Volume 10 Number 23 1982 Nucleic Acids Research

Asymmetrical distribution of CpG in an 'average' mammalian gene

Michael McClelland and Robert Ivarie

Department of Molecular and Population Genetics, University of Georgia, Athens, GA 30602, USA

Received 30 August 1982; Revised and Accepted 22 October 1982

Abstract

The frequency and distribution of the rare dinucleotide CpG was examined in 15 mammalian genes. CpG is highly methylated at cytosine in mammalian DNA (1,2) and 5-methylcytosine (5mC) is thought to undergo a transition mutation via deamination to produce thymine (3). This would result in the accumulation of TpG and CpA and depletion of CpG during evolution (4). Consistent with this hypothesis, the gene sample of 26,541 dinucleotides contained CpG at 40% the frequency expected by base composition and the CpG transition products, TpG+CpA, were significantly elevated at 124% of expected random frequency. However, because CpG occurs at only 25% of expected random frequency in the genome, the sampled genes were considerably enriched in this dinucleotide. CpGs were asymmetrically distributed in sequences flanking the genes. 5'-flanking sequences were enriched in CpG at 135% of the frequency expected assuming a symmetrical distribution of all the CpGs in the sampled genes ($p < 0.01$), while 3'-flanking regions were depleted in CpG at 40% of expected values ($p < 0.0001$). This asymmetry may reflect the role of 5-methylcytosine in gene expression. In contrast the frequencies of GpC and GpT+ ApC did not differ significantly from that predicted by base composition and these dinucleotides were not asymmetrically distributed.

INTRODUCTION

Mammalian DNA contains approximately 1 mole % 5-methylcytosine (5mC) of which about 90% is found in the relatively rare dinucleotide CpG (reviewed in 1,2). CpG occurs at about 1% of all dinucleotides in the genome compared to 4% measured for GpC and expected from base composition alone. As proposed by Salser (3), spontaneous deamination of 5mC to thymine generating the transition mutation products TpG and CpA may explain the low frequency of CpG in vertebrate DNA. Consistent with the hypothesis, Bird (4) has shown that there is a negative correlation between total genomic frequency of TpG+CpA versus CpG among vertebrate and invertebrate species differing in CpG content. As one test of the hypothesis at

© IRL Press Limited, Oxford, England.

0305-1048/82/1023-7865 $2.00/0

the gene level, we have measured the frequencies of CpG and TpG+CpA in coding and noncoding regions of 15 mammalian genes and report the findings here.

Cytosine methylation is also thought to play a role in the regulation of eukaryotic gene expression. In particular, several lines of evidence suggest that undermethylation of a gene is generally necessary but not sufficient for the gene's expression (reviewed in 1,2). We have, therefore, measured the frequency and distribution of CpG in six regions of the sampled genes for which sequence data were available (5'- and 3'-flanking regions; coding and intervening sequences; and 5' and 3' untranslated mRNA regions).

METHODS

Numbers of dinucleotides were obtained from published sequences (see Table 1) using the SEQ program of the SUMEX MOLGEN computer of Stanford University, Stanford, CA. Expected number of dinucleotides in Table 3 were calculated from the base composition of each region, which assumes a random distribution of the bases. In the χ^2 test, $\chi^2 = \frac{(\text{observed-expected})^2}{\text{expected}}$. Since there was one degree of freedom, a value of 4 was significant at $p < 0.5$ and a value of 7 at $p < 0.01$.

In Table 4 a random distribution of CpG's between regions was assumed. For each gene the size of each region and its C + G content relative to the whole gene were used to calculate expected numbers of dinucleotides which were then summed over all genes to give a cumulative total for each region. This was compared to the total observed number by use of a χ^2 test. Thus expected number of CpG for a region =

$$\sum_{\substack{\text{over}\\ \text{all}\\ \text{genes}}} \left(\text{number of CpG obs.} \times \frac{\text{number of bases in the region}}{\text{number of bases in gene}} \times \frac{p(C) \times p(G) \text{ in the region}}{p(C) \times p(G) \text{ in the gene}}\right)$$

where p(C) and p(G) are the proportion of C and G in the sequence. A similar procedure was followed for the other dinucleotides.

In Table 5 a pairwise comparison of dinucleotide frequencies between the 5' flanking sequences and each of the other gene regions was performed. The expected number in each comparison was

calculated assuming that the 5' flanking sequences had the same expected dinucleotide frequency as each of the other regions. Thus the 5' flanking sequence expected number compared to region$_i$ = number of CpGs observed in region$_i$

$$\times \frac{\text{number of bases in 5' flanking seq.}}{\text{number of bases in region;}}$$

$$\times \frac{p(C) \times p(G) \text{ in 5' flanking seq.}}{p(C) \times p(G) \text{ in region;}}$$

Manipulation of data was performed using an Apple II+ microcomputer and programs in Applesoft Basic written by M.M.

RESULTS

Genes used in the analysis are listed in Table 1 along with the number of base pairs in each of the six regions. Most of the sequences and dinucleotide frequencies were obtained by accessing the Stanford University SUMEX MOLGEN computer data file. Analysis was limited to genes from a single evolutionary order (mammals: human, rats and mice), and to genes transcribed by RNA polymerase II. Furthermore, we avoided using more than two members of closely related gene families and chose those for which 5' and 3' flanking sequences were available. Although a large nucleotide sequence database exists, most of it consists of cDNAs and closely related gene families such as the immunoglobulins. Distantly related genes were, however, used such as the human α2 and β globin genes estimated to have diverged 500 million years (Myr) ago (26). The α and β globin genes share about 50% sequence homology in coding regions, but have little homology in their intervening and flanking sequences (18,26). A total of 11 genes fitting the above criteria were available. In addition four cDNAs were included in the sample to increase the size of the 5' and 3' untranslated regions.

CpG Versus TpG+CpA Frequencies

Table 3 contains the number of CpG and GpC dinucleotides and the number of TpG, CpA, GpT and ApC in the total sequence sample and in each of the gene regions. Statistical significance was evaluated by a χ^2 test (see METHODS) against frequencies expected from the base composition of each region and of the total sequence. Base compositions are given in Table 2. 698 CpGs were found in the total sample which is 40% of the 1764 expected from base composition. This low value is significant at $p < 0.0001$. By contrast,

Table 1. Genes Used in the Analysis

			Total Gene (bp)	Gene Region[a] 5'flank (bp)	5'UT (bp)	Coding (bp)	Intervening (bp)	3'UT (bp)	3'Flank (bp)	Ref.
Human:	1.	β globin	2165	214	50	444	980	131	344	5,6
	2.	$\alpha_2$globin	1054	98	37	428	257	109	125	7
	3.	preproinsulin	1789	262	59	333	965	72	98	8,10
	4.	leukocyte α_1 interferon	1179	153	67	570	---	---	389	11
	5.	pro-opiomelanocortin	1203	---	---	666	---	164	51	12
	6.	delta globin	1976	122	50	444	1017	130	213	13
	7.	prechorionic gonadotropin α_1-subunit (cDNA)	621	---	51	351	---	220	---	14,15
	8.	fibroblast interferon B1	1835	286[b]	73	564	---	203	711	22,24
	9.	HLA	4118	244[b]	---	1076	1622	576	599	23
Mouse:	1.	metallothionein I	1560	300	73	186	694	132	175	16
	2.	immunoglobin γ_1 C-region	1823	---	---	975	575	93	82	17
	3.	α globin	1441	372	32	429	256	101	251	18
	4.	histone H4	968	228	29	312	---	59	340	19
	5.	α-fetoprotein (cDNA)	2012	---	41	1818	---	152	---	20
Rat:	1.	pregrowth hormone	2243	229	60	651	1269	101	33	21

[a] 5'flank, 5'-sequences preceding mRNA cap site; 5'UT, untranslated region of mRNA from cap site to start codon; coding and intervening sequences, or exons and introns, respectively; 3'UT untranslated region of mRNA from stop codon to poly-A addition site; 3'flank, 3'sequences after poly-A addition site.

[b] The cap site has not yet been mapped for the HLA gene; the whole region upstream from the start codon was classified as 5'flanking region.

Table 2. Base compositions

Gene regions[a]	Proportion of each base				C+G%
	(A)	(C)	(G)	(T)	
5'flanking	.289	.236	.263	.212	49.9
5'untranslated	.276	.319	.191	.213	51.0
coding sequences	.235	.281	.275	.210	55.5
intervening sequences	.226	.250	.245	.258	49.5
3'untranslated	.247	.244	.198	.310	44.3
3'flanking	.259	.214	.223	.304	43.7
total sequence	.244	.257	.252	.248	50.9
genomic average[b]	.288	.206	.213	.293	41.9

[a] Coding strand values;
[b] Taken from the values compiled by Shapira (30).

the "internal control" dinucleotide GpC was not significantly different from expected random frequency ($p > 0.1$). The CpG transition products TpG + CpA occurred at 124% of expected frequency in the total sample, a value significantly greater than expected ($p < 0.001$) while GpT + ApC were 83% of expected, significantly lower than random. Scarcity of CpG and increased abundance of TpG and CpA were found in all regions of the gene sample. Overall these results support the hypothesis (3,4) that a proportion of 5mCpGs have mutated to TpG/CpA during the course of mammalian gene evolution.

It has been proposed (1,3) that CpGs would be retained preferentially in coding regions where selection against mutation to TpG/CpA would be strong, and would be depleted in noncoding DNA where selection would be less likely to occur. By the same argument, TpG and CpA should be depleted in coding DNA and enriched in noncoding DNA. These expectations were met in part by the data in Table 3 in that on average CpG was more enriched in coding sequences (47%) than most noncoding DNA, e.g. intervening sequences (34%). However, the transition products were also enriched in coding sequences (137%) compared to intervening sequence DNA (118%), as was GpC at 106% vs 82%. A likely explanation for these findings is preferential codon usage (31). For instance, TpG and CpA each occurred at elevated but significantly different frequencies in coding DNA at 152% and 120%, respectively, whereas in intervening sequences, they each occurred at nearly equal frequencies (114% and 110%, respectively). A similar enrichment for these

Table 3. Observed and expected frequency of dinucleotides[a] (footnotes are below Table 4)

Region	Size[b]	CpG[c]	exp[d]	%of exp[e]	χ^2	GpC	%of exp	χ^2	TpG	exp	%of exp	χ^2	CpA	exp	%of exp	χ^2	GpT	%of exp	χ^2	ApC	%of exp	χ^2
5'flanking	2797	98	184	53	40	180	98	<4	148	147	101	<4	189	172	110	<4	127	86	<4	129	75	11
5'untranslated	633	12	37	32	17	36	97	<4	32	25	130	<4	73	55	134	6	25	101	<4	62	114	<4
coding	9244	336	722	47	207	762	106	<4	808	521	155	158	722	596	121	27	430	83	16	518	87	10
intervening seq.	7635	165	491	34	216	403	82	16	532	457	116	12	493	413	119	15	382	84	12	356	86	8
3'untranslated	2575	36	127	28	65	125	98	<4	215	155	139	24	145	146	99	<4	120	78	8	137	94	<4
3'flanking	3347	26	164	16	116	151	95	<4	284	223	128	17	208	177	118	5	192	86	4	146	83	5
Total	26,541	698	1764	40	644	1698	96	<4	2056	1607	128	126	1874	1609	117	44	1307	81	56	1367	85	37

Table 4. Statistical significance of the observed sum of dinucleotides in the given region of all examined genes normalized to the total sequence sample[a]

Region	Size[b]	CpG[c]	exp[d]	%of exp[e]	χ^2	GpC	exp	%of exp	χ^2	TpG	exp	%of exp	χ^2	CpA	exp	%of exp	χ^2	GpT	exp	%of exp	χ^2	ApC	exp	%of exp	χ^2
5'flanking	2797	98	73	135	9	180	177	102	<4	148	190	78	9	189	200	95	<4	127	148	87	<4	129	146	88	<4
5'untranslated	633	12	15	80	<4	36	36	101	<4	32	32	104	<4	73	64	115	<4	25	20	124	<4	62	46	133	5
coding	9244	336	286	118	9	762	695	110	6	808	667	121	30	722	694	104	<4	430	424	102	<4	518	506	102	<4
intervening seq.	7635	165	194	85	4	403	473	85	10	532	585	91	5	493	481	103	<4	382	372	103	<4	356	351	101	<4
3'untranslated	2575	36	50	72	4	125	122	102	<4	215	198	109	<4	145	170	85	<4	120	126	96	<4	137	120	114	<4
3'flanking	3347	26	65	40	23	151	158	96	<4	284	285	100	<4	208	206	101	<4	192	181	106	<4	146	150	97	<4
Total	26,541	698				1698				2056				1874				1307				1367			

[a] Deviations from expected are underlined if they differ by more than 25% from expected or if $p < 0.01$; [b] number of dinucleotides; [c] observed number; [d] expected number from base composition (see METHODS for derivation of expected numbers); [e] ratio of observed to expected number.

dinucleotides in coding DNA from birds, mammals and tumor viruses has been reported by Nussinov (32). There does, however, seem to be an inverse correlation between the average CpG and the average TpG and CpA frequencies in noncoding DNA; the 5' flanking sequences have higher CpG levels than the 3' flanking sequences. For TpG and CpA this situation is reversed. Intervening sequences and noncoding regions with intermediate levels of CpG have intermediate levels of TpG and CpA.

Asymmetrical Distribution of Dinucleotides

Although the frequency of CpG in the gene sample was only 40% of expected, this amounted to 2.6% of all dinucleotides in the sample, i.e., 2.6-fold greater than the genomic frequency of 1%. Furthermore CpG's were not uniformly distributed among the gene regions as the data in Table 3 shows. This disparity is illustrated graphically in Figure 1 where the ratio of observed to

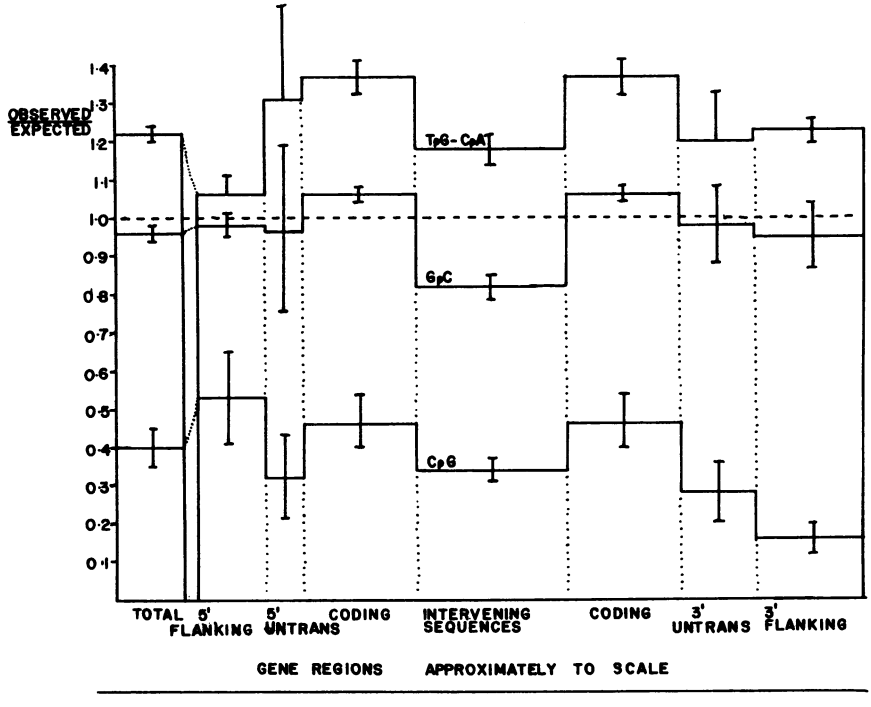

Figure 1. The observed/expected ratios for CpG, GpC and TpG and CpA were plotted against each of the six gene regions. Error bars were calculated as a variance from the means given in Table 3.

expected frequency of CpG, GpC and of TpG + CpA in the genome, total sequences, and the average for each gene region are plotted as a histogram. The 5' flanking region contained a high density of CpGs (3.5%) while the 3' flanking region was comparatively depleted compared to all other regions (0.8%). In fact, only the 3' flanking sequences contained CpG near the genomic frequency of 1%. Coding DNA and the untranslated noncoding DNAs were more abundant in CpG than the genome, containing 3.6% and 2.0% CpG respectively. Enrichment of CpG levels in globin coding sequences has also been reported by Salser (3).

The expected CpG content of each region sample was calculated assuming a random symmetrical distribution of all 698 CpGs in the total sequence sample and corrected for the G+C content of each gene region (see METHODS). The results are presented in Table 4. A χ^2 test indicated that the scarcity of CpG in the 3' region sample was significant to $p < 0.001$ and the abundance of the dinucleotide in the 5' region was significant to $p < 0.01$. In contrast, the observed frequencies of GpC, CpA, GpT and ApC in these regions were not significantly different from those expected from base composition. However, TpG was significantly less common than expected in the 5' flanking sequence. The statistical significance of increases in the frequencies of GpC and TpG ($p < .05$) in coding sequences (Table 4) may be due to the much larger sample sizes for these regions and biases in codon preference (31). ApC was higher than expected in the 5' untranslated region and GpC was depleted in intervening sequences. Some other dinucleotides also show consistent deviations from expected in some regions (M. McClelland, in preparation).

Table 5. 5' flanking sequence normalized to each of the other gene regions

5'flanking sequence dinucleotide	# obs	Normalized to 5' untranslated # exp	Normalized to 5' untranslated %of exp	Normalized to 5' untranslated χ^2	Normalized to coding # exp	Normalized to coding %of exp	Normalized to coding χ^2	Normalized to intervening # exp	Normalized to intervening %of exp	Normalized to intervening χ^2	Normalized to 3' untranslated # exp	Normalized to 3' untranslated %of exp	Normalized to 3' untranslated χ^2	Normalized to 3' flanking # exp	Normalized to 3' flanking %of exp	Normalized to 3' flanking χ^2	Normalized to total # exp	Normalized to total %of exp	Normalized to total χ^2
CpG	98	60	165	25	86	115	<4	62	159	21	52	188	40	30	326	154	73	135	9
GpC	180	179	101	<4	194	93	<4	151	119	6	181	100	<4	175	103	<4	177	102	<4
TpG	148	192	77	10	228	65	28	171	87	<4	205	87	<4	200	74	14	188	79	9
CpA	189	232	81	9	208	91	<4	205	92	<4	171	111	<4	217	87	<4	200	95	<4
GpT	127	150	85	<4	121	105	<4	123	103	<4	124	103	<4	135	94	<4	120	106	<4
ApC	129	197	65	24	149	86	<4	148	87	<4	161	80	7	152	85	<4	146	88	<4

Deviations from expected are underlined if they differ by more than 25% from expected or if $p < 0.01$.

The above test suffered the disadvantage that nearly 50% of the total sequence used in the normalization was coding DNA in which selection for sequence is different from that in noncoding DNA. Since coding DNA contained 3.6% CpG while noncoding untranslated regions each contained CpG at about 2.0%, the comparison with the total sequence was biased for higher CpG content contributed by the coding sequences. Accordingly, statistical significance was reassessed, using the same normalization method, in a pairwise comparison of CpG in the 5'-flanking region to each of the other gene regions. Table 5 shows that CpG enrichment in the 5' flanking region was statistically significant to $p < 0.001$ when compared to the untranslated RNA and 3' flanking sequences, but it was not significant compared to coding DNA, $p > 0.05$. In contrast GpC abundance did not differ from that expected by base composition when the 5' flanking region was compared to any other region ($p > 0.1$). Enrichment for CpG in the 5' flanking region and its scarcity in the 3'flanking region may well be a general feature of an average mammalian gene region transcribed by RNA polymerase II.

DISCUSSION

In this study the frequency and distribution of CpG and related dinucleotides was measured in fifteen mammalian genes and cDNAs. CpG was consistently lower than expected random frequency in the total sequence sample while its transition mutation products TpG and CpA (3) were generally more abundant than expected. These observations support the finding of a negative correlation between these dinucleotides in the total genomic DNA among species as reported by Bird (4). Furthermore, since 90% of 5mC is contained in CpG in vertebrates (1,2), these observations are also consistent with the hypothesis that 5mC is hypermutable via deamination to thymine (3).

The level of CpG of 2.6% in the total sample was much higher than the mammalian genomic average of 1.0%. However the difference between the gene sample and the genome could not be accounted for by variation in G+C content, nor was it confined to coding regions where selection for amino acid sequences takes place. One possible explanation for the observed disparity is that CpGs in genes in the germline remain relatively undermethylated relative to genomic

sequences and are less mutable as a consequence. A second, though not mutually exclusive, explanation may be that genes have maintained CpG levels by selection because of the role of 5mC in regulation (1,2,33-38). This could also account for the significant asymmetry in the frequency of CpG in the 5' and 3' flanking regions (Tables 4 and 5). Indeed, a 200 bp segment upstream from the mRNA transcription start site for the chicken adult β-globin gene was found to be hypersensitive to DNA nucleases and contained 70% C+G and 17 CpGs (33). Undermethylation of this region has been correlated with expression of the β-globin gene (34,35). However, it should be noted that our sample may not be typical of all mammalian genes because it contained predominantly tissue-specific genes whose expression has been correlated with changes in cytosine methylations (34,35,38). Such genes may have a higher average CpG content than consitutively expressed genes.

Accurate measurements of 5mC levels in various human cell types and tissues (including sperm) have recently been reported by Ehrlich et al. (39). Repeated sequences comprised 40% of the genome and contained about 1.5 mole % 5mC while unique DNA comprised 60% of the genome and contained about 0.6 mole % 5mC. Since genes analyzed here are found in unique sequences, these findings imply that the high CpG levels in the sampled genes may not necessarily mean a higher level of cytosine methylation. Furthermore, Sano and Sager (40) have observed that CCGG is highly methylated in bovine satellite DNA whereas $\overset{T}{A}CG\overset{A}{T}$ is less methylated and varies in a tissue-specific manner which implies that levels of CpG methylation can be overestimated when assayed by MspI/HpaII restriction enzyme analysis.

Our findings and those of Erhlich et al. imply that a substantial fraction of mammalian single-copy DNA may be low in CpG; recall that a large fraction of CpGs must be methylated since about 90% of 5mC is in CpG and the genome contains 5mC and CpG at about 1 mole % each. It follows that repeated DNA which contains 1.5% 5mC may have about 1.2% CpG (90% of 1.5%). Since genes contain 2.6% CpG then most unique DNA external to genes must be very low in CpG. Only then will enough CpG be available for 1.2 mole % CpG in repeated DNA (40% of genome) and 2.6 mole % in genes (> 5% of genome) while maintaining a genomic average near 1% for both 5mC

and CpG. While this may seem unrealistic, a 600 bp segment downstream from the poly A site in the human HLA gene contains only one CpG (23). In addition, as shown here, 3' sequences are depleted in CpGs and Bird has reported that large segments of the sea urchin genome are free of 5mC (41).

It is worth noting that our results represent average values and many individual genes differ markedly from them. For instance, human α_2 globin gene contains 84 CpGs in 1054 bp (or 8%) while the β globin gene contains only 10 CpGs in 2165 bp (or 0.5%). This is a 16-fold difference in CpG content between two genes that evolved from a common ancestor, share about 50% coding sequence homology, and are expressed in the same cell type albeit at different times. A similar disparity has been found in coding DNA of rabbit α and β globin genes (3). It seems unlikely that the considerable differences between these two distantly related genes arose by chance. Some selective mechanism acting on CpG levels around these genes must have taken place. Work is in progress evaluating variation in CpG frequencies and distributions within gene families and their pseudogenes, as one approach to this problem.

The enrichment in CpG of mammalian genes compared to the genome and the asymmetrical distribution of CpG between gene regions may be a general feature of mammalian DNA and reflect the role of CpG in mammalian gene regulation. Indeed these features are not found in bacterial or insect genes which do not have high levels of CpG methylation and an equally striking but very different distribution of CpG is found in higher plants which do have high levels of 5mCpG (43)(M. McClelland, in preparation).

ACKNOWLEDGMENTS

The authors thank Barbara Rutledge, John McDonald and Wyatt Anderson for critical comment on the manuscript and Colleen McElfresh for help in preparing the manuscript. This work has been supported by NIH research grants CA28826 and GM27973 to RDI. M.M. was supported by a University of Georgia Fellowship.

REFERENCES

1. Razin, A. and Riggs, A.D. (1980). Science 210:604-610.
2. Ehrlich, M. and Wang, R.Y.-H. (1981). Science 212:1350-1357.
3. Salser, W. (1977). Cold Spring Harbor Symp. Quant. Biol. 42:985-1002.

4. Bird, A.P. (1980). Nucleic Acids Research 8:1499-1504.
5. Lawn, L.M., Efstratiashus, A. O'Connell, C., and Maniatis, T. (1980). Cell 21:647-651.
6. Spritz, R.A., Jagadeesuwaran, P., Choudary, P.V., Bird, P.A., Elder, J.T., DeRiel, J.K., Manley, J.L. Gefter, M.L., Forget, B.G., and Weissman, S.M. (1981). Proc. Nat. Acad. Sci., U.S.A. 78:2455-2459.
7. Proudfoot, N.J. and Maniatis, T. (1980). Cell 21:537-544.
8. Bell, G.I., Pictet, R.L., Rutter, W.J., Cordell, B., Tischer, E., and Goodman, H.M. (1980). Nature 284:26-32.
9. Ulrich, A., Dull, T.J., Gray, A., Brosins, J. and Sures, I. (1980). Sci. 209:612-614.
10. Bell, G.I., Pictet, R., and Rutter, W.J. (1980). Nucleic Acids Res. 8:4091-4109.
11. Nagata, S., Mantei, N. and Weissmann, C. (1980). Nature 287: 401-408.
12. Chang, A.C.Y., Cochet, M., and Cohen, S.N. (1980). Proc. Nat. Acad. Sci. U.S.A. 77:4890-4894.
13. Spritz, R.A., Deriel, J.K., Forget, B.G., and Weissman, S.M. (1980). Cell 21:639-646.
14. Shine, J., Seeburg, P.H., Martial, J.A., Baxter, J.D., and Goodman, A.M. (1977). Nature 270:494-499.
15. Fiddes, J.C. and Goodman, H.M. (1979). Nature 281:351-356.
16. Glanville, N., Durnam, D.M. and Palmiter, R.D. (1981). Nature 292:267-269.
17. Honjo, T., Obata, M., Yamawaki-Katasks, Y., Kataoks, T., Kawakaski, T. Takahashi, N., and Mano, Y. (1979). Cell 18: 559-568.
18. Nishioka, Y. and Leder, P. (1979). Cell 18:875-882.
19. Seiler-Tuyns, A. and Birnstiel, M.L. (1981). J. Mol. Biol. 151:607-625.
20. Law, S.W. and Dugaiczyk, A. (1981). Nature 291:201-205.
21. Page, G.S., Smith, S., and Goodman, H.M. (1981). Nucleic Acids Res. 9:2087-2104.
22. Ohno, S. and Taniguchi, I. (1981). Proc. Natl. Acad. Sci. U.S.A. 78:5305-5309.
23. Malissen, M., Malissen, B., and Jordan, B.R. (1982). Proc. Nat. Acad. Sci., U.S.A. 79:893-897.
24. Lawn, R.M., Adelman, J., Franke, A.E., Houck, C.M., Grass, M., Najarian, R., and Goddel, D.V. (1981). Nucleic Acids Res. 9:1045-1052.
25. Quinto, C., Quiraga, M., Swain, W.F., Nikovits, Jr., W.C., Standing, D.N., Pictet, R.T., Velenzuela, P., and Rutter, W.J. Proc. Nat. Acad. Sci., U.S.A. 79:31-35.
26. Efstratiadis, A., Posakony, J.W., Maniatis, T., Lawn, R.M., O'Connell, C., Spritz, R.A., DeRiel, J.K., Forget, B.G., Weissman, S.M., Slightom, J.A., Blechl, A.E., Smithies, Baralle, F.E., Shouldero, C.C., and Proudfoot, N.J. (1980). Cell 21:653-668.
27. Niall, H.D., Hogan, M.L., Sauer, R., Rosenblum, I.Y. and Greenwood, F.C. (1971). Proc. Nat. Acad. Sci., U.S.A. 68:866.
28. Bewley, T.A., Dixon, J.S., and Li, C.H. (1972). Int. J. Peptide Prot. Res. 4:281.
29. Cooke, N.E., Coit, D., Weiner, R.I., Baxter, J.D., and Martial, J.A. (1980). J. Biol. Chem. 255:6502-6510.
30. Shapira, H.S. (1976). Handbook of Biochem. and Molec. Biol.: Nucleic acids (ed. G.D. Fasman) CRC Press, Cleveland, Ohio 2:241-275.

31. Grantham, R., Goutier, C., Jouy, M., Jacobzone, M., and Mercies, R. (1981). Nucleic Acids Res. 9:r43-r73.
32. Nussinov, R. (1981). J. Mol. Evol. 17:237-244.
33. McGhee, J.D., Wood, W.I., Dolan, M., Engel, J.D., and Felsenfeld, G. (1981). Cell 27:45-55.
34. McGhee, J.D. and Ginder, G.D. (1979). Nature 280:419-420.
35. Ginder, G.D. and McGhee, J.D. (1981). In Organization and Expression of Globin Genes, G. Stamatoyannopoulos, G. and Nienhuis, A.W., eds. A.R. Liss: N.Y. pp. 191-201.
36. Ivarie, R.D., Morris, J.A. and Martial, J.A. (1982). Molec. Cell. Biol. 2:179-189.
37. Ivarie, R.D. and Morris, J.A. (1982). Proc. Natl. Acd. Sci., U.S.A. 79:2967-2970.
38. van der Ploeg, L.H.T. and Flavell, R.A. (1980). Cell 19:947-958.
39. Ehrlich, M., Gama-Sosa, M.A., Huang, L-H, Midgett, R.M., Kuo, K.C., McCune, R.A., and Gehrke, C. (1982). Nucleic Acids Res. 10:2709-2721.
40. Sano, H. and Sager, R. (1982). Proc. Nat. Acad. Sci., U.S.A. 79:3584-3588.
41. Bird, A.P., Taggart, M.H., and Smith, B.A. (1979). Cell 17: 889-901.
42. Setlow, P. (1976). Handbook of Biochem. Molec. Biol.: Nucleic Acids (ed. G. Fasman) CRC Press, Cleveland, Ohio 2:312.
43. Gruenbaum, Y., Naveh-Many, T., Cedar, H. and Razin, A. (1981). Nature 292:860-862.