotonous decay of 5mC and of CpG levels in vertebrate genomes, as well as the hypothesis of a very high methylation and CpG levels in the ancestral vertebrates (see Fig. 5).

4.2. The time of the methylation/CpG transition in vertebrate genomes

A rough estimate of the time at which the methylation/CpG levels changed in the evolution of vertebrates may be obtained from the data on reptilian genomes. Indeed, the results available so far suggest that reptiles are characterized by a lower methylation level close to those of mammals and birds (with the possible single exception of *T. horsfieldi*). The present results place the methylation transition at the appearance of the common ancestor of reptiles and mammals, namely at the appearance of amniotes.

One should, however, keep in mind that the number of reptilian DNAs analyzed so far is small, that rather different levels were found in the few reptiles investigated and that no estimate is available for the CpG shortage in reptilian DNAs. Under these circumstances, one should consider that the question as to whether the low methylation level is present in all or only in some reptiles is still open. Therefore, two possibilities should still be considered, namely, that either the common reptilian ancestor of mammals and birds had already undergone a decrease in methylation which was then transmitted by descent to extant reptiles, mammals and birds; or that the methylation change occurred independently in reptiles (possibly, only in some of them), in mammals and in birds. Needless to say, the first explanation is the more parsimonious one and definitely the more acceptable one at present. The genomes of mammals and birds differ from those of fishes and amphibians in several major respects (see Bernardi, 1995, for a review), namely in that they are characterized not only (i) by lower methylation and lower CpG levels, as reported in the

present work; but also (ii) by the presence of very GC-rich isochores, of CpG islands (Aïssani and Bernardi, 1991a,b) and of T(elomeric) chromosomal bands; and (iii) by an increase of karyotypic changes and species formation (Bush et al., 1977; Bernardi, 1993).

The present results indicate that these changes were not concomitant. Indeed, low methylation levels appear to be already present in reptiles. In contrast, reptilian genomes lack CpG islands as well as GC-rich isochores (Aïssani and Bernardi, 1991a,b). Moreover, T bands are absent in reptilian chromosomes. Finally, karyotypic change and species formation are slow in this vertebrate class (Bush et al., 1977; see also Bernardi, 1993).

4.3. The causes of the transition in methylation and CpG levels: a speculation

At this point in time, it is difficult to draw any conclusion about the causes of the transition in methylation and CpG levels which seems to have occurred at the time of the appearance of reptiles. One could, however, speculate that the lower 5mC level of warm-blooded vertebrates is due to a higher deamination rate related to their higher body temperature, the deamination of 5mC residues in double-stranded DNA having a strong temperature dependence (Shen et al., 1994). The fact that reptiles are similar to warm-blooded vertebrates in their genome methylation may be accounted for by the high body temperature which is reached in these species. It should be mentioned, however, that the deamination rate is much higher than the mutation rate at CpG doublets and that the latter is largely dependent upon the efficiency of the mismatch repair mechanism (Shen et al., 1994; Yang et al., 1996) which might, however, be overburdened by the 5-fold increased production of mismatches due to the increase in temperature from approx. 20 to 37°C. Under this hypothesis, a methylation decrease would not require the permanent higher body temperature of warm-blooded vertebrates to become effective, whereas other properties such as the formation of GC-rich isochores, CpG islands, and T bands might require it, as suggested by Bernardi and Bernardi (1986) and Bernardi (1993).

Acknowledgement

We thank Giuseppe Geraci, Ettore Olmo and Aharon Razin for their comments. We also thank Dr. J. Pettigrew and Dr. I. Ruiz for the generous gift of some of the DNA samples used in this work.

References

Aïssani, B., Bernardi, G., 1991a. CpG islands: features and distribution in the genome of vertebrates. Gene 106, 173–183.

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

Bachmann, K., 1972. Genome size in mammals. Chromosoma 37, 85–93.

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

Bernardi, G., 1993. Genome organization and species formation in vertebrates. J. Mol. Evol. 37, 331–337.

Bernardi, G., 1995. The human genome organization and evolutionary history. Annu. Rev. Genet. 29, 445–476.

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

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

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.

Bird, A., 1987. CpG islands as gene markers in the vertebrate nucleus. Trends Genet. 12, 342–347.

Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B., Aparicio, S., 1993. Characterization of the pufferfish (*Fugu*) genome as a compact model vertebrate genome. Nature 366, 265–268.

Bush, G.L., Case, S.M., Wilson, A.C., Patton, J.L., 1977. Rapid speciation and chromosomal evolution in mammals. Proc. Natl. Acad. Sci. USA 74, 3942–3946.

Cacciò, S., Jabbari, K., Matassi, G., Gamonprez, F., Desgrès, J., Bernardi, G., 1997. Methylation patterns in the isochores of vertebrate genomes. Gene, in press.

Clark, S.J., Harrison, J., Frommer, M., 1995. CpNpG methylation in mammalian cells. Nature Genet. 10, 20–27.

Clay, O., Cacció, S., Zoubak, S., Mouchiroud, D., Bernardi, G., 1996. Human coding and noncoding DNA: compositional correlations. Mol. Phyl. Evol. 5, 2–12.

Cooper, D.N., Krawczak, M., 1989. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum. Genet. 83, 181–188.

Cooper, D.N., Krawczak, M., 1996. Mutational processes in pathology and evolution. In: Jackson, M., Strachan, T., Dover, G. (Eds.), Human Genome Evolution, pp. 1–33.

Cortadas, J., Macaya, G., Bernardi, G., 1977. An analysis of the bovine genome by density gradient centrifugation fractionation in Cs_2SO_4/3,6 bis (acetato-mercurimethyl) dioxane density gradient. Eur. J. Biochem. 76, 13–19.

Coulondre, C., Miller, J.H., Farabaugh, P.J., Gilbert, W., 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. Nature 274, 775–780.

Cuny, G., Soriano, P., Macaya, G., Bernardi, G., 1981. The major components of the mouse and human genomes: preparation, basic properties and compositional heterogeneity. Eur. J. Biochem. 111, 227–233.

Desgrès, J., Keith, G., Kuo, K.C., Gehrke, C.W., 1989. Presence of phosphorylated O-ribosyl-adenosine in T-y-stem of yeast methionine initiator tRNA. Nucleic Acids Res. 17, 865–882.

Doskocil, J., Sorm, F., 1962. Distribution of 5-methylcytosine in pyrimidine sequences of deoxyribonucleic acids . Biochim. Biophys. Acta 55, 953

Ehrlich, M., Gama-Sosa, M.A., Huang, L.H., Midgett, R.M., Kuo, K.C., McCune, R.A., Gehrke, C., 1982. Amount and distribution of 5-methylcytosine in human DNA from different types of tissues of cells. Nucleic Acids Res. 10, 2709–2721.

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

Forget, B.G., Cavalesco, C., De Riel, J.K., Tuan, D., Biro, A.P., Wilson, J.T., Wilson, L.B., Weismann, S.M., 1979. Structure of the human globin genes. J. Supramol. Struct. Suppl. 3, 44

Gama-Sosa, M.A., Midgett, R., Slagel, V.A., Githens, S., Kuo, K.C., Gehrke, C.W., Ehrlich, M., 1983a. Tissue-specific differences in DNA methylation in various mammals. Biochim. Biophys. Acta 740, 212–219.

Gama-Sosa, M.A., Wang, R., Kuo, K.C., Gehrke, C.W., Ehrlich, M., 1983b. The 5-methylcytosine content of highly repeated sequences in human DNA. Nucleic Acids Res. 11, 3087–3095.

Gehrke, C.W., MacCune, R.A., Gama-Sosa, M., Ehrlich, M., Kuo, K.C., 1984. Quantitative reverse phase high performance liquid chromatography of major and modified nucleosides in DNA. J. Chromatogr. 301, 199–219.

Ginatulin, A.A., 1984. Structure, Organization, Evolution of Vertebrate Genome. Nauka, Moscow.

Grippo, P., Iaccarino, M., Parisi, E., Scarano, E., 1968. Methylation of DNA in developing sea urchin embryos. J. Mol. Biol. 36, 195–208.

Gruenbaum, Y., Stein, R., Cedar, H., Razin, A., 1981. Methylation of CpG sequences in eukaryotic DNA. FEBS Lett. 124, 67–71.

Josse, J., Kaiser, A.D., Kornberg, A., 1961. Enzymatic synthesis of deoxyribonucleic acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. J. Biol. Chem. 236, 864–875.

Kadi, F., Mouchiroud, D., Sabeur, G., Bernardi, G., 1993. The compositional patterns of the aviain genomes and their evolutionary implications. J. Mol. Evol. 37, 544–551.

Karlin, S., Doerfler, W., Cardon, L.R., 1994. Why is CpG suppressed in the genomes of virtually all eukaryotic viruses but not in those of large eukaryotic viruses? J. Virol. 68, 2889–2897.

Kopecka, H., Macaya, G., Cortadas, J., Thiery, J.P., Bernardi, G., 1978. Restriction enzyme analysis of satellite DNA components from the bovine genome. Eur. J. Biochem. 84, 189–195.

Kramer, B., Kramer, W., Fritz, H.J., 1984. Different base/base mismatches are corrected with different efficiencies by the methyl-directed DNA mismatch repair system of *E. coli*. Cell 38, 879–887.

Kunnath, L., Locker, J., 1982. Characterization of DNA methylation in the rat. Biochim. Biophys. Acta 699, 264–271.

Lennon, G.G., Fraser, N.W., 1983. CpG frequency in large DNA segments. J. Mol. Evol. 19, 286–288.

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

Matassi, G., Melis, R., Kuo, K.C., Macaya, G., Gehrke, C.W., Bernardi, G., 1992. Large-scale methylation patterns in the nuclear genomes of plants. Gene 122, 239–245.

McGeoch, D.J., 1970. Some base sequence characteristics of deoxyribonucleic acid. PhD thesis, Institute of Biochemistry, University of Glasgow.

Montero, L.M., Filipski, J., Gil, P., Capel, J., Martinez-Zapater, J.M., Salinas, J., 1992. The distribution of 5-methylcytosine in the nuclear genome of plants. Nucleic Acids Res. 20, 3207–3210.

Olmo, E., Capriglione, T., Odierna, G., 1989. Genome size evolution in vertebrates: trends and constraints. Comp. Biochem. Physiol. 73B, 739–745.

Olofsson, B., Bernardi, G., 1983. Organization of nucleotide sequences in the chicken genome. Eur. J. Biochem. 130, 241–245.

Pizon, V., Cuny, G., Bernardi, G., 1984. Nucleotide sequence organization in the very small genome of a tetraodontid fish *Arothron diadematus*. Eur. J. Biochem. 140, 25–30.

Pollock Jr., J.M., Swihart, M., Taylor, J.H., 1978. Methylation of DNA in early development: 5-methyl cytosine content of DNA in sea urchin sperm and embryos. Nucleic Acids Res. 5, 4855–4861.

Romanov, G.A., Vanyushin, B.F., 1981. Methylation of reiterated sequences in mammalian DNAs. Effects of the tissue type, age, malignancy and hormonal induction. Biochim. Biophys. Acta 653, 204–218.

Russell, G.J., 1974. Characterization of Deoxyribonucleic Acids by Doublet Frequency Analysis. Ph.D. Thesis, University of Glasgow, Glasgow.

Russell, G.J., Subak-Sharpe, J.H., 1977. Similarity of the general designs of protochordates and invertebrates. Nature 266, 533–536.

Russell, G.J., Walker, P.M.B., Elton, R.A., Subak-Sharpe, J.H., 1976. Doublet frequency analysis of fractionated vertebrate nuclear DNA. J. Mol. Biol. 108, 1–23.

Sabeur, G., Macaya, G., Kadi, F., Bernardi, G., 1993. The isochore patterns of mammalian genomes and their phylogenetic implications. J. Mol. Evol. 37, 93–108.

Salser, W., 1977. Globin mRNA sequences: analysis of base pairing and evolutionary implications. Cold Spring Harbor Symp. Quant. Biol. 40, 985–1002.

Serrano, J., Kuehl, D.W., Naumann, S., Analytical procedure and quality assurance criteria for the determination of major and minor deoxynucleosides in fish tissue DNA by liquid chromatography–ultraviolet spectroscopy and liquid chromatography–thermo. 1993. J. Chromatogr. 615, 203–213.

Shen, J.C., Rideout, W.M., Jones, P.A., 1994. The rate of hydrolytic deamination of 5-methylcytosine in double-stranded DNA. Nucleic Acids Res. 22, 972–976.

Siegfried, Z., Cedar, H., 1997. DNA methylation: a molecular lock. Curr. Biol. 7, R305–R307.

Sinsheimer, R.L., 1955. The action of pancreatic deoxyribonuclease II. Isomeric dinucleotides,. J. Biol. Chem. 215, 579

Skalka, A., Audree, V.F., Hurwitz, G., 1966. The effect of histones on the enzymatic synthesis of ribonucleic acid. J. Biol. Chem. 241, 588–596.

Soriano, P., Macaya, G., Bernardi, G., 1981. The major components of the mouse and human genomes: reassociation kinetics. Eur. J. Biochem. 115, 235–239.

Subak-Sharpe, J.H., 1967. Doublet patterns and evolution of viruses. Brish Med. Bull. 23, 161–168.

Subak-Sharpe, H., Burk, R.R., Crowford, L.V., Morrisson, J.M., Hay, J., Keir, A.M., 1966. An approach to evolutionary relationship of mammalian DNA viruses through analysis of the pattern of nearest neighbor base sequence. Cold Spring Harbor Symp. Quant. Biol. 31, 737–738.

Subak-Sharpe, J.H., Elton, R.A., Russell, G.J., 1974. Evolutionary implications of doublet analysis. Symp. Soc. Gen. Microbiol. 24, 131–150.

Swartz, M.N., Trautner, T.A., Kornberg, A., 1962. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. J. Biol. Chem. 237, 1961–1967.

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

Van der Ploeg, L.H.T., Flavell, R.A., 1980. DNA methylation in the human β-globin locus in erythroid and non-erythroid tissues. Cell 19, 947–958.

Vanyushin, B.F., Mazin, A.L., Vasilyev, V.K., Belozersky, A.N., 1973. The content of 5-methylcytosine in animal DNA: the species and tissue specificity. Biochim. Biophys. Acta 299, 397–403.

Vanyushin, B.F., Tkacheva, S.G., Belozersky, A.N., 1970. Rare bases in animal DNA. Nature 225, 948–949.

Venturini, G., Raffaella, D., Ernest, C., 1986. Size and structure of the bird genome-I DNA content of 48 species of neognathae. Comp. Biochem. Physiol. 85B, 61–65.

Woodcock, D.M., Crowther, P.J., Diver, W.P., 1987. The majority of methylated deoxycytosines in human DNA are not in the CpG dinucleotide. Biochim. Biophys. Res. Commun. 145, 888–894.

Yang, A.S., Gonzalgo, M.L., Zingg, J.M., Millar, R.P., Buckley, J.D., Jones, P.A., 1996. The rate of CpG mutation in Alu repetitive elements within the p53 tumor suppressor gene in the primate germline. J. Mol. Biol. 258, 240–250.