CpG Islands in Vertebrate Genomes

M. Gardiner-Garden[1] and M. Frommer[1,2]

[1]*The Kanematsu Laboratories, Royal Prince Alfred Hospital*
Missenden Road, Camperdown
N.S.W. 2050, Australia

[2]*CSIRO Division of Molecular Biology, P.O. Box 184*
North Ryde, N.S.W. 2113, Australia

(Received 3 November 1986, and in revised form 25 February 1987)

Although vertebrate DNA is generally depleted in the dinucleotide CpG, it has recently been shown that some vertebrate genes contain CpG islands, regions of DNA with a high G+C content and a high frequency of CpG dinucleotides relative to the bulk genome. In this study, a large number of sequences of vertebrate genes were screened for the presence of CpG islands. Each CpG island was then analysed in terms of length, nucleotide composition, frequency of CpG dinucleotides, and location relative to the transcription unit of the associated gene. CpG islands were associated with the 5′ ends of all housekeeping genes and many tissue-specific genes, and with the 3′ ends of some tissue-specific genes. A few genes contained both 5′ and 3′ CpG islands, separated by several thousand base-pairs of CpG-depleted DNA. The 5′ CpG islands extended through 5′-flanking DNA, exons and introns, whereas most of the 3′ CpG islands appeared to be associated with exons. CpG islands were generally found in the same position relative to the transcription unit of equivalent genes in different species, with some notable exceptions.

The locations of G/C boxes, composed of the sequence GGGCGG or its reverse complement CCGCCC, were investigated relative to the location of CpG islands. G/C boxes were found to be rare in CpG-depleted DNA and plentiful in CpG islands, where they occurred in 3′ CpG islands, as well as in 5′ CpG islands associated with tissue-specific and housekeeping genes. G/C boxes were located both upstream and downstream from the transcription start site of genes with 5′ CpG islands. Thus, G/C boxes appeared to be a feature of CpG islands in general, rather than a feature of the promoter region of housekeeping genes.

Two theories for the maintenance of a high frequency of CpG dinucleotides in CpG islands were tested: that CpG islands in methylated genomes are maintained, despite a tendency for ^{5m}CpG to mutate by deamination to TpG+CpA, by the structural stability of a high G+C content alone, and that CpG islands associated with exons result from some selective importance of the arginine codon CGX. Neither of these theories could account for the distribution of CpG dinucleotides in the sequences analysed. Possible functions of CpG islands in transcriptional and post-transcriptional regulation of gene expression were discussed, and were related to theories for the maintenance of CpG islands as "methylation-free zones" in germline DNA.

1. Introduction

In vertebrate DNA, the dinucleotide CpG occurs at only 0·25 to 0·2 of the frequency expected from the base composition (Josse *et al.*, 1961; Swartz *et al.*, 1962). However, the extent of this "CpG depletion" is not uniform throughout vertebrate genomes (Smith *et al.*, 1983; Lennon & Fraser, 1983; Adams & Eason, 1984). The suggestion that CpG dinucleotides might be asymmetrically distributed within a gene was put forward by McClelland & Ivarie (1982), who analysed a composite DNA sequence derived from a number of mammalian genes. Subsequently, two groups of workers identified unusual stretches of vertebrate DNA with close to the expected frequency of CpG dinucleotides. Tykocinski & Max (1984) found regions of this type associated with several otherwise CpG-depleted genes: chicken α2(I) collagen, mouse

0022-2836/87/140261-22 $03.00/0

dihydrofolate reductase (DHFR)† and a number of mammalian major histocompatibility complex (MHC) class I and II genes. Bird *et al.* (1985) isolated, from a mouse genomic DNA library, three regions consisting of single-copy DNA with numerous *Hpa*II sites. Each region, about 1 kb in length, contained *Hpa*II sites at a frequency that would occur only in a sequence with an unusually high G+C content and around the expected frequency of CpG dinucleotides.

Stretches of DNA with a high G+C content, and a frequency of CpG dinucleotides close to the expected value, appear as CpG clusters within the CpG-depleted bulk DNA, and are now generally known as CpG islands. Recently, regions with the sequence characteristics of CpG islands have been found associated with a number of genes, most of which are "housekeeping" genes (for reviews, see Cooper & Gerber-Huber, 1985; Bird, 1986). Almost all CpG islands identified to date are associated with the 5′ ends of genes. However, two regions with clustered *Hpa*II sites have been found near the 3′ end of the human glucose-6-phosphate dehydrogenase gene (Toniolo *et al.*, 1984) and a region with the expected frequency of CpG dinucleotides has been found in and around an internal exon of some mammalian MHC class II genes (Tykocinski & Max, 1984).

The most commonly accepted explanation for the CpG depletion of vertebrate genomes relates to the tendency of 5-methylcytosine (^{5m}C) to mutate by deamination to thymine (Coulondre *et al.*, 1978). Vertebrate DNA is highly methylated at cytosine in CpG dinucleotide pairs, so CpG dinucleotides in methylated regions of germline DNA should tend to mutate to TpG and its complement CpA (Salser, 1977; Bird, 1980). Indeed, in all animal groups studied, (1) both total genomic DNA (Bird, 1980) and most specific DNA sequences (Tykocinski & Max, 1984; Smith *et al.*, 1983; McClelland & Ivarie, 1982) (2) the extent of CpG depletion, show a positive correlation between the level of CpG methylation and the extent of TpG+CpA elevation. If the relationship between CpG depletion, CpG methylation and ^{5m}C deamination is real, then regions with the expected frequency of CpG dinucleotides must either be maintained in an unmethylated state in germline DNA (Tykocinski & Max, 1984) or be protected from deamination. Resistance of methylated CpG islands to deamination could result from the DNA structure (Adams & Eason, 1984) or from a high selective importance of ^{5m}CpG dinucleotides in these regions (Cooper & Gerber-Huber, 1985).

The methylation status of most CpG islands is not known, but all CpG islands for which methylation levels have been estimated contain, at the most, very low levels of ^{5m}C. The three CpG islands identified by Bird *et al.* (1985) in mouse genomic DNA are largely or completely unmethylated at all *Hpa*II sites in all tissues examined, including sperm. CpG islands which are hypomethylated in a number of tissues have also been identified in or near a few known genes: human and mouse hypoxanthine phosphoribosyl transferase (HPRT) in the active X chromosome (Yen *et al.*, 1984; Wolf *et al.*, 1984*a*; Lock *et al.*, 1986), hamster adenine phosphoribosyl transferase (APRT; Stein *et al.*, 1983), mouse DHFR (Stein *et al.*, 1983; Tykocinski & Max, 1984), chicken α2(I) collagen (McKeon *et al.*, 1982; Tykocinski & Max, 1984), and human glucose-6-phosphate dehydrogenase (Toniolo *et al.*, 1984; Wolf *et al.*, 1984*b*; Battistuzzi *et al.*, 1985).

We have screened an extensive set of vertebrate genomic DNA sequences, containing genes transcribed by RNA polymerase II, and have identified CpG islands within these sequences. We have studied the extent and location of each CpG island relative to the transcription unit and the exon structure of the associated gene, compared CpG islands associated with equivalent genes from different species, studied the association of CpG islands with housekeeping and tissue-specific genes, analysed the distribution of G/C boxes within and outside CpG islands, and tested various theories for the maintenance of CpG islands. We discuss these results in terms of theories for the function of CpG islands.

2. Methods

The GenBank DNA sequence data bank (December 1985 issue) was screened for all vertebrate genomic sequences containing genes or parts of genes transcribed by RNA polymerase II. Of these, sequences which extended more than 200 bp upstream from the translation start site were selected for analysis, with the exception of sequences of genes such as immunoglobulins that require rearrangement of genomic DNA for transcription.

The ratio observed/expected (Obs/Exp or O/E) CpG was calculated as follows:

$$\text{Obs/Exp CpG} = \frac{\text{Number of CpG}}{\text{Number of C} \times \text{number of G}} \times N,$$

where N is the total number of nucleotides in the sequence being analysed. A moving average value for percentage (%) G+C and for Obs/Exp CpG was calculated for each sequence, using a 100 bp window (N = 100) moving across the sequence at 1 bp intervals.

For this study, CpG-rich regions were defined as stretches of DNA where both the moving average of %G+C was greater than 50, and the moving average of Obs/Exp CpG was greater than 0·6. Since one of the

† Abbreviations used: DHFR, dihydrofolate reductase; MHC, major histocompatability complex; kb, 10^3 base-pairs; HPRT, hypoxanthine phosphoribosyl transferase; APRT, adenine phosphoribosyl transferase; bp, base-pair(s); Obs/Exp or O/E, ratio of observed value/expected value; EGF receptor, epidermal growth factor receptor; GAPDH, glyceraldehyde-3-phosphate dehydrogenase; HMG CoA reductase, 3-hydroxy-3-methylglutaryl coenzyme A reductase; PGK, 3-phosphoglycerate kinase; CRF, corticotropin releasing factor; IGF-II, insulin-like growth factor II; snRNA, small nuclear RNA; AIDS retrovirus LTR, acquired immune deficiency syndrome retrovirus long terminal repeat.

primary aims of the study was to identify genes with 5′ CpG islands, it was important to avoid misclassifying any gene which might contain an unsequenced CpG island in the DNA immediately upstream from the transcription start site. Therefore, of the sequences which contained no CpG-rich regions greater than 200 bp in length, those sequences that extended less than 200 bp upstream from the transcription start site were eliminated from the study. The study included a number of sequences that were not in the December 1985 issue of GenBank. These were published sequences which satisfied the selection criteria and contained CpG-rich regions more than 200 bp in length. Where two genes were identical or very similar in sequence, one was omitted. However, if two sequences were very similar within exons, but differed within introns or at the 5′ or 3′ end (e.g. human $\alpha 1$ and $\alpha 2$ globins), they were included as two separate entries in the study.

Analysis was carried out using software available on MBIS (Bucholtz & Reisner, 1986) and some graphics and statistical programs developed by Simon Worthington.

3. Results and Discussion

All previously identified CpG islands include several hundred base-pairs of G+C-rich DNA with a CpG content close to the level expected from the base composition. As the significance of CpG islands is unclear, a precise definition of the sequence requirements for a CpG island is not possible. For the purpose of this survey, regions of DNA with a moving average of %G+C over 50 and Obs/Exp CpG over 0·6 have been classed as CpG-rich regions. CpG-rich regions over 200 bp in length are unlikely to have occurred by chance alone, so, as a working definition, have been labelled as CpG islands.

(a) *Sequences analysed*

To identify, locate and characterize any CpG islands present, we plotted Obs/Exp CpG and %G+C against the position in the sequence for all sequences in the survey. We also plotted the position of each CpG dinucleotide along the sequence, relative to the exon–intron boundaries of the gene. We plotted the position of each GpC dinucleotide as a comparison, since the frequency of GpC dinucleotides is generally indicative of the G+C content of a region. Figures 1 and 2 show examples of such plots for the various classes of genes discussed below.

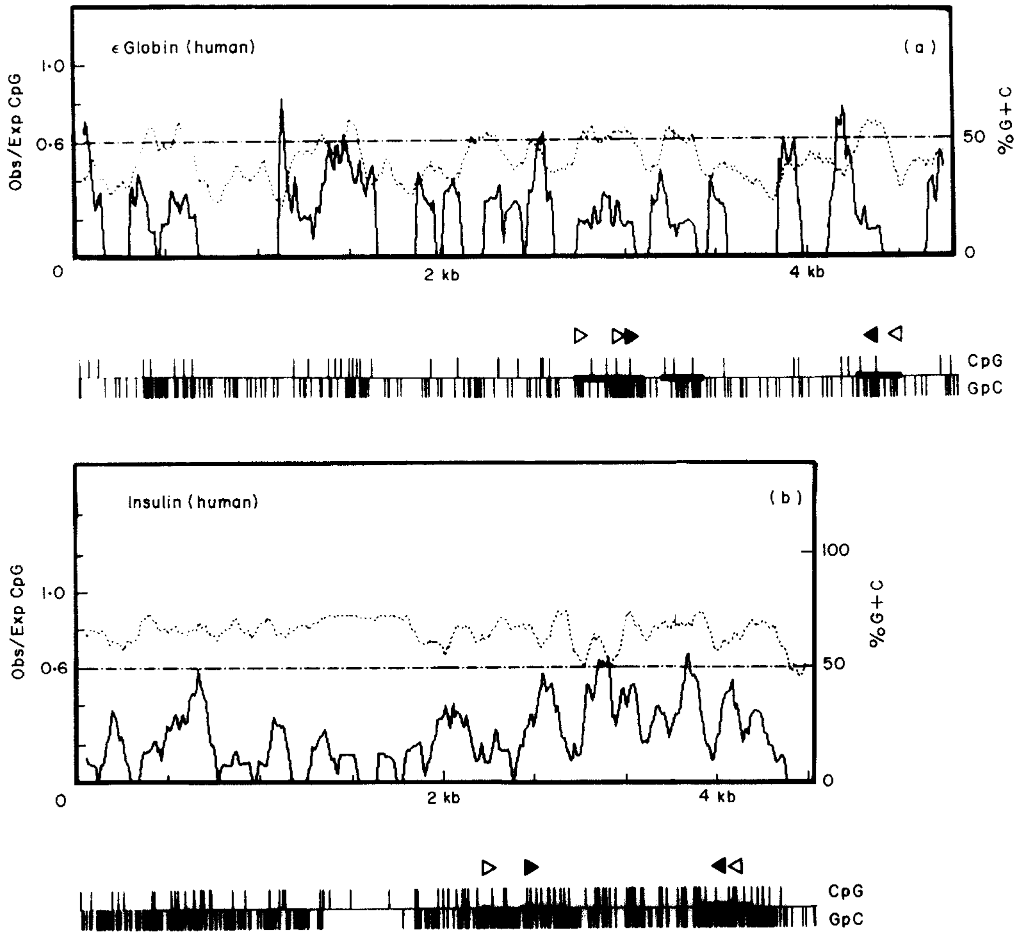

Figure 1. Analysis of the distribution of CpG dinucleotides in two CpG-depleted genes. Moving averages of %G+C and Obs/Exp CpG were calculated as described in Methods; each point on the graph represents the average values for 10 adjacent 100 bp windows; values for %G+C are plotted as a broken line and values for Obs/Exp CpG are plotted as a continuous line. Underneath the graph, the position of each CpG and GpC dinucleotide relative to position in the sequence is indicated by a vertical line; exons are marked by heavy horizontal bars, mRNA start and stop sites by open triangles, and peptide initiation and termination codons by filled triangles.

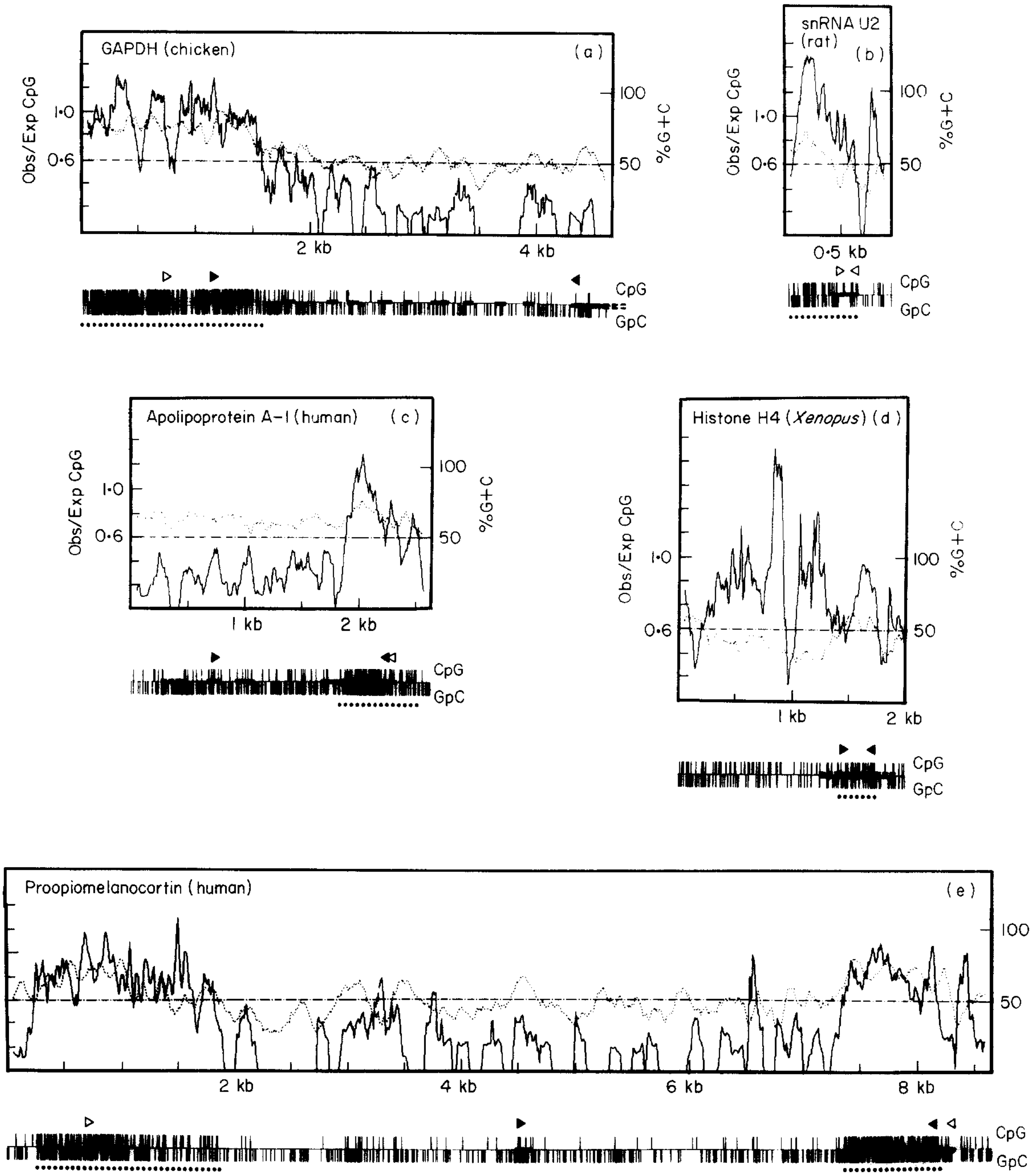

Figure 2. Analysis of the distribution of CpG dinucleotides in several genes with CpG islands. %G+C, Obs/Exp CpG, and the location of individual CpG and GpC dinucleotides along the sequence are plotted as described in the legend to Fig. 1; the extent of each CpG island is shown by a dotted line beneath the plots; where the full extent of an exon is not known, the horizontal bar representing the exon is flanked by two dashed lines.

In Tables 1 and 2 we have listed all genes included in the survey. For each gene, we have listed the total length of sequenced DNA and have described the portion of the gene contained in the sequence. Where no CpG island was present in the sequence (Table 1), we have included a calculation of %G+C and Obs/Exp CpG for the total sequence. Where a CpG island was present in the sequence (Table 2), we have described the location of each CpG island relative to the exons and introns of the associated gene, and have noted where sufficient sequence is available to define the approximate boundaries of the CpG island. We have included a calculation of the %G+C and Obs/Exp CpG for each CpG island and for the remainder of the sequence.

(i) *CpG-depleted genes*

We found many gene sequences which contain no CpG islands (Table 1). We have called these CpG-depleted genes (e.g. Fig. 1). Even though DNA sequences are largely non-random due to structural and functional constraints, we expected to find many small CpG-rich regions arising by chance. The

Table 1

CpG-depleted sequences

Gene		Extent of total sequence	Obs/Exp CpG	%G+C	N (bp)	References†
α-1 Antitrypsin (S variant)	Human	Complete	0·23	51	12,222	HUMA1ATP
Apo very low density lipoprotein II	Chicken	Complete	0·12	43	3401	CHKAPOLP2
Asialoglycoprotein receptor RHL-1	Rat	Complete	0·21	54	4093	RATRHL1
Atrial natriuretic factor	Human	Complete	0·32	53	2710	HUMANFA
Chymotrypsin B	Rat	Complete‡	0·28	52	5809	RATCTRPB
α Crystallin	Mouse	5′ Flank to IVS3	0·28	54	2380	MUSCRYA5A
γ Crystallin	Rat	Complete	0·31	49	3430	RATCRYG
Cytochrome P-450C	Rat	Complete	0·18	48	6917	RATCYP45C
Elastase I	Rat	5′ Flank to 3′ flank (IVSs 2,4–7 incomplete)	0·35	54	4640	RATELAI1–6
Elastase II	Rat	5′ Flank to 3′ flank (IVSs 2–7 incomplete)	0·32	53	4663	RATELAII1–7
Factor IX	Human	Complete	0·20	39	38,059	HUMFIXG
α Fetoprotein	Mouse	5′ Flank to IVS1	0·38	39	1032	MUSAFP14Z
γ,γ′Fibrinogen	Human	Complete	0·13	37	10,564	HUMFBRG
α Fibrinogen	Rat	5′ Flank to IVS1	0·17	43	2223	RATFBA5E
β Fibrinogen	Rat	5′ Flank to IVS1	0·44	37	2223	RATFBB5E
γ Fibrinogen	Rat	5′ Flank to IVS3	0·29	46	2222	RATFBG5E
β Globin	Bovine	5′ Flank to E3 non-coding	0·17	45	2072	BOVHBB
γ Globin	Bovine	5′ Flank to 3′ non-coding (mRNA end not sequenced)	0·25	47	2048	BOVHBG
α^D Globin	Chicken	Complete	0·25	62	1468	CHKHBADA1
β Globin	Chicken	5′ Flank to 3′ non-coding (mRNA end not known)	0·34	59	2157	CHKHBBCOM
ε Globin	Chicken	Complete	0·28	56	1955	CHKHBBR2
ε_1 Globin	Goat	Complete‡	0·13	44	2221	GOTHBBEI
ε_2 Globin	Goat	Complete‡	0·22	45	2278	GOTHBBEII
δ Globin-β globin cluster	Human	Complete	0·17	38	16,489	HUMHBB5
ε Globin	Human	Complete‡	0·21	41	4805	HUMHBB2
G_γ Globin-A_γ globin cluster	Human	Complete	0·19	41	11,376	HUMHBB3
α Globin	Mouse	Complete‡	0·26	55	1441	MUSHBA
βh0 Globin	Mouse	Complete	0·09	43	2135	MUSHBBH0
βh1 Globin	Mouse	Complete	0·09	42	1926	MUSHBBH1
β Globin, minor	Mouse	Complete	0·14	43	1830	MUSHBBMIN
β_1 Globin, type 1 allele	Rabbit	Complete	0·14	43	1827	RABHBB1A1
β_4 Globin	Rabbit	Complete	0·15	40	2630	RABHBB4
β_1 Globin	*Xenopus*	Complete	0·28	35	2972	XENHBBI
Growth Hormone	Rat	Complete	0·36	52	2557	RATGH1
Growth hormone releasing factor	Human	5′ Flank to 3′ flank (IVSs incomplete)	0·31	50	1157	HUMGRFP1–5
Hepatic product spot 14	Rat	Complete	0·37	52	1606	RATSPOT14
Insulin	Human	Complete	0·22	66	4044	HUMINS1
α Interferon, leukocyte a	Human	Complete	0·09	36	1733	HUMIFNAA
α Interferon, leukocyte h2	Human	Complete	0·02	36	1612	HUMIFNAH2
β Interferon, fibroblast	Human	Complete	0·13	40	1835	HUMIFNB1F
γ Interferon, immune	Human	Complete	0·17	37	5961	HUMIFNG
Interleukin 2	Human	Complete	0·10	32	5561	HUMIL2B
Kallikrein-1, glandular	Mouse	Complete	0·12	48	9433	MUSGKAL1
Luteinizing hormone	Rat	Complete	0·30	58	1798	RATLHB
MHC class II, HLA-DR-α	Human	Complete	0·27	43	5724	HUMMHDRHA
Myoglobin	Human	5′ Flank to 3′ flank (IVSs incomplete)	0·19	51	6975	HUMMG1–3
Myoglobin	Seal	5′ Flank to IVS2 (IVS1 incomplete)	0·31	57	1101	SEAMG1–2
Myosin, cardiac α, heavy chain	Rat	(a) 5′ Flank to IVS3 (b) Last intron to 3′ flank	0·30	53	3848	RATMYHAB2
Myosin, fast alkali light chains, 1f & 3f	Chicken	5′ Flank to 3′ flank (IVS1, chain 1f, and IVS2, chains 1f and 3f, incomplete)	0·31	45	6680	CHKMLC131–133
Opsin	Human	5′ Flank to 3′ non-coding (mRNA end not known)	0·26	55	6953	HUMOPS
Ovalbumin	Chicken	Complete	0·16	38	9206	CHKOVAL
Parathyroid hormone	Bovine	Complete	0·21	31	3154	BOVPTHG
Parathyroid hormone	Rat	5′ Flank to 3′ flank (IVS1 incomplete)	0·20	40	2457	RATPTH1–3
Phosphoenolpyruvate carboxykinase, GTP	Rat	5′ Flank to IVS2	0·36	54	1196	RATPEC
Placental lactogen	Human	Complete	0·33	54	2967	HUMPLA
Prolactin	Bovine	5′ Flank to E2 (IVS1 incomplete)	0·15	40	1289	BOVPRLP1–2
Prolactin	Rat	5′ Flank to IVS1	0·18	38	1064	RATPRLHR1

Table 1 *(continued)*

Gene		Extent of total sequence	Obs/Exp CpG	%G+C	N (bp)	References†
Prostatic steroid binding protein c1	Rat	5′ Flank to 3′ flank (IVS1 incomplete)	0·34	38	728	RATPSBC11–12
Prostatic steroid binding protein c2	Rat	5′ Flank to IVS1	0·00	39	379	RATPSBC21
Prostatic steroid binding protein c3(1)	Rat	5′ Flank to 3′ flank (IVSs incomplete)	0·18	42	1287	RATPSBPA1–3
Renin 1, kidney	Mouse	5′ Flank to IVS1	0·13	51	641	MUSREN1K
Renin 2, submaxillary gland	Mouse	5′ Flank to IVS1	0·11	48	640	MUSREN2SM
Renin	Human	5′ Flank to 3′ flank (IVSs incomplete)	0·28	55	2724	HUMREN01–10
Uteroglobin	Rabbit	Complete	0·28	54	3709	RABUG
Vitellogenin A2	*Xenopus*	5′ Flank to IVS3	0·30	34	1028	XENVIT
Vitellogenin, major vtgii	Chicken	5′ Flank to IVS3	0·28	41	1632	CHKVITII2
x Gene, ovalbumin gene family	Chicken	(a) 5′ Flank to IVS1 (b) IVS4 to 3′ non-coding (IVSs incomplete)	0·16	42	2940	CHKX1–4
y Gene, ovalbumin gene family	Chicken	Complete	0·13	37	8372	CHKY

E, exon; IVS, intervening sequence or intron.

† GenBank code.

‡ The location of the mRNA end is putative.

majority of CpG-depleted genes, in our survey, do contain small CpG-rich regions, mostly less than 50 bp in length. Only five of the CpG-depleted genes contain CpG-rich regions approaching 200 bp in length that are at all comparable, in terms of %G+C and Obs/Exp CpG, with previously identified CpG islands and with CpG islands identified in this study (Table 2). These regions are found in the genes coding for chicken β globin and ε globin (5′ ends), as well as human factor IX (intron 6), mouse α globin (exon 2) and rat chymotrypsin B (exon 6–intron 6). It is possible that a number of these small CpG-rich regions have not arisen by chance, and may even fulfil the same function as CpG islands.

We note that, although most of the sequences listed in Table 1 include an entire transcription unit, some sequences are incomplete, and may be found to contain CpG-rich regions in the currently unsequenced parts of the gene. However, by virtue of the design of this study, we can be certain that none of the genes listed in Table 1 has CpG islands, as they are currently understood, at their 5′ ends.

(ii) *Genes with 5′ CpG islands*

Table 2 lists sequences which contain CpG islands. The genes in Table 2A to D have CpG islands that begin upstream from the translation start site. We have called these 5′ CpG islands (e.g. Fig. 2(a)). Table 2A lists protein-coding genes where the CpG island begins in the non-transcribed 5′-flanking DNA. The bulk of CpG islands identified by our screening procedure belong to this category. Table 2B lists two genes where the CpG island begins between the transcription and translation start sites. Table 2C lists genes where the CpG island begins upstream from the translation start site, but the transcription start site either has not been definitely located or has not been sequenced. For most of the genes in Table 2C, the CpG island begins well upstream from the translation start site. Table 2D lists two genes which contain a long CpG-rich region beginning upstream from the translation start site, and another CpG rich region separated from it by a large stretch of unsequenced DNA. These regions could be part of the same CpG island or part of two separate CpG islands. Table 2E lists small nuclear RNA (snRNA) genes, all of which contain 5′ CpG islands beginning upstream from the transcription start site (e.g. Fig. 2(b)). These were included in the analysis since they are transcribed by RNA polymerase II, even though they are not translated.

(iii) *Genes with 3′ CpG islands*

Although our study was primarily designed to identify CpG islands at the 5′ ends of genes, we found a number of genes with CpG islands that lie entirely downstream from the translation start site. We have called these 3′ CpG islands. Table 2F lists genes which contain 3′ CpG islands (e.g. Fig. 2(c)). Table 2G lists genes which contain both 5′ CpG islands and 3′ CpG islands, clearly separated by at least 1 kb of CpG-depleted sequence (e.g. Fig. 2(e)).

(iv) *Histones*

Histone genes have proved to be unusual in several respects (e.g. Fig. 2(d)), so they are listed separately in Table 2H.

(b) *Extent and location of CpG islands*

(i) *CpG islands are generally over 500 bp in length and include regions with a very high G+C content and the expected number of CpG dinucleotides*

We identified the presence of CpG islands using the criterion of 200 bp or more of CpG-rich DNA,

Table 2
Sequences containing CpG islands

		Total sequence		CpG island sequence				Remaining sequence			
Gene		Extent	*N* (bp)	Extent	O/E CpG	% G+C	*N* (bp)	O/E CpG	% G+C	*N* (bp)	References†
A. *Genes with 5′ CpG islands starting upstream from the transcription start site*											
α Actin, cardiac	Chicken	Complete‡	6109	⟨5′ Flank to IVS2⟩	0·95	70	1070	0.23	40	5039	Eldridge *et al.* (1985)
α Actin, skeletal	Chicken	Complete	2426	← 5′ Flank to IVS2⟩	0·95	73	500	0·36	53	1926	CHKACTA
β Actin	Chicken	Complete	5046	⟨5′ Flank to IVS2⟩	1·01	73	1390	0·26	51	3656	CHKACB
β Actin	Rat	Complete‡	4098	. . .5′ Flank to E2⟩	0·97	67	1250	0·35	49	2848	RATACTB
Adenosine deaminase	Human	5′ Flank to 3′ flank‡ (IVS 1–6, 8–11 incomplete)	2229	← 5′ Flank to IVS1 →	1·03	79	272	0·32	57	1957	Valerio *et al.* (1985)
5-Aminolevulinate synthase	Chicken	5′ Flank to 3′ flank‡ (IVS7 incomplete)	8542	⟨5′ Flank to IVS1⟩	0·86	69	590	0·22	47	7952	Maguire *et al.* (1986)
APRT	Mouse	Complete	3066	⟨5′ Flank to IVS2⟩	0·88	66	560	0·24	54	2506	Dush *et al.* (1985)
Argininosuccinate synthetase	Human	5′ Flank to 3′ flank (IVS 1–4, 7–9 incomplete)	2843	. . .5′ Flank to IVS1 →	0·73	72	945	0·50	56	1898	Jinno *et al.* (1985), HUMAS2–8
c-fos	Human	Complete	4165	. . .5′ Flank to E2⟩	0·76	63	1900	0·25	49	2265	HUMCFOS
c-fos	Mouse	5′ Flank to 3′ non-coding (mRNA end unknown)	3548	← 5′ Flank to IVS2⟩	0·67	60	1520	0·36	48	2028	MUSFOS
c-Ha-ras1	Human	Complete	6453	← 5′ Flank to E4⟩	0·63	71	3460	0·31	65	2993	HUMRASH, Ishii *et al.* (1985)
c-myc	Chicken	Complete	4984	← 5′ Flank to E3 non-coding⟩	0·98	66	3450	0·57	45	1534	CHKMYC
c-myc	Human	Complete	8082	⟨5′ Flank to IVS2⟩	0·79	61	2510	0·49	46	4572	HUMMYCC
c-myc	Mouse	5′ Flank to 3′ flank (IVS2 incomplete)	4257	← 5′ Flank to IVS2 →	0·77	59	3306	0·58	46	951	MUSCMYC1–2
c-sis	Human	5′ Flank to IVS2 (IVS1 incomplete)	1788	← 5′ Flank to IVS1 →	0·80	70	1192	0·43	61	596	HUMSISA1–2
Collagen, α2(I)	Chicken	(a) 5′ Flank to IVS6 (IVS 1, 2, 4, 5 incomplete) (b) IVS8 to IVS10 (IVS9 incomplete) (c) IVS15 to IVS22 (IVS20 incomplete) (d) IVS23 to IVS25 (IVS24 incomplete) (e) IVS40 to E42 (IVS41 incomplete) (f) E42 to 3′ non-coding (IVS 42–44, 46–49 incomplete)	8727	← 5′ Flank to IVS1 →	1·00	63	680	0·24	42	8047	CHKC1A201–226
Collagen (II)	Rat	5′ Flank to IVS1	1404	⟨5′ Flank to IVS1 →	0·66	66	854	0·30	57	550	Kohno *et al.* (1985)
DHFR	Human	5′ Flank to 3′ flank (IVS2–5 incomplete)	5866	⟨5′ Flank to IVS1. . .	0·79	69	1040	0·26	40	4826	HUMFOL1-5
DHFR	Mouse	5′ Flank to 3′ flank (IVSs incomplete)	3403	. . .5′ Flank to IVS1 →	0·85	64	1095	0·33	38	2308	McGrogan *et al.* (1985), Dynan *et al.* (1986), MUSFOL2–7
EGF receptor	Human	5′ Flank to IVS1	561	← 5′ Flank to IVS1 →	0·99	76	561	—	—	—	Ishii *et al.* (1985)
Enkephalin	Human	5′ Flank to 3′ flank (IVS2 incomplete)	7616	(1) ← 5′ Flank to IVS2⟩ (2) ⟨Downstream from gene. . .	0·84 0·74	64 62	2020 390	0·30	39	5206	HUMENKA, HUMENKPH1–2

Table 2 *(continued)*

		Total sequence		CpG island sequence				Remaining sequence			
Gene		Extent	*N* (bp)	Extent	O/E CpG	% G+C	*N* (bp)	O/E CpG	% G+C	*N* (bp)	References†
Enkephalin	Rat	5′ Flank to 3′ flank (IVSs incomplete)	2655	. . .5′ Flank to IVS1 →	0·93	63	616	0·39	48	2039	RATENK1 3
ρ Globin	Chicken	Complete	1554	. . .5′ Flank to E2⟩	0·63	62	580	0·21	57	974	CHKHBBR1
α^A Globin	Duck	Complete	1145	. . .5′ Flank to E1⟩	0·78	74	320	0·39	59	825	DUKHBADA2
α1 Globin	Goat	Complete‡	1894	⟨5′ Non-coding to 3′ flank‡. . .	0·73	68	920	0·25	54	974	GOTHBAI
α2 Globin	Goat	Complete‡	1691	. . .5′ Non-coding to 3′ flank‡. . .	0·71	66	1400	0·18	55	291	GOTHBAII
α Globin cluster	Human		12,847	(1) . . .far upstream of gene⟩	0·73	68	520	0·26	54	9707	HUMHBA4
α2		Complete		(2) ⟨5′ Flank to 3′ flank⟩	0·82	71	1340				
α1		Complete		(3) ⟨5′ Flank to E3 non-coding⟩	0·82	72	1280				
GAPDH†	Chicken	Complete‡	4645	← 5′ Flank to IVS2⟩	0·91	74	1600	0·22	50	304	Stone *et al.* (1985)
Heat shock protein hsp70	Human	Complete	2691	. . .5′ Flank to coding⟩	0·82	63	2210	0·20	46	481	Hunt & Morimoto (1985)
Heat shock protein hsp70	*Xenopus*	Complete‡	2574	. . .5′ Flank to coding⟩	1·04	50	500	0·26	45	2074	Bienz (1984)
HMG CoA reductase	Hamster	5′ Flank to E2 (IVS1 incomplete)	1101	← 5′ Flank to E1⟩	0·88	68	690	0·49	51	411	Reynolds *et al.* (1984)
HPRT	Mouse	5′ Flank to E9 non-coding	2575	⟨5′ Flank to IVS1 →	0·87	72	320	0·35	37	2255	MUSHPRT1–9
int-1	Human	Complete‡	4522	← 5′ Flank to 3′ non-coding⟩	0·78	65	3580	0·27	55	942	van Ooyen *et al.* (1985)
int-1	Mouse	Complete‡	4511	. . .5′ Flank to 3′ non-coding⟩	0·66	61	3370	0·32	52	1141	MUSINT1
Metallothionein-1A	Human	Complete	2941	. . .5′ Flank to IVS1⟩	0·82	66	770	0·21	48	2171	HUMMET1A
Metallothionein-II	Human	Complete	1703	⟨5′ Flank to IVS1⟩	0·92	69	501	0·39	50	1202	HUMMET2
Metallothionein-I	Mouse	Complete	1560	← 5′ Flank to IVS1⟩	0·92	57	620	0·39	52	940	MUSMETI
Metallothionein-II	Mouse	Complete	1400	← 5′ Flank to IVS2⟩	0·76	60	900	0·38	51	500	MUSMETII
Oxytocin-neurophysin	Bovine	Complete	1167	← 5′ Flank to E3 non-coding. . .	0·81	73	1050	0·39	51	117	BOVOT
Oxytocin-neurophysin	Rat	Complete	1053	. . .E1 non-coding to E3 non-coding. . .	0·70	66	800	0·07	47	253	RATOXTNP
PGK	Human	5′ Flank to IVS1	812	← 5′ Flank to IVS1. . .	1·06	65	680	0·37	58	132	Singer-Sam *et al.* (1984)
Ribosomal protein L30	Mouse	Complete	3475	← 5′ Flank to IVS2⟩	0·86	63	770	0·31	38	2705	Wiedemann & Perry (1984)
Ribosomal protein L32	Mouse	Complete	3757	⟨5′ Flank to IVS1⟩	0·97	63	770	0·25	48	2987	MUSRPL3A
Ribosomal protein S16	Mouse	Complete	2647	← 5′ Flank to IVS2⟩	0·83	60	1060	0·28	46	1587	Wagner & Perry (1985)
Somatostatin-I	Human	Complete	2667	⟨5′ Flank to IVS1⟩	0·80	64	340	0·33	42	2327	HUMSOMI
Somatostatin-14	Rat	Complete	2021	⟨5′ Flank to IVS1⟩	0·75	61	550	0·33	46	1471	RATSOM14M
Superoxide dismutase-1	Human	5′ Flank to 3′ flank (IVSs incomplete)	2289	← 5′ Flank to IVS1 →	0·96	66	538	0·21	38	1751	Levanon *et al.* (1985)
Triosephosphate isomerase	Human	5′ Flank to 3′ flank (IVSs incomplete)	1835	. . .5′ Flank to IVS1 →	0·94	69	452	0·31	52	1383	Brown *et al.* (1985)
β Tubulin	Human	Complete	5117	← 5′ Flank to IVS2⟩	0·75	56	2480	0·32	48	2637	HUMTBBM40
Urokinase	Human	Complete	7259	⟨5′ Flank to IVS2⟩	0·82	73	680	0·27	52	6579	Riccio *et al.* (1985)
Vimentin	Hamster	5′ Flank to 3′ flank (IVS1, 2, 4, 5, 7, 8 incomplete)	4090	. . .5′ Flank to IVS1. . .	0·77	63	840	0·27	43	3250	HAMVIM1–7
B. *Genes with 5′ CpG islands starting between the transcription and translation start sites*											
CRF	Human	Complete	2685	⟨E1 to E2⟩	0·83	63	1340	0·36	42	1345	HUMCRF
Urokinase	Porcine	Complete‡	7143	⟨IVS1 to IVS2⟩	0·74	65	400	0·22	51	6743	Nagamine *et al.* (1984)
C. *Genes with 5′ CpG islands where the transcription start site is not known*											
c-Ki-ras2	Human	E1 to 3′ non-coding (mRNA start unsequenced)	3918	← E1 to IVS1 →	0·87	77	741	0·15	33	3177	HUMRASK21 26

Carbonic anhydrase II	Human	5′ Non-coding to IVS2 (mRNA start unknown)	2150	← 5′ non-coding to IVS1⟩	0·87	72	1350	0·35	46	800	Venta *et al.* (1985)
Carbonic anhydrase II	Mouse	5′ Non-coding to IVS2 (IVS1 incomplete) (mRNA start unknown)	1298	← 5′ Non-coding to IVS1 →	0·82	70	552	0·43	45	746	Venta *et al.* (1985)
Cytochrome *c*, allele CC10	Chicken	5′ Non-coding to 3′ flank (mRNA start unknown)	1620	← 5′ Non-coding⟩	0·87	73	300	0·28	38	1320	CHKCYC10
MHC, class I HLA	Human	5′ Non-coding to 3′ flank‡ (mRNA start unknown)	4123	← 5′ Non-coding to IVS3⟩	0·86	66	1270	0·15	51	2853	HUMMH
MHC, class I HLA-A3	Human	5′ Non-coding to 3′ non-coding (mRNA start & end unknown)	3717	. . .5′ Non-coding to IVS3⟩	0·82	68	1190	0·17	53	2527	HUMMHA3
Thymidine kinase	Chicken	5′ Flank to 3′ non-coding (mRNA start putative & mRNA end unknown)	3008	. . .5′ Flank to IVS3⟩	0·90	75	1300	0·23	51	1708	CHKTK
D. *Genes with 5′ CpG islands separated by unsequenced DNA from CpG-rich regions further downstream*											
IGF-II	Human	E1 non-coding to 3′ non-coding (mRNA start unsequenced & mRNA end unknown)	2127	(1) ← E1 non-coding to IVS1 → (2) . . .IVS3 to 3′ non-coding. . .	0·96 0·72	72 68	1101 360	0·32	51	498	Dull *et al.* (1984)
Retinol binding protein	Human	5′ flank to IVS5 (IVS4 incomplete)	2127	(1) ⟨5′ Flank to IVS3. . . (2) . . .IVS4 to IVS5. . .	0·85 0·65	72 59	1040 300	0·38	50	787	D'Onofrio *et al.* (1985)
E. *Genes with 5′ CpG islands—small nuclear RNA genes*											
snRNA, U1-52a	Chicken	Complete	690	← 5′ Flank to 3′ flank →	0·99	67	690	—	—	—	CHKUG1521
snRNA, U1-52b (comp)	Chicken	Complete	653	← 5′ Flank to 3′ flank →	0·98	66	653	—	—	—	CHKUG1522
snRNA, U1-52c (comp)	Chicken	Complete	601	← 5′ Flank to 3′ flank →	0·78	61	601	—	—	—	CHKUG1523
snRNA, U1	Human	Complete	806	. . .5′ Flank to 3′ flank →	0·74	55	736	0·20	53	70	HUMUG1
snRNA, U1 clone 6-6B	Rat	Complete	1160	⟨5′ Flank to transcribed⟩	1·03	60	450	0·51	47	710	RATUG12
snRNA, U2	Rat	Complete	930	← 5′ Flank to 3′ flank⟩	0·95	57	620	0·56	37	310	RATUG2A
snRNA, U2	*Xenopus*	Complete	831	← 5′ Flank to transcribed⟩	0·67	66	370	0·30	52	461	XENUG2
F. *Genes with 3′ CpG islands*											
α Actin, skeletal	Mouse	Complete‡	4007	⟨E3 to E7 non-coding⟩	0·67	61	1310	0·30	54	2697	Hu *et al.* (1986)
α Actin, skeletal	Rat	5′ Flank to E7 non-coding	3033	⟨E2 to E7 non-coding →	0·62	59	1803	0·23	55	1230	RATACTSK
Apolipoprotein A-1	Human	Complete (mRNA start putative)	2637	⟨E4 to 3′ flank. . .	0·85	66	630	0·24	61	2007	HUMAPOAI1
Apolipoprotein E	Human	Complete	5515	⟨IVS3 to E4 non-coding⟩	0·76	73	910	0·31	57	4605	Das *et al.* (1985)
α^A Globin	Chicken	Complete	1216	⟨E2 to E3 non-coding. . .	0·64	65	460	0·33	61	756	CHKHBADA2
ζ Globin	Human	Complete	2685	⟨IVS1 to 3′ flank. . .	0·86	76	1060	0·18	57	1625	HUMHBA1
MHC, class II HLA-DC-β	Human	Complete‡ (mRNA start putative)	7272	⟨IVS1 to IVS2⟩	0·77	67	990	0·21	44	628	HUMMHDCB
MHC, class II HLA-DC-3β	Human	5′ Flank to E5 non-coding (mRNA end unknown)	8090	(1) ⟨IVS1 to IVS2⟩ (2) ⟨IVS2⟩	0·85 0·72	67 61	630 270	0·22	44	7190	HUMMHDC3B
MHC, class II H2-1A-β haplotype b	Mouse	Complete‡ (mRNA start putative)	10,000	⟨IVS1 to IVS2⟩	0·86	68	980	0·21	48	9020	MUSMHAB3
Vasopressin neurophysin II	Bovine	Complete	2427	⟨IVS1 to 3′ flank →	0·80	76	837	0·29	60	1590	BOVVP
Vasopressin neurophysin II	Rat	Complete‡	2358	⟨IVS1 to E3. . .	0·71	67	550	0·34	57	1808	RATVPNPA
G. *Genes with both 5′ CpG islands and 3′ CpG islands*											
Proopiomelanocortin	Bovine	5′ Flank to 3′ flank (IVSs incomplete)	2807	(1) ← 5′ Flank to IVS1 → (2) ← IVS2 to E3 non-coding. . .	0·90 0·89	66 69	354 890	0·38	50	1563	BOVPOMC3-7
Proopiomelanocortin	Human	Complete	8658	(1) ⟨5′ Flank to IVS1⟩ (2) ⟨IVS2 to E3 non-coding⟩	0·81 0·83	61 70	1610 830	0·20	46	6218	HUMPOMC

Table 2 *(continued)*

		Total sequence		CpG island sequence				Remaining sequence			
Gene		Extent	*N* (bp)	Extent	O/E CpG	% G+C	*N* (bp)	O/E CpG	% G+C	*N* (bp)	References†
Proopiomelanocortin	Mouse	5′ Flank to 3′ flank (IVSs incomplete)	1650	(1) . . .5′ Flank to E1 . . . (2) . . . E3. . .	0·72 0·85	65 64	240 550	0·30	54	860	MUSPOMC1–3
Proopiomelanocortin	Rat	5′ Flank to 3′ flank (IVSs incomplete)	1916	(1) ⟨5′ Flank to E1 → (2) . . .E3. . .	0·74 0·84	64 64	402 560	0·20	54	974	RATPOMC1–3
β Tubulin, brain	Human	5′ Non-coding to 3′ flank‡ (mRNA start unknown)	8874	(1) ⟨5′ non-coding to IVS1⟩ (2) ⟨IVS3⟩ (3) ⟨IVS3 to E4⟩	0·98 0·82 0·73	76 55 61	330 360 1450	0·30	54	6734	HUMTBB5
H. *Histone genes*											
Histone H1	Chicken	5′ Flank to 3′ non-coding (mRNA end unknown)	1098	← 5′ Flank to coding⟩	1·08	66	610	0·62	56	488	CHKH11A1
Histone H2a	Chicken	5′ Non-coding to 3′ flank‡ (mRNA start unknown)	843	. . .5′ Non-coding to coding. . .	1·03	69	470	0·84	47	373	CHKH2A
Histones H2a (comp)/H2b	Chicken	Complete‡ (mRNA starts putative)	1564	. . .Coding (H2a) to coding (H2b). . .	1·09	62	1040	0·68	46	524	CHKH2A2B
Histone H4	Chicken	5′ Flank to 3′ non-coding (mRNA end unknown)	675	. . .5′ Flank to 3′ non-coding. . .	1·09	68	460	0·44	46	215	CHKH43D8
Histone H5	Chicken	Complete	1754	⟨5′ Flank to coding⟩	0·73	65	900	0·20	52	854	CHKH5M
Histone H5	Duck	Complete	1175	. . .5′ Flank to coding⟩	0·77	68	820	0·42	43	355	DUKH5
Histone H2a	Human	Complete‡	866	⟨Coding⟩	1·08	62	340	0·27	41	526	Zhong *et al.* (1983)
Histone H2b	Human	Complete‡	843	⟨5′ Flank to 3′ non-coding. . .	0·90	56	451	0·82	47	392	Zhong *et al.* (1983)
Histone H4	Human	5′ Non-coding to 3′ flank (mRNA start unknown)	727	← 5′ Non-coding to 3′ flank →	1·07	55	727		—		HUMH4
Histone H4	Mouse	Complete	968	. . .5′ Flank to 3′ flank. . .	0·86	59	570	0·42	44	398	MUSHIST4
Histone cluster	*Xenopus*		14,949					0·47	42	12,659	Perry *et al.* (1985)
H4		Complete‡		⟨5′ Non-coding to 3′ non-coding⟩	0·90	56	340				
H2a (comp)		Complete‡		⟨5′ Non-coding to coding⟩	0·87	49	400				
H2b		Complete‡		⟨Coding⟩	0·79	54	360				
H1b		Complete‡		⟨Coding to 3′ non-coding⟩	0·65	60	420				
H3		Complete‡		⟨5′ Non-coding to 3′ non-coding⟩	0·77	56	440				
H4		Complete‡ (mRNA starts putative)		⟨5′ Non-coding to 3′ non-coding. .	0·97	55	330				
Histone cluster	*Xenopus*		8599					0·57	45	5219	Perry *et al.* (1985)
H1a		Complete‡		⟨Coding (H1a) to 3′ non-coding (H2b)⟩	0·72	57	1450				
H2b		Complete‡									
H2a		Complete‡		⟨5′ Non-coding to coding⟩	0·98	57	700				
H3		Complete‡		⟨5′ Non-coding to 3′ non-coding⟩	0·75	58	770				
H4		Complete‡ (mRNA starts putative)		⟨5′ Non-coding to 3′ non-coding⟩	0·90	57	460				

IVS, intron; E, exon.

comp, complimentary strand.

← or →, the CpG island extends to the 5′ or 3′ end of the sequence, respectively; ⟨ or ⟩, there is more than 250 bp of CpG-depleted sequence between the 5′ (⟨) or 3′(⟩) boundary of the CpG island and the 5′ or 3′ end of the sequence, respectively; . . ., the CpG island extends to within 250 bp of the 5′ or 3′ boundary of the sequence.

† GenBank code.

‡ The location of the mRNA end is putative.

but in some cases the exact boundaries of the island were difficult to determine, and in a large number of cases one or both boundaries were outside the sequenced region of the gene. Nevertheless, it was immediately apparent that almost all the CpG islands in our survey are much greater than 200 bp in length (Table 2). Only three 5′ CpG islands, those associated with the genes coding for human brain β tubulin, human somatostatin-1 and porcine urokinase plasminogen activator are clearly less than 500 bp in length. The only two 3′ CpG islands that are clearly less than 500 bp in length occur in the genes for human MHC class II (HLA-DC-3β) and human brain β tubulin. Both are located within short interspersed repeated sequences of the *Alu* family. Five *Xenopus* histone genes have CpG islands less than 500 bp in length. Including all members of sequence families, at least three-quarters of the CpG islands in our survey are over 500 bp in length and at least one third are over 1 kb in length. The CpG islands associated with the human β tubulin gene and the oncogenes chicken, human and mouse c-myc, human c-Ha-ras1, and human and mouse int-1 are over 2 kb in length, but are broken by significant regions of lower G+C content and lower Obs/Exp CpG (e.g. Fig. 3(c)).

We classified DNA with a moving average of %G+C over 50 and Obs/Exp CpG over 0·6 as CpG-rich DNA. A small number of genes, including *Xenopus* U2 snRNA, chicken ρ globin and *Xenopus* heat shock protein hsp70, have 5′ CpG islands that might be classed as "borderline" based on these levels of %G+C and Obs/Exp CpG (e.g. Fig. 3(g)). The small 3′ CpG islands associated with *Alu* repeats in human MHC class II (HLA-DC-3β) and human brain β tubulin, mentioned previously, could also be classed as borderline in this regard. However, most of the CpG islands in the survey include stretches of highly CpG-rich DNA, with a moving average of %G+C over 60 and Obs/Exp CpG over 0·8, and are therefore much longer and stronger than the criteria used to identify them.

(ii) *Most 5′ CpG islands extend well into the gene*

By locating approximate boundaries of CpG islands, we have been able to study the position of CpG islands relative to the transcription start site, the translation start site and the position of exons and introns. We have found that 5′ CpG islands are strikingly non-uniform in this regard.

Table 3 lists, for all the genes in our survey where the transcription start site is known, the length of sequence upstream from the transcription start site and the extent of the CpG island upstream and downstream from the transcription start site. Most 5′ CpG islands begin before the transcription start site, but the length upstream varies enormously from less than 100 bp in mouse ribosomal protein L32, and human and rat somatostatin, to more than 2 kb in human c-myc. We found no case where the CpG island lies entirely within the 5′ untranscribed flank of the gene. A number of 5′ CpG islands extend far into the gene. For instance, the 5′ CpG islands associated with goat and human α globin, human and mouse int-1, and bovine and rat oxytocin-neurophysin all extend into the 3′ untranslated portion of the gene (Fig. 3(a), (c) and (e)). In fact, within the limited set of genes where sufficient sequence is available to allow a comparison to be made, most 5′ CpG islands extend further downstream into the transcribed portion of the gene than upstream into the 5′ untranscribed flanking DNA.

Almost all the 5′ CpG islands associated with protein-coding genes extend past the translation start site (no data given). This occurs even when the translation start site is a considerable distance from the transcription start site, and is separated from it by an intron, as in the case of the genes for chicken cardiac α actin, chicken β actin, human enkephalin, chicken glyceraldehyde-3-phosphate dehydrogenase (GAPDH) and chicken, mouse and human c-myc (e.g. Fig. 2(a)). The only genes for which the complete upstream sequence is available, and the island ends before the translation start site, are chicken cytochrome *c*, mouse ribosomal protein L32 and human proopiomelanocortin.

Tykocinski & Max (1984) noted that the CpG-rich regions in various MHC class I genes occurred in both exons and introns near the 5′ end of the gene. We have found that 5′ CpG islands can be located in exons, introns and 5′-flanking DNA, and are not consistently stronger (in terms of %G+C and Obs/Exp CpG) in any particular region. The only clear exceptions are the islands associated with the human and mouse int-1 genes (Fig. 3(c)). In both these cases, the island is stronger in the exons than the introns. In the case of mouse int-1, the association with exons is so pronounced and the CpG island so broken that it could almost be classed as three CpG islands, the first including some 5′ untranscribed flanking DNA and exon 1, the second encompassing exons 2 and 3, and the third encompassing exon 4.

(iii) *CpG islands associated with snRNA genes are strongest before the transcription start*

Not all genes containing CpG islands are translated. All snRNA genes in the survey have 5′ CpG islands starting upstream from the transcription start site. Cooper & Gerber-Huber (1985) suggested that the CpG-rich regions associated with snRNAs are due to RNA secondary structure requirements. Our study shows that this cannot be the entire reason for CpG islands in snRNAs, as the strongest part of the CpG island occurs upstream from the transcription start site in all cases (e.g. Figs 2(b) and 3(f)).

(iv) *3′ CpG islands mostly lie within exons, or closely surround and include exons*

We have identified a considerable number of 3′ CpG islands in the mammalian genes in our survey, despite the fact the criteria for selection of sequences did not require the presence of the translated portion or 3′ end of the gene. Therefore,

Table 3

Position of CpG island and G/C boxes relative to the transcription start site (▷) of the associated gene

		Length (bp)			Number of G/C boxes			
		Sequence	CpG island					
		before	before	after	> 250 bp before	≤ 250 bp before	≤ 250 bp after	> 250 bp after
Gene		▷	▷	▷	▷	▷	▷	▷
A. *Genes with 5′ CpG islands starting upstream from the transcription start site*								
α Actin, cardiac	Chicken	1090	380	690	0,0	0	2	1,0
α Actin, skeletal	Chicken	91	>91	409	—	†1	1	0
β Actin	Chicken	543	183	1207	1	0,6	3	6,0
β Actin	Rat	234	≥24	1126		†0,0	0	6,0
Adenosine deaminase	Human	131	>131	>142		†5	0*	2
5-Aminolevulinate synthase	Chicken	996	136	454	0	0,3	2	1,0
APRT	Mouse	840	130	430	1	0,3	0	0,1
Argininosuccinate synthetase	Human	760	≥600	>345	0,2	2	2	0
c-fos	Human	739	≥639	1261	0,0	1	1	4,0
c-fos	Mouse	133	>133	1387		†1	0	3,0
c-Ha-ras1	Human	536–560	>536	2900	2	8	2	6,1
c-myc	Chicken	1203	>1203	2247	7	4	1	8,0
c-myc	Human	2327	2077	2993	0,3	1	1	12,4
c-myc	Mouse	424	>424	>2882	0	3	0	7,0
c-sis	Human	580	>580	>612	3	2	2	0,1
Collagen, α2(I)	Chicken	403	>403	>277	1	0	0	0,1
Collagen, (II)	Rat	995	455	>409	1	4	0	0
DHFR	Human	1251	531	≥509	0,5	5	1	0,0
DHFR	Mouse	951	≥811	>284	0,7	4	0	0,0
EGF receptor	Human	208–362	>208	>352	—	2	3	0
Enkephalin	Human	949	>949	1071	1	0	1	3
Enkephalin	Rat	444	≥334	>282	0,0	0	0	0
ρ Globin	Chicken	200	≥120	460	—	†0,0	0	0,0
α^A Globin	Duck	330	≥200	120	0	0,5	0,0	0
α1 Globin	Goat	876	86	≥834	0	0,1	0	1,1
α2 Globin	Goat	704	≥564	≥836	0,1	2	1	1,0
α2 Globin	Human	3422	505	835	1,2	6	1	3,0
α1 Globin	Human	2978	546	734	1,2	6	1	2,2
GAPDH	Chicken	689	>689	911	2	6	4	4,0
Heat shock protein hsp70	Human	275	≥125	2085	0	1,0	0	2,0
Heat shock protein hsp70	*Xenopus*	361	≥261	239	0,0	0	0,0	1
HMG CoA reductase	Hamster	292–384	>292	306	1	4	2	1,0
HPRT	Mouse	845	125	>195	1	0,3	0†	0
int-1	Human	280	>280	3400	0	3	2	5,1
int-1	Mouse	497	≥447	2923	0,0	2	0	2,0
Metallothionein-1A	Human	860	≥660	110	0,0	5	0,0	1
Metallothionein-II	Human	765	395	105	0,0	2	0,0	0
Metallothionein-I	Mouse	300	>300	320	0	2	0	2
Metallothionein-II	Mouse	374	>374	526	0	1	0	1
Oxytocin-neurophysin	Bovine	209	>209	≥841	—	†1	3	5,0
Oxytocin-neurophysin	Rat	219	≥20	≥790	—	†0,0	1	0,0
PGK	Human	435–445	>435	≥235	0	3	0,0	0
Ribosomal protein L30	Mouse	460	>460	310	0	0	0	1,0
Ribosomal protein L32	Mouse	367	87	683	1	0,0	2	1,0
Ribosomal protein S16	Mouse	347	>347	713	0	2	1	0,0
Somatostatin-I	Human	1125	65	275	0	0,0	0	0,0
Somatostatin-14	Rat	748	98	452	0	0,0	1	0,0
Superoxide dismutase-1	Human	292	>292	>246	0	3	0†	0
Triosephosphate isomerase	Human	333	≥303	>149	0,0	5	1†	0
β Tubulin	Human	106	>106	2374	—	†2	0	9,1
Urokinase	Human	799	189	491	0	0,3	1	0,0
Vimentin	Hamster	140	≥80	>760	—	†0	0	0,2
B. *Genes with 5′ CpG islands—small nuclear RNA genes*								
snRNA, U1-52a	Chicken	441	>441	>249	3	4	0†	—
snRNA, U1-52b (comp)	Chicken	373	>373	>280	1	3	0	0
snRNA, U1-52c (comp)	Chicken	328	>328	>273	0	4	0	0
snRNA, U1	Human	432	≥362	>374	0,1	0	0	0

Table 3 *(continued)*

		Length (bp)			Number of G/C boxes			
		Sequence before	CpG island before	CpG island after	> 250 bp before	≤ 250 bp before	≤ 250 bp after	> 250 bp after
Gene		▷	▷	▷	▷	▷	▷	▷
snRNA, U1 clone 6–6B	Rat	689	299	161	0,**0**	**0**	**0**,0	0
snRNA, U2	Rat	419	>419	201	**0**	**2**	**0**,0	0
snRNA, U2	*Xenopus*	359	>359	11	**0**	**0**	**0**,0	0
C. *Genes with both 5′ CpG islands and 3′ CpG islands*								
Proopiomelanocortin	Bovine	203	>203	>151		**†1**	**1†**	0,**4**,0
Proopiomelanocortin	Human	680	430	1180	0,1	**0**	**1**	**4**,0,**3**,0
Proopiomelanocortin	Mouse	308	≥208	≥22	0	0,1	**0†**	0,1,0
Proopiomelanocortin	Rat	711	241	>161	0	0,1	**0†**	0,1,0
D. *Genes with 5′ CpG islands separated by unsequenced DNA from CpG-rich regions further downstream*								
Retinol binding protein	Human	784	434	≥606	0,**0**	**1**	**1**	**1**,0,**0**,0
E. *Genes with 5′ CpG islands starting between the transcription and translation start sites*								
CRF	Human	333	N/A	N/A	0	0	0,**0**	**3**,0
Urokinase	Porcine	974	N/A	N/A	1	3	0,**0**	**0**,0
F. *Genes with 3′ CpG islands only*								
α Actin, skeletal	Mouse	753	N/A	N/A	0	2	0	0,**2**,0
α Actin, skeletal	Rat	190	N/A	N/A		†2	0	0,**2**
Apolipoprotein E	Human	1046	N/A	N/A	2	2	0	3,**4**,0
α^A Globin	Chicken	300	N/A	N/A	0	2	0	0,**1**,0
ζ Globin	Human	769	N/A	N/A	0	0	1	0,**25**,0
MHC, class II HLA-DC-3β	Human	645	N/A	N/A	0	0	0	0,**4**,4,1,0
Vasopressin-neurophysin II	Bovine	315	N/A	N/A	0	0	2	1,**12**
Vasopressin-neurophysin II	Rat	367	N/A	N/A	0	0	0	2,**0**,0
G. *Histone genes*								
Histone H1	Chicken	169	>169	441	—	**†1**	**1**	**0**,1
Histone H4	Chicken	246	≥96	≥374	—	**†1,2**	**1**	**0**,1
Histone H5	Chicken	778	218	682	1	0,**2**	**0**	**0**,1
Histone H5	Duck	151	≥101	719	—	†0,**2**	**0**	**0**,0
Histone H2a	Human	206	N/A	N/A	—	†0	0,**0**	**0**,1
Histone H2b	Human	323	23	≥427	0	0,**0**	**1**	**0**,0
Histone H4	Mouse	228	≥78	≥492	—	†0,**1**	**0**	**0**,0

This Table examines genes for which the transcription start site has been definitely located.

Where a range is given for the length of sequence upstream from the transcription start site, multiple transcription start sites of similar or unknown priority exist, and the number of G/C boxes is recorded relative to the most upstream transcription start site.

> Indicates that the CpG island extends to the 5′ or 3′ boundary of the sequence; ≥ indicates that the CpG island extends to within 250 bp of the 5′ or 3′ boundary of the sequence.

Underlining of the length of CpG island, upstream and downstream from the transcription start site, indicates that a comparison between the upstream and downstream lengths of CpG island is possible.

For each gene region studied, the number of G/C boxes outside the CpG island (normal typeface) and the number of G/C boxes within the CpG island (bold typeface) are listed. Numbers of G/C boxes, within particular gene regions, are listed in 5′ to 3′ order of occurrence.

N/A The length of CpG island upstream or downstream from the transcription start site is not applicable, as the CpG island lies entirely downstream from the transcription start site.

comp, complimentary strand.

† The complete sequence within 250 bp of the transcription start site is not available.

we believe that 3′ CpG islands are not uncommon, in mammalian genomes at least. We also found some unusual genes, bovine, human, mouse and rat proopiomelanocortin and human brain β tubulin, which have both 5′ CpG islands and 3′ CpG islands, separated by about 5 kb of CpG-depleted DNA (e.g. Fig. 2(e)).

Most of the 3′ CpG islands in our survey extend to the 3′ end of the gene, but a few lie between the initiation and termination codons (Table 2F, G and H). The 3′ CpG islands, unlike the 5′ CpG islands, mostly lie within exons, or consist of a stretch of DNA which encompasses an exon or exons (e.g. Figs 2(c), (e) and 3(b)). The only two examples of CpG islands totally within an intron are the small, weak islands associated with *Alu* family repeats in human MHC class II (HLA-DC-3β) and human brain β tubulin.

(v) *Some histone sequences contain unusual regions with the expected frequency of CpG dinucleotides despite low G+C content*

In our study, we found that regions of high Obs/Exp CpG (over 0·6) are almost always located in regions of high G+C content (over 50), that is, in CpG islands. The only regions over 200 bp in length, with a moving average of Obs/Exp CpG over 0·6 and a moving average of %G+C under 50, occur in the genes for rat γ crystallin, *Xenopus* β-1 globin, mouse ribosomal protein L30, human c-myc and a number of histones. Of these, the only large regions of low G+C content and high Obs/Exp CpG, comparable in length to the majority of CpG islands in our survey, occur in *Xenopus* histone gene clusters (e.g. Fig. 2(d)); regions over 500 bp in length occur upstream from *Xenopus* histone H4, H3 and H2a genes.

Whether or not these few regions of low G+C content and high Obs/Exp CpG are part of the same phenomenon as CpG islands is unclear, though it appears they may be in the case of histone genes. Most histone genes, for which the transcription start site has been determined, have 5′ CpG islands beginning in the 5′ untranscribed flanking DNA (Table 2H), and the histone gene regions with low G+C content and high Obs/Exp CpG also lie upstream from the transcription start site. For instance, each of the large regions of low G+C content and high Obs/Exp CpG found in the *Xenopus* histone gene clusters closely precedes a 5′ CpG island. Human histone H2b, chicken histones H2a and H2b, and mouse histone H4 have small regions of low G+C content and high Obs/Exp CpG, located either immediately upstream from a 5′ CpG or within a broken 5′ CpG island. A number of histone genes have 3′ CpG islands, as we have defined them, and all of these genes, with the exception of human histone H2a, have a region of low G+C content and high Obs/Exp preceding the translation start site. So all the histone genes in our survey have a CpG island within the protein-coding portion of the gene, and all the histone genes, with the exception of human histone H2a, have a region of high Obs/Exp CpG before the translation start site. This region of high Obs/Exp CpG is usually contained within a CpG island, but may consist, at least in part, of a region with a low G+C content.

Since very few *Xenopus* sequences are available, we cannot be certain whether these unusual regions are characteristic of *Xenopus* genes in general, characteristic of vertebrate histone genes in general, or are found both in *Xenopus* genes and histone genes.

(vi) *CpG islands are found in genes with a range of tissue specificities*

CpG islands are not limited to any one class of gene. Nevertheless, we have found a number of relationships between the occurrence or location of CpG islands and the extent of tissue-specific expression of the associated genes. Firstly, all housekeeping genes in the survey, including ubiquitously expressed genes for metabolic enzymes, structural proteins and snRNAs, have 5′ CpG islands that begin upstream from the transcription start site, where this has been established. The only exceptions are some histone genes, which may be ubiquitously expressed and which contain CpG islands beginning downstream from the transcription start site. However, as discussed earlier, these histone genes do contain, in their 5′-flanking DNA, unusual regions with the expected frequency of CpG dinucleotides despite a low G+C content.

Secondly, the genes with 5′ CpG islands include housekeeping genes, widely expressed genes, and highly tissue-specific genes expressed only in terminally differentiated cells. In contrast, the genes with 3′ CpG islands and the CpG-depleted genes all encode products that are tissue-specific. We are unable to identify any functional characteristics that differentiate, as a class, tissue-specific genes with 5′ CpG islands from genes with 3′ CpG islands or CpG-depleted genes. In addition, there is no obvious difference, in terms of the length or strength of CpG islands, between housekeeping and tissue-specific genes.

(vii) *Some housekeeping genes and other widely expressed genes with 5′ CpG islands lack a TATA box, but no tissue-specific genes with 5′ CpG islands lack a TATA box*

A number of workers have noted that some housekeeping genes have G+C-rich promoters that lack the TATA box normally found in genes transcribed by RNA polymerase II (for a review, see Dynan, 1986). Each of these G+C-rich promoters forms part of a 5′ CpG island. We have analysed all genes in our survey, where the transcription start site is known, for the presence or absence of a TATA box in the appropriate region. A number of genes with 5′ CpG islands lack a TATA box or similar A+T-rich sequence (Table 4). In agreement with the observations by Dynan (1986), this group of genes is comprised mainly of housekeeping genes, but includes a few widely expressed genes. We have shown that a number of tissue-specific genes with 5′ CpG islands have promoter regions as G+C-rich as those of house-

Table 4
Genes with no TATA box

Housekeeping genes	Other widely expressed genes
Adenosine deaminase (human)	c-Ha-ras1 (human)
APRT (mouse)	c-sis (human)
DHFR (mouse)	EGF receptor (human)
HMG CoA reductase (hamster)	
HPRT (mouse)	
PGK (human)	
Ribosomal protein L32 (mouse)	
Ribosomal protein L30 (mouse)	
snRNA U1 (chicken, human, rat)	
snRNA U2 (rat, *Xenopus*)	
β Tubulin (human)	

keeping genes. Interestingly, none of these tissue-specific genes with 5′ CpG islands lacks a TATA box or similar sequence. So the absence of a TATA box does appear to be related to some aspect of the regulation of widely expressed genes, and is not simply related to the presence of a CpG island in the promoter region of these genes.

(viii) *CpG islands are not always conserved between species*

Sequence characteristics of functional significance tend to be conserved for equivalent genes from different species, and the presence of a CpG island in a certain part of a gene is often conserved in this way. However, in a number of cases we found clear differences between species in the position, length or strength of a CpG island (Fig. 3). A number of examples are described below.

(1) The chicken skeletal α actin gene has a strong 5′ CpG island whereas both mouse and rat skeletal α actins have a weak and broken 3′ CpG island near the end of the gene (Fig. 3(g)).

(2) The bovine and rat oxytocin-neurophysin genes both have 5′ CpG islands, but the rat island extends only a few base-pairs upstream from the transcription start site, whereas the bovine island includes more than 200 bp of untranscribed 5′-flanking DNA (Fig. 3(a)). Similarly, the human and porcine urokinase plasminogen activator genes both have 5′ CpG islands but only the human island extends upstream from the transcription start site (Fig. 3(d)). So only the presence of a CpG island within the transcribed portion of the gene is conserved in oxytocin-neurophysin and urokinase plasminogen activator.

(3) As mentioned previously, the human and mouse int-1 genes both have long, broken 5′ CpG islands, stronger in exons than in introns. In mouse int-1, this effect is far more pronounced because almost the entire length of the introns between the CpG-rich exons is CpG-depleted (Fig. 3(c)). Similarly, the bovine and rat vasopressin-neurophysin genes have both 3′ CpG islands that appear to be exon-related, but in rat vasopressin-neurophysin the intron between the CpG-rich exons is markedly CpG-depleted so the effect is more pronounced (Fig. 3(b)). Thus, only the presence of the island within the exon is conserved in int-1 and vasopressin-neurophysin.

(4) CpG islands associated with the α globin genes are extremely variable (Fig. 3(e)). The human and goat $\alpha 1$ and $\alpha 2$ globin genes have strong 5′ CpG islands, whereas some α globin genes from mouse and chicken are entirely CpG-depleted.

In general, we observe that, where differences in the length or strength of CpG islands occur, the bovine, chicken or human CpG islands are more pronounced than those of mouse, rat or *Xenopus*.

(c) *CpG islands and G/C boxes*

Bird (1986) has suggested that CpG islands may bind ubiquitious transcription factors. The transcription factor Spl is a possible candidate. Spl facilitates the transcription *in vitro* of several viral and vertebrate genes, including the simian virus 40 (SV40) early genes (Dynan & Tjian, 1983; Gidoni *et al.*, 1984), monkey sequence 7·02 (Gidoni *et al.*, 1984; Dynan *et al.*, 1985), herpes simplex virus (HSV) thymidine kinase (Jones *et al.*, 1985), HSV immediately-early genes *3* and *4/5* (Jones & Tjian, 1985), mouse DHFR (Dynan *et al.*, 1986), and AIDS retrovirus LTR (Jones *et al.*, 1986). In these cases, Spl binds to a segment of DNA, containing G/C boxes, within the promoter of the gene.

A G/C box consists of the hexanucleotide sequence GGGCGG or its reverse complement CCGCCC. Kadonaga *et al.* (1986) have derived a decanucleotide consensus sequence for Spl binding, G/TGGGCGGPuPuPy, that contains a G/C box. With the exception of two of the three binding sites in AIDS retrovirus LTR, Spl actually binds to the G/C box. In AIDS retrovirus, the two unusual binding sites do not contain a perfect G/C box but do contain a strong homology (8 out of 10 nucleotides) to the decanucleotide consensus sequence for Spl binding (Jones *et al.*, 1986).

CpG islands, because of their unusual sequence composition (high G+C content, high CpG content), might be expected to contain G/C boxes at a higher frequency than occurs in bulk DNA. So we investigated the number and location of G/C boxes relative to the transcription start site, and the association of G/C boxes with CpG islands, in the genes in our survey (Table 3).

(i) *G/C boxes are rare in CpG-depleted genes*

We found that G/C boxes are rare in all regions of the CpG-depleted genes in the survey. The only CpG-depleted genes that have a G/C box in what we will loosely call the "promoter region", that is, in the 250 bp immediately upstream from the transcription start site, are the genes for rat α and γ fibrinogen, chicken α^D and β globin, rat hepatic product spot 14, human insulin, and bovine and rat parathyroid hormone. In each case only a single G/C box is present in the promoter region.

(ii) *G/C boxes are commonly found upstream from the transcription start site of genes with 5′ CpG islands*

Almost all G/C boxes found upstream from the transcription start site of the genes in our survey occur in CpG islands (Table 3). The majority of the genes with 5′ CpG islands beginning upstream from the transcription start site have at least one G/C box within the promoter region (Table 3A, B, C, D and G). Few of the genes with CpG islands starting downstream from the transcription start site have G/C boxes in the promoter region (Table 3E, F and G). In most cases, the presence or absence of G/C boxes in the promoter region appears to be conserved between species. Unexpectedly, this also appears to be true for the three cases, urokinase plasminogen activator, α globin and skeletal α actin

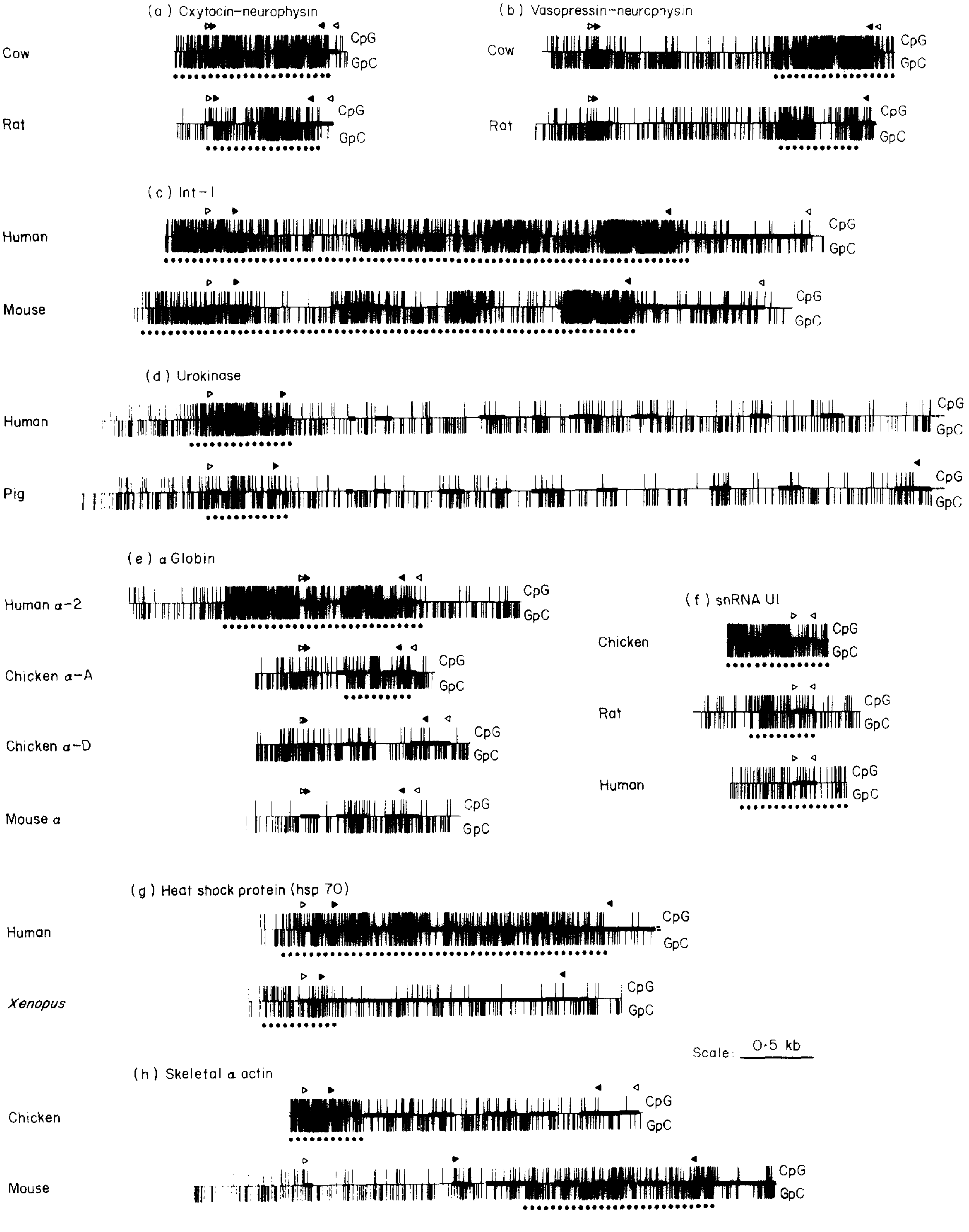

Figure 3. Distribution of CpG dinucleotides along the sequences of equivalent genes in a number of vertebrate species. The location of individual CpG and GpC dinucleotides along each sequence and the extent of each CpG island are marked as described in the legends to Figs 1 and 2; a single broken line at the end of a diagram indicates that the full sequence analysed has not been included in the diagram.

genes, where the presence of a CpG island upstream from the transcription start site is not conserved.

(iii) *G/C boxes are also found downstream from the transcription start site of genes with CpG islands*

We also found many G/C boxes downstream from the transcription start site of genes with 5′ CpG islands, again almost always within the CpG island (Table 3A, B, C and D). Clustered G/C boxes, within 250 bp downstream from the transcription start site, occur in the genes coding for chicken β actin, chicken GAPDH, hamster HMG CoA reductase and bovine oxytocin-neurophysin. Clusters further than 250 bp downstream from the transcription start site occur in the genes coding for chicken and rat β actin (intron 1), human β tubulin (intron 1), human c-Ha-ras1 (intron 3, exon 4), chicken c-myc (exon 2), human c-myc (exon 2, intron 2), chicken GAPDH (intron 2), and bovine oxytocin-neurophysin (intron 1). Most of the 3′ CpG islands contain G/C boxes (Table 3C, F and G). Clusters occur in human MHC class II (HLA-DC-β; introns 1 and 2) and bovine vasopressin-neurophysin (introns 1 and 2), and a large, concentrated cluster occurs within the repeated sequence that forms human ζ globin intron 1.

The random nature of the DNA sequence in CpG islands cannot be assumed. For example, Tautz *et al.* (1986) have demonstrated a non-random distribution of trinucleotide and tetranucleotide motifs in the CpG islands associated with chicken β actin. Therefore, to ascertain whether G/C boxes are present more often than expected, in any particular region, from the nucleotide composition alone would require a mathematical model beyond the scope of this study. Nevertheless, in the case of genes with 5′ CpG islands beginning upstream from the transcription start site, we observe that clusters of G/C boxes are more common in the promoter region than downstream from the transcription start site (Table 3A).

(iv) *G/C boxes are found in the promoter region of both housekeeping and tissue-specific genes*

The promoter regions of all vertebrate housekeeping genes, examined previously, contain G/C boxes (Maguire *et al.*, 1986), and work on Sp1 binding to G/C boxes has been confined to viral genes and vertebrate housekeeping genes. Therefore, we investigated whether G/C boxes are associated with CpG islands generally, or are more clearly associated with a particular class of genes such as housekeeping genes.

Most housekeeping genes in our survey have G/C boxes in the promoter region, but there are some notable exceptions. Mouse ribosomal protein L30 and rat U1 snRNA have no G/C boxes in over 450 bp of sequence upstream from the transcription start site. However, most tissue-specific genes with 5′ CpG islands starting upstream from the transcription start site, also have G/C boxes in the promoter region, with prominent clusters occurring in the genes coding for rat collagen type II, and human and duck α globins.

We note that the presence of a G/C box does not necessarily imply that Sp1 will bind to the gene (Kadonaga *et al.*, 1986). Furthermore, the absence of G/C boxes in a region does not imply that Sp1 will not bind to the region, especially if a sequence homologous to the decanucleotide Sp1 binding consensus sequence, such as that found in AIDS retrovirus LTR, is present. No studies have been carried out to determine whether Sp1 will bind to the two unusual sites, in the absence of the central G/C box, in the AIDS retrovirus, but there is no evidence to the contrary.

Our study indicates that future studies of Sp1 binding and regulation should not be confined to the promoter regions of genes, nor to housekeeping genes, but should focus on all genes with CpG islands. In particular, the G/C box clusters downstream from the transcription start site and the G/C box clusters in the tissue-specific genes listed above, warrant investigation.

(d) *Maintenance of CpG islands*

(i) *DNA regions with a high G+C content are not necessarily CpG-rich*

One theory for the maintenance of CpG islands in methylated genomes is that the lack of CpG depletion is a direct result of the high G+C content of the DNA. This theory is based on the suggestion by Adams & Eason (1984) that mutation by deamination of ^{5m}CpG to TpG occurs less frequently in G+C-rich regions, because of the inherent stability of these regions. Adams & Eason (1984) found that vertebrate sequences with a low G+C content (40% G+C or less) were CpG-depleted, whereas sequences with a very high G+C content (around 70%) had close to the expected frequency of CpG dinucleotides. However, if we take into account the presence of CpG islands in a wide range of vertebrate genes, we can propose another theory to account for the apparent relationship between G+C content and Obs/Exp CpG: that the relationship is caused by varying lengths of CpG island DNA in the sequences analysed, with an absence of CpG islands in regions of low G+C content (40% or less), and a preponderance of CpG islands in regions of over 70% G+C. This theory does not require that the two characteristics of CpG islands, high G+C content and high Obs/Exp CpG, be causally related. It predicts that, if CpG island sequences and depleted sequences were separated, no relationship between G+C content and Obs/Exp CpG would be found. On the other hand, if maintenance of a high frequency of CpG dinucleotides is dependent largely or entirely on G+C content, an increase in Obs/Exp CpG with increasing G+C content should be observed both in the set of CpG island sequences and the set of CpG-depleted sequences.

To test these predictions, we first determined the

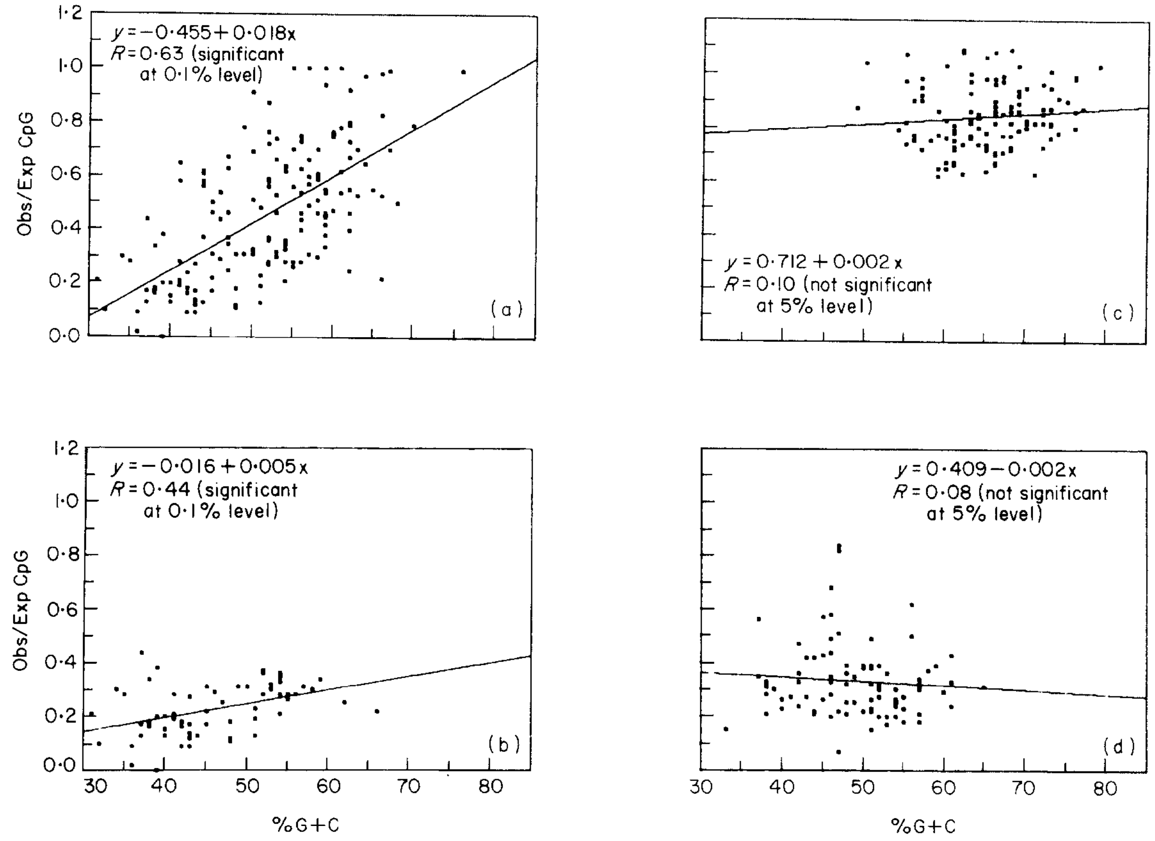

Figure 4. Relationship between Obs/Exp CpG and %G+C. The value for Obs/Exp CpG was plotted against %G+C for: (a) total sequences of all genes in the survey; (b) sequences of CpG-depleted genes, as listed in Table 1; (c) CpG island sequences, as listed in Table 2; (d) remaining CpG-depleted sequences, from genes with CpG islands, as listed in Table 2. The regression line, linear regression equation, correlation coefficient (*R*), and significance level of the correlation for each set of points are shown.

total %G+C and Obs/Exp CpG for each sequence listed in Tables 1 and 2. We found that there is a strong positive correlation between the total values for %G+C and Obs/Exp CpG (Fig. 4(a)), in agreement with the observations by Adams & Eason (1984). We then separated each sequence with a CpG island into CpG island region(s) and CpG-depleted region(s), and calculated the %G+C and Obs/Exp CpG for the CpG island sequence and the remaining CpG-depleted sequence. We found no correlation either between %G+C and Obs/Exp CpG values for the CpG island parts of the sequence (Fig. 4(c)) or between %G+C and Obs/Exp CpG values for the CpG-depleted parts of the sequences (Fig. 4(d)). We did find a positive correlation between %G+C and Obs/Exp CpG values for the completely CpG-depleted genes in our survey (Fig. 4(b)). However, the rate of increase of Obs/Exp CpG with increasing G+C content was very much less in the CpG-depleted genes (Fig. 4(b)) than in the total set of genes (Fig. 4(a)). The reason for this apparent difference in the relationship between %G+C and Obs/Exp CpG in the two sets of CpG-depleted sequences (Fig. 4(b) and (d)) is unclear. We note that the set of sequences from entirely CpG-depleted genes contains more 5′ regic ıs of genes than does the set of CpG-depleted sequences from genes with CpG islands. Therefore, the positive correlation between %G+C and Obs/Exp CpG in the CpG-depleted genes might be due to some slight "CpG island character" in the 5′ regions of some of these genes. Alternatively, the effect could indicate a small influence of G+C content on Obs/Exp CpG in CpG-depleted genes.

These results suggest that there is no relationship between %G+C and Obs/Exp CpG within separated CpG island and CpG-depleted sequences that could account for the strong relationship between %G+C and Obs/Exp CpG in the total set of sequences, so we believe that this strong relationship is due to varying lengths of CpG island DNA in the sequences. The results are consistent with the theory that G+C content and Obs/Exp CpG are independently maintained in CpG islands, and are not consistent with the theory that G+C content alone determines the extent of CpG depletion.

However, our study indicates that high G+C content may still be involved in the maintenance of regions of DNA with a high frequency of CpG dinucleotides. Figure 4 shows that, at G+C contents between 30% and 50%, the vast majority of the sequences are CpG-depleted, and our extensive analysis of individual sequences showed that regions greater than 200 bp in length with a moving average of %G+C under 50 and Obs/Exp CpG over 0·6 are very rare (see section (b) (v), above). This suggests that a high G+C content may be generally necessary to prevent CpG depletion. At G+C contents between 50% and 65%, DNA sequences can have high or low values for Obs/Exp CpG. For instance, many completely CpG-depleted genes are G+C-rich (Fig. 4(b)), and CpG-depleted regions with a high G+C content occur in many genes with CpG islands (Fig. 4(d); e.g. apolipoprotein A1, Fig. 2(c)). Again this indicates that a high G+C content is not sufficient to prevent CpG depletion. At a G+C content over 65%, essentially all the sequences have high Obs/Exp CpG values. Human insulin is the only CpG-depleted sequence, in our study, with a G+C content over 65%, and is the only sequence that contains regions greater than 200 bp in length with a moving average of Obs/Exp CpG under 0·6 and a moving average of %G+C over 70 (Fig. 1(b)). So we cannot exclude the possibility that at this very high G+C content, the stability of the DNA duplex could act alone to prevent deamination of ^{5m}CpG.

To summarize, our results are not consistent with one of the models we tested: that regions with a high frequency of CpG dinucleotides are maintained entirely by resistance to deamination of ^{5m}CpG in regions of high G+C content. Our results are consistent with an alternative model: that a high G+C content is generally required to prevent CpG depletion, but is not sufficient to prevent CpG depletion, except perhaps at the highest levels of G+C. Our results are also consistent with the other model we tested: that regions of DNA with the expected frequency of CpG dinucleotides almost always occur in CpG islands for reasons other than the high G+C content of CpG islands, that is, the characteristics of CpG islands, high Obs/Exp CpG and high G+C content are not causally related.

(ii) *CpG islands concentrated in exons are not the result of arginine codon usage*

The apparent relationship between some CpG islands and exons raises the possibility that these islands may be due, at least in part, to coding effects. Neutral mutations of CpG dinucleotides can occur when the C is in position 2 or 3 of a codon. In methylated regions of DNA, such neutral mutations could result from deamination of ^{5m}C. However, no neutral mutations of CpG dinucleotides can occur when the C is in position 1 of the codon. Arginine is the only amino acid that has a codon starting with the dinucleotide CpG. So, CpG islands associated with exons coding for arginine-rich peptides could result from a strong requirement for the arginine codon CGX.

To test the effect of arginine codon usage, we first calculated the percentage of arginine codons in each exon, or part thereof, that lay within exon-associated CpG islands. Many of these regions did code for arginine-rich peptides (Table 5, column (a)), with the great majority of arginines coded by CGX, rather than AGA or AGG (Table 5, column (b)). However, when we calculated the %G+C and Obs/Exp CpG for these exon regions, omitting CpG dinucleotides at the beginning of arginine codons, we found that, with very few exceptions (human apolipoprotein E, exon 4; human and mouse int-1, exon 2), the exons still had a G+C content greater than 50% and Obs/Exp CpG greater than 0·6 (Table 5, columns (e) and (f). So arginine codon usage is not a cause of these exon-associated CpG islands, although it could contribute to their strength. Therefore, the high preference for the arginine codon CGX is presumably a result of CpG islands in the coding regions of some genes. Furthermore, it seems possible that in a basic peptide coded by a CpG island, a high arginine content may be the result of the presence of the island, since the other basic amino acids do not permit a high content of CpG dinucleotides.

4. Conclusion

CpG islands are clearly ubiquitous in vertebrate genomes. Although many tissue-specific genes do not have CpG islands, it is becoming apparent that all widely expressed genes and many tissue-specific genes have CpG islands either at their 5′ ends or 3′ ends, or both. Thus, it seems possible that CpG islands will prove to be associated with the great majority of genes. However, we are unable to estimate the proportion of genes with or without CpG islands from the biased sample of genes currently available for study. Bird *et al.* (1985) have estimated from studies of the unmethylated component of mouse DNA that there may be as many as 30,000 CpG islands in the mouse genome.

The majority of CpG islands in our survey are associated with the 5′ ends of genes (5′ CpG islands). However, we identified a number of CpG islands which lie entirely downstream from the translation start site of the associated genes (3′ CpG islands). We found no consistent differences between 5′ and 3′ CpG islands in terms of %G+C, Obs/Exp CpG or the occurrence of G/C boxes. However, in contrast to the 5′ CpG islands, the 3′ CpG islands tend to be more pronounced in exons. We do not know if 5′ and 3′ CpG islands have the same function, or if they are maintained by the same mechanism. We consider it likely that the position of the CpG island is critical in housekeeping genes. All housekeeping genes in our survey have 5′ islands beginning upstream from the transcription start site, whereas the various tissue-specific genes have 5′ or 3′ CpG islands or are CpG-depleted across their length. We cannot exclude the possibility that some 3′ CpG islands are 5′ CpG islands relative to as yet unidentified transcripts which overlap the known transcription unit. Considering the recent discoveries of overlapping transcripts arising from opposite strands of DNA in some eukaryotic genes (Henikoff *et al.*, 1986; Spencer *et al.*, 1986; Williams & Fried, 1986), and of divergent transcription from some viral and vertebrate promoter regions within

Table 5

Effect of Arg codon usage on CpG islands associated with exons

						Total sequence analysed			Sequence omitting CpG dinucleotides in Arg codons	
Gene		Exon	Codons	N (bp)	(a) % Arg codons†	(b) Arg codon usage CGX/total‡	(c) Obs/Exp CpG	(d) % G+C	(e) Obs/Exp CpG	(f) % G+C
Apolipoprotein A-1	Human	E4	139–267	387	8·5	10/11	0·95	68	0·80	66
Apolipoprotein E	Human	E4	80–317	714	21·7	30/30	0·80	72	0·52	70
ζ Globin	Human	E2	33–100	204	2·9	2/2	1·09	68	1·04	67
		E3	101–142	126	4·8	2/2	0·97	68	0·88	66
Histone H2a	Human	N/A	3–115	339	11·5	13/13	1·08	62	0·81	59
Histone H1a	*Xenopus*	N/A	4–210	621	1·4	3/3	0·69	61	0·65	60
Histone H1b	*Xenopus*	N/A	5–144	420	1·4	2/2	0·64	60	0·61	59
Histone H2b	*Xenopus*	N/A	6–124	347	6·7	6/8	0·78	54	0·64	53
int-1	Human	E2	36–119	252	10·7	9/9	0·78	61	0·46	58
		E3	120–208	266	11·2	10/10	1·01	66	0·78	63
		E4	209–370	486	12·3	19/20	1·15	70	0·96	67
int-1	Mouse	E2	36–119	252	10·7	9/9	0·78	62	0·49	59
		E3	120–208	266	11·2	10/10	0·94	65	0·68	62
		E4	209–370	396	11·1	18/18	1·10	68	1·09	65
MHC, class II HLA-DC-β	Human	E2	37–126	269	13·3	9/12	1·06	61	0·82	58
MHC, class II HLA-DC-3β	Human	E2	37–126	269	13·3	9/12	0·97	60	0·70	57
MHC, class II H2-IA-β haplotype b	Mouse	E2	32–122	273	14·3	10/13	1·08	64	0·85	61
Proopiomelanocortin	Bovine	E3	45–265	663	8·6	17/19	0·94	68	0·78	66
Proopiomelanocortin	Human	E3	45–267	669	8·5	17/19	0·84	68	0·68	67
Proopiomelanocortin	Mouse	E3	47–232	558	9·1	14/17	0·85	64	0·68	62
Proopiomelanocortin	Rat	E3	45–231	561	9·6	14/18	0·83	64	0·66	62
β Tubulin, brain	Human	E4	93–444	1055	4·8	17/17	0·78	64	0·67	63
Vasopressin-neurophysin	Bovine	E2	41–107	201	4·5	3/3	0·81	74	0·74	73
		E3	108–166	176	10·2	6/6	1·02	76	0·84	74
Vasopressin-neurophysin	Rat	E2	45–111	201	4·5	3/3	0·78	72	0·70	71
		E3	112–144	88	18·2	5/6	1·08	68	0·88	64

N/A Not applicable because the gene is intronless.

† Number of Arg codons, as a percentage of the total number of codons.

‡ Number of CGX codons/total number of Arg codons.

CpG islands (Harvey *et al.*, 1982; Saffer & Singer, 1984; Vigneron *et al.*, 1984; Crouse *et al.*, 1985; Farnham *et al.*, 1985; Perry *et al.*, 1985; Mitchell *et al.*, 1986), we suggest that any search for transcripts arising from 3′ CpG islands must take into account the possibility of transcription from both strands of DNA.

The mechanism for the maintenance of CpG islands within the CpG-depleted bulk DNA is unclear. We have shown that CpG islands are not maintained, despite a tendency for ^{5m}C to mutate to T, by the structural stability of a high G+C content alone, nor are they maintained by some selective importance of the arginine codon CGX.

Accurate methylation studies of CpG islands are difficult due to the high frequency of CpG dinucleotides. However, all CpG islands studied to date appear to be unmethylated, except for those in the inactive X chromosome in somatic cells. So we believe that, on current evidence, the best theory for the maintenance of CpG islands is that the CpG dinucleotides are unmethylated, in germline DNA at least. CpG islands are not intrinsically unmethylatable (Compere & Palmiter, 1981; Lieberman *et al.*, 1983; Wolf *et al.*, 1984*a*,*b*; Yen *et al.*, 1984; Lock *et al.*, 1986), so we propose that some mechanism must exist to keep the CpG dinucleotides unmethylated in the germline. CpG islands may bind transcription factors that obstruct the methylating enzyme (Bird, 1986). Binding of such transcription factors is consistent with the observation that all housekeeping genes in our survey have 5′ CpG islands, since one would expect housekeeping genes to be transcribed in germline cells. However, many tissue-specific genes also have CpG islands. On current evidence concerning their expression it is unlikely, albeit possible, that these tissue-specific genes are transcribed in germline cells. Perhaps a more probable mechanism, whereby the CpG islands associated with tissue-specific genes could be maintained in an unmethylated state by binding of transcription factors, is that these genes require for transcription both ubiquitous factors which bind to CpG islands and tissue-specific factors. Alternatively, CpG islands may remain unmethylated because they bind a specific "anti-methylase" protein that is not involved in transcription, and is produced in high concentration in germline cells.

As all housekeeping genes have CpG islands, it seems likely that CpG islands are essential for the regulation of expression of vertebrate housekeeping genes. However, some tissue-specific genes are associated with CpG islands in one vertebrate species and not in another. For instance, the human and goat α globin genes have CpG islands, whereas at least one mouse α globin gene does not. We consider it unlikely that α globin genes in different mammalian species are controlled by completely different mechanisms. In such cases, it seems that CpG islands are not essential for appropriate expression of the associated gene. We propose that these variable CpG islands may confer a higher level of stability or tight control to the associated gene, for instance by involving a wider range of transcription factors or by preventing methylation of control sequences. Either gain or loss of the sequence characteristics of a CpG island could occur by accumulation of point mutations and by DNA slippage, a process which Tautz *et al.* (1986) suggest may be responsible for the generation of variable regions of "cryptic simplicity" in coding and non-coding DNA.

All CpG islands in our survey either extend into or lie within the transcribed portion of the associated gene, and in some cases only the CpG island region within the transcribed portion or within the exons of a gene is conserved between species. These observations suggest two directions for future work on the function of CpG islands. Firstly, one must consider the possibility that any transcription factors which recognize CpG islands could bind upstream and/or downstream from the transcription start site of the associated gene, depending on the location of the CpG island. For instance, possible Spl binding sites within the transcribed DNA require investigation. Secondly, the mRNA that arises from transcribed regions of CpG islands could form unusual three-dimensional structures or contain protein binding sites involved in post-transcriptional regulation of the associated gene. Indeed, post-transcriptional regulation is a feature of a number of genes with 5′ CpG islands: c-fos (Mitchell *et al.*, 1985; Meijlink *et al.*, 1985; Treisman, 1985), c-myc (Blanchard *et al.*, 1985; Knight *et al.*, 1985; Dean *et al.*, 1986), collagen (Focht & Adams, 1984; Stepp *et al.*, 1986), DHFR (Leys *et al.*, 1984; Yoder & Berget, 1985), GAPDH (Piechaczyk *et al.*, 1984), histones (Heintz *et al.*, 1983; Sittman *et al.*, 1983; Schiimperli, 1986), thymidine kinase (Groudine & Casimir, 1984) and tubulin (Cleveland & Havercroft, 1983). A number of CpG islands, in particular 3′ CpG islands, are concentrated in exons. The exon association could relate to involvement of processed mRNA in post-transcriptional gene regulation.

We thank Carolyn Bucholtz and Alex Reisner for assistance with computer analyses, Simon Worthington for development of some specialized computer programs, and Richard Cowan for helpful discussions. Computer facilities and laboratory space were generously provided by Alex Reisner and the CSIRO Division of Molecular Biology, North Ryde, N.S.W. This work was supported by a project grant (860266) from the National Health and Medical Research Council of Australia.

References

Adams, R. L. P. & Eason, R. (1984). *Nucl. Acids Res.* **12**, 5869–5877.

Battistuzzi, G., D'Urso, M., Toniolo, D., Persico, G. M. & Luzzatto, L. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 1465–1469.

Bienz, M. (1984). *EMBO J.* **3**, 2477–2483.

Bird, A. P. (1980). *Nucl. Acids Res.* **8**, 1499–1504.

Bird, A. P. (1986). *Nature (London)*, **321**, 209–213.

Bird, A., Taggart, M., Frommer, M., Miller, O. J. & Macleod, D. (1985). *Cell*, **40**, 91–99.

Blanchard, J.-M., Piechaczyk, M., Dani, C., Chambard, J.-C., Franchi, A., Pouyssegur, J. & Jeanteur, P. (1985). *Nature (London)*, **317**, 443–445.

Brown, J. R., Daar, I. O., Krug, J. R. & Maquat, L. E. (1985). *Mol. Cell. Biol.* **5**, 1694–1706.

Bucholtz, C. A. & Reisner, A. H. (1986). *Nucl. Acids Res.* **14**, 265–272.

Cleveland, D. W. & Havercroft, J. C. (1983). *J. Cell Biol.* **97**, 919–924.

Compere, S. J. & Palmiter, R. D. (1981). *Cell*, **25**, 233–240.

Cooper, D. N. & Gerber-Huber, S. (1985). *Cell Different.* **17**, 199–205.

Coulondre, C., Miller, J. H., Farabaugh, P. J. & Gilbert, W. (1978). *Nature (London)*, **274**, 775–780.

Crouse, G. F., Leys, E. J., McEwan, R. N., Frayne, E. G. & Kellems, R. E. (1985). *Mol. Cell. Biol.* **5**, 1847–1858.

Das, H. K., McPherson, J., Bruns, G.A.P., Karathanasis, S. K. & Breslow, J. L. (1985). *J. Biol. Chem.* **260**, 6240–6247.

Dean, M., Levine, R. A. & Campisi, J. (1986). *Mol. Cell. Biol.* **6**, 518–524.

D'Onofrio, C., Colantuoni, V. & Cortese, R. (1985). *EMBO J.* **4**, 1981–1989.

Dull, T. J., Gray, A., Hayflick, J. S. & Ullrich, A. (1984). *Nature (London)*, **310**, 777–781.

Dush, M. K., Sikela, J. M., Khan, S. A., Tischfield, J. A. & Stambrook, P. J. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 2731–2735.

Dynan, W. S. (1986). *Trends Genet.* **2**, 196–197.

Dynan, W. S. & Tjian, R. (1983). *Cell*, **32**, 669–680.

Dynan, W. S., Saffer, J. D., Lee, W. S. & Tjian, R. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 4915–4919.

Dynan, W. S., Sazer, S., Tjian, R. & Schimke, R. T. (1986). *Nature (London)*, **319**, 246–248.

Eldridge, J., Zehner, Z. & Paterson, B. M. (1985). *Gene*, **36**, 55–63.

Farnham, P. J., Abrams, J. M. & Schimke, R. T. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 3978–3982.

Focht, R. J. & Adams, S. L. (1984). *Mol. Cell. Biol.* **4**, 1843–1852.

Gidoni, D., Dynan, W. S. & Tjian, R. (1984). *Nature (London)*, **312**, 409–413.

Groudine, M. & Casimir, C. (1984). *Nucl. Acids Res.* **12**, 1427–1446.

Harvey, R. P., Robins, A. J. & Wells, J. R. E. (1982). *Nucl. Acids Res.* **10**, 7851–7863.

Heintz, N., Sive, H. L. & Roeder, R. G. (1983). *Mol. Cell. Biol.* **3**, 539–550.

Henikoff, S., Keene, M. A., Fechtel, K. & Fristrom, J. W. (1986). *Cell*, **44**, 33–42.

Hu, M. C.-T., Sharp, S. B. & Davidson, N. (1986). *Mol. Cell. Biol.* **6**, 15–25.

Hunt, C. & Morimoto, R. I. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 6455–6459.

Ishii, S., Xu, Y.-H., Stratton, R. H., Roe, B. A., Merlino, G. T. & Pastan, I. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 4920–4924.

Jinno, Y., Matuo, S., Nomiyama, H., Shimada, K. & Matsuda, I. (1985). *J. Biochem.* **98**, 1395–1403.

Jones, K. A. & Tjian, R. (1985). *Nature (London)*, **317**, 179–182.

Jones, K. A., Yamamoto, K. R. & Tjian, R. (1985). *Cell*, **42**, 559–572.

Jones, K. A., Kadonaga, J. T., Luciw, P. A. & Tjian, R. (1986). *Science*, **232**, 755–759.

Josse, J., Kaiser, A. D. & Kornberg, A. (1961). *J. Biol. Chem.* **236**, 864–875.

Kadonaga, J. T., Jones, K. A. & Tjian, R. (1986). *Trends Biochem. Sci.* **11**, 20–23.

Knight, E., Jr, Anton, E. D., Fahey, D., Friedland, B. K. & Jonak, G. J. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 1151–1154.

Kohno, K., Sullivan, M. & Yamada, Y. (1985). *J. Biol. Chem.* **260**, 4441–4447.

Lennon, G. G. & Fraser, N. W. (1983). *J. Mol. Evol.* **19**, 286–288.

Levanon, D., Lieman-Hurwitz, J., Dafni, N., Wigderson, M., Sherman, L., Bernstein, Y., Laver-Rudich, Z., Danciger, E., Stein, O. & Groner, Y. (1985). *EMBO J.* **4**, 77–84.

Leys, E. J., Crouse, G. F. & Kellems, R. E. (1984). *J. Cell. Biol.* **99**, 180–187.

Lieberman, M. W., Beach, L. R. & Palmiter, R. D. (1983). *Cell*, **35**, 207–214.

Lock, L. F., Melton, D. W., Caskey, C. T. & Martin, G. R. (1986). *Mol. Cell. Biol.* **6**, 914–924.

Maguire, D. J., Day, A. R., Borthwick, I. A., Srivastava, G., Wigley, P. L., May, B. K. & Elliott, W. H. (1986). *Nucl. Acids Res.* **14**, 1379–1391.

McClelland, M. & Ivarie, R. (1982). *Nucl. Acids Res.* **10**, 7865–7877.

McGrogan, M., Simonsen, C. C., Smouse, D. T., Farnham, P. J. & Schimke, R. T. (1985). *J. Biol. Chem.* **260**, 2307–2314.

McKeon, C., Ohkuko, H., Pastan, I. & de Crombrugghe, B. (1982). *Cell*, **29**, 203–210.

Meijlink, F., Curran, T., Miller, A. D. & Verma, I. M. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 4987–4991.

Mitchell, P. J., Carothers, A. M., Han, J. H., Harding, J. D., Kas, E., Venolia, L. & Chasin, L. A. (1986). *Mol. Cell. Biol.* **6**, 425–440.

Mitchell, R., Zokas, L., Schreiber, R. D. & Verma, I. M. (1985). *Cell*, **40**, 209–217.

Nagamine, Y., Pearson, D., Altus, M. S. & Reich, E. (1984). *Nucl. Acids Res.* **12**, 9525–9541.

Perry, M., Thomsen, G. H. & Roeder, R. G. (1985). *J. Mol. Biol.* **185**, 479–499.

Piechaczyk, M., Blanchard, J. M., Marty, L., Dani, C., Panabieres, F., El Sabouty, S., Fort, P. & Jeanteur, P. (1984). *Nucl. Acids Res.* **12**, 6951–6963.

Reynolds, G. A., Basu, S. K., Osborne, T. F., Chin, D. J., Gil, G., Brown, M. S., Goldstein, J. L. & Luskey, K. L. (1984). *Cell*, **38**, 275–285.

Riccio, A., Grimaldi, G., Verde, P., Sebastio, G., Boast, S. & Blasi, F. (1985). *Nucl. Acids Res.* **13**, 2759–2771.

Saffer, J. D. & Singer, M. F. (1984). *Nucl. Acids Res.* **12**, 4769–4788.

Salser, W. (1977). *Cold Spring Harbor Symp. Quant. Biol.* **42**, 985–1002.

Schiimperli, D. (1986). *Cell*, **45**, 471–472.

Singer-Sam, J., Keith, D. H., Tani, K., Simmer, R. L., Shively, L., Lindsay, S., Yoshida, A. & Riggs, A. D. (1984). *Gene*, **32**, 409–417.

Sittman, D. B., Graves, R. A. & Marzluff, W. F. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 1849–1853.

Smith, T. F., Waterman, M. S. & Sadler, J. R. (1983). *Nucl. Acids Res.* **11**, 2205–2220.

Spencer, C. A., Gietz, R. D. & Hodgetts, R. B. (1986). *Nature (London)*, **322**, 279–281.

Stein, R., Sciaky-Gallili, N., Razin, A. & Cedar, H. (1983). *Proc. Nat. Acad. Sci., U.S.A.* **80**, 2422–2426.

Stepp, M. A., Kindy, M. S., Franzblau, C. & Sonenshein, G. E. (1986). *J. Biol. Chem.* **261**, 6542–6547.

Stone, E. M., Rothblum, K. N., Alevy, M. C., Kuo, T. M. & Schwartz, R. J. (1985). *Proc. Nat. Acad. Sci., U.S.A.* **82**, 1628–1632.

Swartz, M. N., Trautner, T. A. & Kornberg, A. (1962). *J. Biol. Chem.* **237**, 1961–1967.

Tautz, D., Trick, M. & Dover, G. A. (1986). *Nature (London)*, **322**, 652–656.

Toniolo, D., D'Urso, M., Martini, G., Persico, M., Tufano, V., Battistuzzi, G. & Luzzato, L. (1984). *EMBO J.* **3**, 1987–1995.

Treisman, R. (1985). *Cell*, **42**, 889–902.

Tykocinski, M. L. & Max, E. E. (1984). *Nucl. Acids Res.* **12**, 4385–4396.

Valerio, D., Duyvesteyn, M. G. C., Dekker, B. M. M., Weeda, G., Berkvens, T. M., van der Voorn, L., van Ormondt, H. & van der Eb, A. J. (1985). *EMBO J.* **4**, 437–443.

van Ooyen, A., Kwee, V. & Nusse, R. (1985). *EMBO J.* **4**, 2905–2909.

Venta, P. J., Montgomery, J. C., Hewett-Emmett, D. & Tashian, R. E. (1985). *Biochim. Biophys. Acta*, **826**, 195–201.

Vigneron, M., Barrera-Saldana, H. A., Baty, D., Everett, R. E. & Chambon, P. (1984). *EMBO J.* **3**, 2373–2382.

Wagner, M. & Perry, R. P. (1985). *Mol. Cell. Biol.* **5**, 3560–3576.

Wiedemann, L. M. & Perry, R. P. (1984). *Mol. Cell. Biol.* **4**, 2518–2528.

Williams, T. & Fried, M. (1986). *Nature (London)*, **322**, 275–279.

Wolf, S., Jolly, D. J., Lunnen, K. D., Friedmann, T. & Migeon, B. R. (1984*a*). *Proc. Nat. Acad. Sci., U.S.A.* **81**, 2806–2810.

Wolf, S. F., Dintzis, S., Toniolo, D., Persico, G., Lunnen, K. D., Axelman, J. & Migeon, B. R. (1984*b*). *Nucl. Acids Res.* **12**, 9333–9348.

Yen, P. H., Patel, P., Chinault, A. C., Mohandas, T. & Shapiro, L. J. (1984). *Proc. Nat. Acad. Sci., U.S.A.* **81**, 1759–1763.

Yoder, S. S. & Berget, S. M. (1985). *J. Virol.* **54**, 72–77.

Zhong, R., Roeder, R. G. & Heintz, N. (1983). *Nucl. Acids Res.* **11**, 7409–7425.

Edited by S. Brenner