Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 957–961, February 1992
Genetics

Analysis of CpG suppression in methylated and nonmethylated species

(deamination/mCpG/dinucleotide/methylation/mutation)

DANIEL F. SCHORDERET* AND STANLEY M. GARTLER

Departments of Medicine and Genetics, University of Washington, Seattle, WA 98195

Contributed by Stanley M. Gartler, October 21, 1991

ABSTRACT **The development of nearest-neighbor analysis led to the finding that the frequency of the dinucleotide CpG is markedly depressed in vertebrates. One explanation of this suppression is that methylation of CpG found in vertebrates represents a mutational hot spot through deamination of methylcytidine to thymidine. We have examined the role of methylated CpG as a factor in CpG suppression by comparing CpG distributions in coding regions of 121 genes from six species, three with methylated DNA and three with nonmethylated DNA. Overall base composition shows that all species exhibit CpG suppression, with the methylated forms showing significantly greater suppression than nonmethylated forms. When the data are analyzed by CpG position, the mean values of the methylated forms exhibit greater suppression than nonmethylated forms at positions I–II and II–III, but there is considerable overlap of suppression scores for individual species. At position III–I, CpG suppression is marked in all methylated species, and it is reversed in all nonmethylated species. Our analysis supports the hypothesis that CpG patterns at positions II–III and III–I in methylated forms are affected by mutation acting through deamination of methylcytidine to thymidine. We speculate that the excess of CpGs at position III–I in nonmethylated forms may be related to a requirement for minimal thermal stability of the DNA.**

The development of nearest-neighbor analysis (1, 2) led to the finding that certain dinucleotides were present at frequencies that deviated significantly from random expectations. Significant deviations exist for CG vs. GC, AC vs. CA, TA vs. AT, and GT vs. TG, with the pattern of deviations being essentially uniform over a wide range of life forms (3). The frequencies of complementary dinucleotides (e.g., CG·GC) usually differ by no more than 10–20%. However, for the CG vs. GC pair in vertebrates, the difference is of the order of several hundred percent, with the dinucleotide CpG being markedly suppressed. Various ideas have been proposed to explain this CpG suppression. One explanation is based on the evidence that methylated CpG (mCpG) represents a mutational hot spot through deamination of mC to T (4–6); the C in a GpC dinucleotide is rarely, if ever, methylated (7). Other explanations consider restrictions imposed by the translation apparatus (8, 9), chromosomal structural restraints (10), and the possibility that CpG is depressed as a result of a universal coding rule (11).

To further examine the role of mCpG as a factor in CpG suppression, we have compared CpG distributions in coding regions of 121 genes from six species, three with methylated DNA and three with nonmethylated DNA. Overall base composition shows that all species exhibit CpG suppression, with the methylated forms showing significantly greater suppression than nonmethylated forms. When the data are analyzed by CpG position a more complicated and informative picture emerges. At the first two nucleotides of a codon (position I–II) and the second two nucleotides (position II–III), the mean values of the methylated forms exhibit greater suppression than the means of the nonmethylated forms, but there is considerable overlap of suppression scores for individual species. At position III–I, CpG suppression is marked in all methylated species, and it is reversed in all nonmethylated species. To more directly examine these patterns, we identified several genes whose sequences were available in both methylated and nonmethylated species and found that all the CpGs at position II–III and 95% of the CpGs at position III–I in nonmethylated forms had changed in methylated species, with most of the changes involving the substitution of T for C. Although our results support the mCpG mutational hot spot explanation of CpG suppression in methylated species at positions II–III and III–I, other factors must be invoked to explain patterns and CpG suppression observed in some nonmethylated species.

METHODS AND MATERIALS

We selected genes that had a complete coding sequence in GenBank from *Homo sapiens*, *Mus musculus*, and *Plasmodium falciparum* (12) (methylated forms) and *Saccharomyces cerevisiae* (13), *Drosophila melanogaster*, and *Caenorhabditis elegans* (nonmethylated forms; Table 1) (14). The correct open reading frame for each gene was determined through the use of GenBank's annotations and from published papers. The total CpG frequency was determined for each gene and the expected values were calculated as the product of the frequency of C in the gene times the frequency of G.

We also determined the CpG frequency with respect to the position of a CpG in the open reading frame. At position I–II a CGN (N is any base) is found only when arginine is specified and two-thirds of the arginine codons are CGN. Therefore, ignoring possible codon bias, the expected frequency of CpG at position I–II is two-thirds of the arginine-specifying codons in the gene. CpG at position II–III can occur only in the case of serine-, threonine-, proline-, and alanine-specifying codons, and the expected frequency of NCG codons is one-sixth of the serine codons plus one-fourth of the threonine, proline, and alanine codons. A CpG at position III–I occurs when a 5′ codon ends in C and the 3′ adjacent codon starts with a G. Any NNC and GNN pair can form a III–I CpG, and, therefore, we calculated the expected value of III–I CpGs as the product of the frequency of C at position III and G at position I.

For each species, a score of CpG suppression was calculated by the formula (observed − expected)/expected, which means that the score is negative when suppression occurs and positive when an excess of CpG is found.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*Present address: Institute of Medical Genetics, University of Geneva, CMU CH-1211, Geneva 4, Switzerland.

Table 1. List of genes analyzed

C. elegans	Ornithine decarboxylase	Creatine kinase M
Collagen 1	Methyltransferase	Esterase D
Collagen 2	3-Phosphoglycerate kinase	Fructose 1,6-bisphosphatase
Glyceraldehyde-3-phosphate dehydrogenase	Phosphatidylserine synthetase	Fumarase, mitochondrial
Myosin heavy chain	Phosphoglycerate mutase	Glyceraldehyde-3-phosphate dehydrogenase
Vitellogenin 4	Polymerase I	Glucose-6-phosphate dehydrogenase
Vitellogenin 5	Porin	Interferon β1
D. melanogaster	Profilin	Intestinal phosphatase, adult
Acetylcholinesterase	Pyruvate kinase	Myosin heavy chain, cardiac
Alcohol dehydrogenase	*S*-Adenosylmethionine synthetase	Orosomucoid
α-Amylase	Thymidylate kinase	Placental anticoagulant protein
Adenine phosphoribosyltransferase	Thymidylate synthase	Phosphoglycerate kinase pseudogene
Calmodulin	Triose phosphate isomerase	Phosphoglycerate kinase, testis
Collagen	tRNA methyltransferase	Phosphoglycerate kinase
Dopa decarboxylase	*H. sapiens*	Serum amyloid P component
Esterase-6	Acyl-CoA dehydrogenase	Steroid sulfatase
Even-skipped	Adenylate kinase	Surface antigen CD2
H1 histone	Alcohol dehydrogenase β1	Thymidine kinase
Alcohol dehydrogenase	Aldehyde dehydrogenase 2	Thymidylate synthase
Phospho*enol*pyruvate carboxykinase	Alkaline phosphatase, placental	Ubiquitin
Protein kinase C	α_1-Antitrypsin	*M. musculus*
raf protooncogene	α_2-Macroglobulin I	Adenosine deaminase
Cu–Zn superoxide dismutase	α_2-Plasmin INH1	Cardiac muscle α-actin
P. falciparum	α-Galactosidase	α-Fetoprotein
Exported antigen Ag 5.1	Amyloid A4	Apolipoprotein E
Circumsporozoite protein	Androgen receptor	Aspartate aminotransferase
Circumsporozoite protein related antigen	Angiotensinogen	Creatine kinase, muscle
Glutamic acid-rich protein	Apolipoprotein(a)	Cysteine-rich intestinal protein
Glycoprotein 185	Arginase, liver	Enkephalin
Hypoxanthine phosphoribosyltransferase	Argininosuccinate lyase	Glucocorticoid receptor
Merozoite major surface antigen	Asialoglycoprotein receptor H1	Glycerophosphate dehydrogenase
Small histidine + alanine-rich protein	Asialoglycoprotein receptor H2	Homeobox gene, chromosome 11
S. cerevisiae	β_1-Adrenergic receptor	Hypoxanthine
α-Galactosidase	β_2-Adrenergic receptor	Hypoxanthine phosphoribosyltransferase
Adenylate kinase	β-Tubulin	Interleukin 1
Alcohol dehydrogenase II	Butyrylcholinesterase, fetal	Interleukin 1 receptor
Arginase	c-*abl*	Muscle α-actin
β-Tubulin	Carbonic anhydrase II	Nicotinic acetylcholine receptor B
Glutamate dehydrogenase	Carbonic anhydrase III	Ornithine decarboxylase
Glycogen phosphorylase	Carcinoembryonic antigen	Placental calcium-binding protein
Glyceraldehyde-3-phosphate dehydrogenase	Collagen α_2 type I	Skeletal muscle α-actin
Hexokinase A	Complement component C5	Skeletal muscle actin
Hexokinase P1	Complement component C9	
Invertase		

We also identified four genes [phosphoglycerate kinase (*PGK1*), arginase, α-amylase, and glyceraldehyde-3-phosphate dehydrogenase] with complete coding sequences in both a methylated and a nonmethylated species. The amino acid sequences derived from these four genes were aligned and the corresponding DNA sequences were analyzed for CpG patterns. The PGK1 gene is known to be methylated in humans (7), and it is assumed that the other three genes are also methylated in the methylated species.

RESULTS

The analysis of total CpG content (Table 2) indicates that all the forms studied show CpG suppression. The scores are all significant at beyond the 0.01 level by χ^2 analysis, and the mean suppression score for the methylated forms (−0.533) is significantly greater than the comparable value for the nonmethylated species (−0.227).

Analysis of CpG suppression by position (Table 3) shows that the methylated species exhibit general CpG suppression, with the mean values at each position being significantly more suppressed than those of the nonmethylated species. However, there is marked overlap of individual suppression values at positions I–II and II–III for methylated and nonmethylated species. It is only at position III–I that the suppression patterns appear to be perfectly correlated with DNA methylation. The methylated species are markedly suppressed while the nonmethylated forms all show an excess of CpGs.

A more direct way of analyzing the above question is to compare coding sequences for the same genes in methylated and nonmethylated species. The sequences coding for PGK1,

Table 2. Total CpG composition and suppression scores

	No. of genes	No. of bases	No. of CpGs		Score
			Observed	Expected	
C. elegans	6	9,531	475	612.96	−0.225*
D. melanogaster	15	15,846	1057	1254.53	−0.157*
S. cerevisiae	25	32,778	1006	1437.48	−0.300*
P. falciparum	8	11,448	122	273.23	−0.553*
H. sapiens	48	73,443	2467	5315.81	−0.536*
M. musculus	19	19,452	644	1315.62	−0.510*

*Significantly different from zero; $P < 0.01$ (χ^2 analysis).

Table 3. CpG composition and score by codon position

	No. of genes	I–II		II–III		III–I	
		No. of CpGs	Score*	No. of CpGs	Score†	No. of CpGs	Score‡
C. elegans	6	138	+0.16	47	−0.70	290	+0.42
D. melanogaster	15	194	+0.21	261	−0.18	602	+0.44
S. cerevisiae	25	95	−0.67	172	−0.71	739	+0.54
P. falciparum	8	19	−0.57	18	−0.90	85	−0.75
H. sapiens	48	643	−0.19	653	−0.58	1171	−0.34
M. musculus	19	210	−0.02	123	−0.64	311	−0.29

*Expected value calculated from the number of arginine codons times 2/3.
†Expected value calculated from the number of threonine, proline, and alanine codons times 1/4 plus the number of serine codons times 1/6.
‡Expected value calculated from the product of the proportion of NNC and GNN codons.

arginase, α-amylase, and glyceraldehyde-3-phosphate dehydrogenase were identified in two or more species. The amino acid sequences they specify were aligned by pairs and the corresponding CpGs were compared, assuming that the nonmethylated forms (*S. cerevisiae, C. elegans, D. melanogaster*) represent phylogenetically older forms than the methylated species (*H. sapiens, M. musculus*). We restricted these comparisons to sequences specifying conserved amino acids, because the species being compared have been separated by long evolutionary periods, and nonconserved sites would most likely represent multiple mutational events. A part of the aligned sequence (codons 164–179) of yeast and human PGK1 is shown in Fig. 1. There are 4 CpGs in the yeast sequence and all have changed in the human sequence, with 3 of the changes being to TpG. There are 11 CpNs other than CpG in the same yeast sequence and only 2 differ.

All four gene comparisons exhibited similar patterns and, therefore, we present the statistical results of these analyses in a combined form (Table 4). There are 130 CpGs in the conserved regions of the four genes from the nonmethylated species, and 115 (88%) have changed at homologous positions in the methylated forms. Seventy-two (63%) of the changes were to TpG, 5 (4%) were to CpA, and the remainder (33%) were to other bases. As controls, we analyzed changes at the remaining dinucleotides containing a 5′ C in the same regions (Table 4). If the 5′ C changed, it was considered a mutation. There were 457 5′ C dinucleotides other than CpGs and 121 (26%) changed, which is significantly less than the changes from a CpG. It is shown (Table 4) that the CpGs in nonmethylated forms are distributed in a very unequal manner: 77% at position III–I, 18% at position I–II, and 5% at position II–III. The CpGs at positions III–I and II–III appear to be extremely unstable as >95% of them have changed in the homologous human genes as compared with 57% at position I–II. The control dinucleotides exhibit a similar pattern of instability, although in all cases the CpGs are significantly more unstable than the other 5′ C dinucleotides.

If the changes we have described are primarily due to the instability of mCpGs, then a comparison of homologous genes in nonmethylated forms should not show a CpG instability. We compared the glyceraldehyde-3-phosphate dehydrogenase genes in yeast and *C. elegans*, two nonmethylated forms, and found that, of 11 CpGs in yeast, a total of 3 (27%) had changed in *C. elegans*, all to TpG. At the control dinucleotides, 8 of 18 had changed (44%), which is not significantly different from changes at the CpG dinucleotides. It was also possible to analyze the alcohol dehydrogenase gene in three Drosophilids (*D. simulans, D. mauritiana*, and *D. melanogaster*) (15): there were 46 sites where at least one species had a CpG and only 3 sites (6%) varied between these species. At the control dinucleotides, 9 of 194 sites (5%) varied between these species, which is not significantly different from the frequency of change at the CpG dinucleotides.

All the deamination-related changes of mCpG at position I–II and half of the changes at positions II–III and III–I lead to nonconservative amino acid replacements, which should be subject to negative selection. The data in Table 1, which show that changes at positions I–II occur much less frequently than changes at position II–III and III–I, support this deduction. A comparison of CpGs in a functional gene and its pseudogene in a methylated species should avoid the negative selection problem and permit an estimate of the maximal instability of mCpG.

Such a comparison is possible with the human functional X chromosome-linked *PGK1* gene and a closely linked PGK-processed pseudogene. There are 24 CpGs in the functional *PGK1* gene, including part of the untranslated 5′ end, and 17 of these (71%) have changed in the pseudogene with 10 changes to CA and 3 changes to TG. Five of the 10 CpG changes in the coding region would be nonconservative if the pseudogene were functional. The control changes are ≈7%, which implies that CpG in this sequence is ≈10 times as unstable as a 5′ C control dinucleotide. Another point that can be made here is the extent of CpG loss as a result of mutation to TpG and CpA. The yeast PGK-coding region has 39 CpGs; the functional human X chromosome-linked gene has 15 CpGs and the X chromosome-linked pseudogene has only 9. Ten of the CpGs in the functional X chromosome-linked *PGK1* gene have been lost in the X chromosome-linked pseudogene but 4 CpGs have appeared at new positions, suggesting an eventual mutational balance.

DISCUSSION

There is considerable evidence that mCpG represents a mutational hot spot (6, 16). Cytosine may be the most unstable base in DNA. It has been estimated that ≈100 cytosines a day are spontaneously deaminated in a human cell; this leads to an unrepaired 5′ mC nucleotide mutation rate of $\approx 1 \times 10^{-5}$. This mutation rate is at least 3 orders of magnitude greater than the usual estimates of mutation rates at the nucleotide level. Under these conditions, methylated sites would have a transient existence, and it was originally assumed that this type of mutation could not be repaired. We now know that in both bacteria and mammalian cells, a specific repair system exists

Human AAT GAT GCT TTT GGC ACT GCT CAC AGA GCC CAC AGC TCC ATG GTA GGA

Yeast AAC GAT GCC TTC GGT ACC GCT CAC AGA GCT CAC TCT TCT ATG GTC GGT

FIG. 1. Yeast and human PGK sequences extending from codon 164 to 179. CpGs are double underlined; CCs, CTs, and CAs are overlined.

Table 4. Distribution of changes at CpGs and control dinucleotides in conserved regions in PGK1, arginase 1, α-amylase 2, and glyceraldehyde-3-phosphate dehydrogenase from species with and without methylation

Position	CpGs in nonmethylated species*	CpGs changed in methylated species†	Control dinucleotides in nonmethylated species‡	Control dinucleotides changed in methylated species
I–II	23	13	86	11
II–III	7	7	194	30
III–I	100	95	177	80

The data are derived from the following pairs: glyceraldehyde-3-phosphate dehydrogenase, *S. cerevisiae–H. sapiens* and *C. elegans–H. sapiens*; PGK1, *S. cerevisiae–H. sapiens*; arginase, *S. cerevisiae–H. sapiens*; α-amylase, *D. melanogaster–H. sapiens*.
**S. cerevisiae, D. melanogaster*, and *C. elegans*.
†*H. sapiens*.
‡CC, CT, and CA.

that preferentially removes thymine from TpG pairs, the expected product of mCpG deamination (17–19). However, the repair system is far from perfect since the majority of human point mutations seem to derive from transitions at CpG dinucleotides (20–24). Direct proof that such transitions involve mCpG was recently provided by Rideout *et al.* (25). There is considerable statistical data (6) that supports this general conclusion as well as our own observations in this report. Almost 90% of the CpGs in the nonmethylated species have mutated in the methylated species and two-thirds have changed to TpG or CpA, the dinucleotides expected from deamination of mCpG. In view of the instability of mCpG, one would expect a marked and general CpG suppression in methylated species compared to nonmethylated forms. However, this is not the case; all species exhibit CpG suppression at position II–III, and CpG distributions at position I–II show marked overlap of suppression values. It is only at position III–I that a consistent difference is seen between methylated and nonmethylated species. In fact, *C. elegans, D. melanogaster*, and *S. cerevisiae* have significant excesses of CpGs at position III–I. These results appear to be at variance with the original nearest-neighbor observations that showed marked CpG suppression for vertebrates as compared to nonvertebrates. However, the nearest-neighbor observations were made on total DNA, while our observations are restricted to coding sequences. We now know that intronic DNA in vertebrates is more CpG suppressed than coding DNA (21–29); 90% of the genes in our methylated species are from vertebrates and vertebrates have much more intronic DNA in their genomes. On the other hand, our nonmethylated forms are all nonvertebrate species and generally have much less intronic DNA.

How can we account for the CpG suppression at positions I–II and II–III in the species with nonmethylated DNA? One possibility is that these species have gone through a methylation period in their evolutionary past and that present day CpG frequencies reflect this history. This idea leaves unexplained the excess of CpGs at position III–I in these forms. Considering the long period that these forms must have existed without methylation and the possibility of reverse transitions (24), the present day CpG frequencies do not offer strong support for this idea. Furthermore, one can argue that early eukaryotes lost the prokaryotic defense mechanism of DNA modification and restriction as a result of the evolution of a superior nuclear envelope. Later some eukaryotes evolved an altered DNA modification system for other purposes.

It seems more likely that CpG suppression at positions I–II and II–III in nonmethylated forms reflect restrictions imposed by the translation apparatus for the purpose of optimizing general translation efficiency in microbial forms. This idea was first put forward by Subak-Sharpe *et al.* (8), was given experimental support by Ikemura (30) and Bennetzen and Hall (31), and was extensively analyzed by Grosjean *et al.* (32, 33) and Grantham *et al.* (34). Ikemura (30) has shown that codon usage in *Escherichia coli* and *S. cerevisiae* is related to the relative abundance of the various tRNA molecules. There is a particularly strong correlation between genes expressed at high levels and the use of the most frequent tRNAs. Hoekema *et al.* (35) have demonstrated a significant decrease of expression in the highly expressed yeast *PGK1* gene when major codons are replaced with synonymous minor ones. Highly expressed genes in yeast tend to avoid CGN and NCG codons, while those genes whose mRNAs represent a small fraction of the total cellular RNA make use of a number of different isoacceptors. While there are some general similarities in the *E. coli* and *S. cerevisiae* codon usage patterns, there are some striking differences, especially with respect to codons containing CpG. The preferred arginine codon in *E. coli* is CGU, while in yeast it is AGA. The preferred codon for proline in *E. coli* is CCG, while in yeast the preferred codon is CCA (CCG appears never to be used). Since *E. coli* has methylated DNA, the above pattern differences imply that methylation was not involved in the evolutionary origin of the isoacceptor patterns.

An explanation of the significant excess of CpGs at position III–I in nonmethylated forms is not immediately apparent, and we can only speculate as to what this excess may mean. One possibility is that this excess reflects a requirement for minimal thermal stability of the DNA. CG pairs are more stable than AT pairs and methylated CGs are more stable than nonmethylated ones (36). The general CpG suppression in methylated forms may be compensated by methylation of remaining CpGs, whereas in nonmethylated forms the CpG suppression found at positions I–II and II–III could be compensated by the excess observed at position III–I.

While CpG suppression is general in methylated forms, there are some regions that escape CpG suppression. Russell *et al.* (9) first noted that some G+C-rich tRNAs and a 5S RNA gene were not CpG suppressed. We have analyzed a further 13 human tRNA genes and found that they also are not CpG suppressed. Furthermore, restriction analysis of three of these tRNA genes indicates that they are methylation free (37). Another example of escape from CpG suppression is the G+C-rich promoter regions of mammalian housekeeping genes. These regions may be up to 1 kilobase long and have a G+C composition ranging from 60% to >80%. They are also methylation free (38, 39). One exception to this pattern is the G+C-rich promoters on the inactive X chromosome, which are methylated as part of their repression mechanism (40). However, even these regions are demethylated through

most, if not all, of their germ cell history (41), thus minimizing mutational damage through the deamination pathway.

There appear to be at least three processes that determine CpG patterns. Optimization of translation efficiency may play a significant role in explaining the variations in CpG distributions at different positions in nonmethylated forms. It seems possible that optimization of translation efficiency may have lost some of its selective role in vertebrates (42, 43). For methylated forms it appears that CpG patterns at positions II–III and III–I are significantly affected by mutation acting through elimination of mCpG. All the CpGs at position II–III in nonmethylated species were changed in the homologous genes of methylated species (Table 4). The small number of observations for this position reflects the fact that even nonmethylated forms are markedly CpG suppressed at position II–III. At position III–I, 95% of the CpGs in nonmethylated forms were changed in methylated species, and these changes were primarily to TpG, the substitution predicted by the deamination model. Changes at position I–II were much less frequent, which must be related to the fact that all deamination-caused substitutions at position I–II lead to missense mutation. Methylation and mutation at CpGs most likely account as well for the marked CpG suppression found in intronic DNA of vertebrates (26–29). Smith *et al.* (27) have argued that the absence of a proportional increase in TpG in intronic DNA indicates that the deamination mutational hot spot theory may not explain CpG suppression in this instance. One would not expect the substituted T to be selectively retained in intronic DNA as it is in the III–I position. The methylation-induced CpG suppression in introns and at position III–I in exonic DNA in methylated species together with the excess of CpG at position III–I in nonmethylated forms may be sufficient to explain the marked difference in CpG suppression between nonvertebrates and vertebrates. A third factor affecting CpG patterns must be a signal(s) protecting G+C-rich regions from methylation and subsequent CpG suppression (44). The methylation of G+C-rich promoters on the mammalian inactive X chromosome (45) and the G+C-rich 5′ regions of some L1 elements (46, 47) indicates that such a signal can be overridden. The elucidation of the signal and its interactions could tell us a good deal about methylation targets.

We are grateful to Drs. Eric Selker and Benjamin Hall for critical comments on our manuscript. This work was supported by National Institutes of Health Grant HD-16659. D.F.S. is supported by the Swiss National Science Foundation (Grant 32-9253.87) and S.M.G. is a recipient of a National Institutes of Health Career Award.

1. Josse, J., Kaiser, A. D. & Kornberg, A. (1961) *J. Biol. Chem.* **236,** 864–875.
2. Swartz, M. N., Trautner, T. A. & Kornberg, A. (1962) *J. Biol. Chem.* **237,** 1961–1967.
3. Setlov, P. (1976) in *Practical Handbook of Biochemistry and Molecular Biology*, ed. Fasman, G. D. (CRC, Cleveland), pp. 312–318.
4. Salser, W. (1977) *Cold Spring Harbor Symp. Quant. Biol.* **42,** 985–1002.
5. Coulondre, C., Miller, J. A., Farabaugh, P. J. & Gilbert, W. (1978) *Nature (London)* **274,** 775–780.
6. Bird, A. P. (1980) *Nucleic Acids Res.* **8,** 1499–1504.
7. Pfeifer, G. P., Steigerwald, S. D., Mueller, P. R., Wold, B. & Riggs, A. D. (1989) *Science* **246,** 810–813.
8. Subak-Sharpe, H., Bürk, R. R., Crawford, L. V., Morrison, J. M., Hay, J. & Keir, H. M. (1966) *Cold Spring Harbor Symp. Quant. Biol.* **31,** 737–748.
9. Russell, G. J., Walker, P. M. B., Elton, R. A. & Subak-Sharpe, J. H. (1976) *J. Mol. Biol.* **108,** 1–23.
10. Lennon, G. G. & Fraser, N. W. (1983) *J. Mol. Evol.* **19,** 286–288.
11. Ohno, S. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 9630–9634.
12. Pollack, Y., Kogan, N. & Golenser, J. (1991) *Exp. Parasitol.* **72,** 339–344.
13. Proffitt, J. H., Davie, J. R., Swinton, D. & Hattman, S. (1984) *Mol. Cell. Biol.* **4,** 985–988.
14. Blumenthal, R. M. (1989) *Focus* **11,** 41–46.
15. Cohn, V. H., Thompson, M. A. & Moore, G. P. (1984) *J. Mol. Evol.* **20,** 31–37.
16. Lindahl, T. & Nyberg, B. (1974) *Biochemistry* **13,** 3405–3410.
17. Brown, T. C. & Jiricny, J. (1987) *Cell* **50,** 945–950.
18. Brown, T. C. & Jiricny, J. (1988) *Cell* **54,** 705–711.
19. Hare, J. T. & Taylor, J. H. (1988) *Gene* **74,** 159–161.
20. Youssoufian, H., Kazazian, H. H., Jr., Phillips, D. G., Aronis, S., Tsiftis, G., Brown, V. A. & Antonarakis, S. E. (1986) *Nature (London)* **324,** 380–382.
21. Cooper, D. N. & Youssoufian, H. (1988) *Hum. Genet.* **78,** 151–155.
22. Koeberl, D. D., Bottema, C. D. K., Buerstedde, J. M. & Sommer, S. S. (1989) *Am. J. Hum. Genet.* **45,** 448–457.
23. Green, P. M., Montandon, A. J., Bentley, D. R., Ljung, R., Nilsson, I. M. & Giannelli, F. (1990) *Nucleic Acids Res.* **18,** 3227–3231.
24. Sved, J. & Bird, A. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 4692–4696.
25. Rideout, W. M., III, Coetzee, G. A., Olumi, A. F. & Jones, P. A. (1990) *Science* **249,** 1288–1290.
26. Fraser, N. W., Burdon, R. H. & Elton, R. A. (1975) *Nucleic Acids Res.* **2,** 2131–2146.
27. Smith, T. F., Waterman, M. S. & Sadler, J. R. (1983) *Nucleic Acids Res.* **11,** 2205–2220.
28. Adams, R. L. P. & Eason, R. (1984) *Nucleic Acids Res.* **12,** 5869–5877.
29. Beutler, E., Gelbart, T., Han, J., Koziol, J. A. & Beutler, B. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 192–196.
30. Ikemura, T. (1981) *J. Mol. Biol.* **151,** 389–409.
31. Bennetzen, J. L. & Hall, B. D. (1982) *J. Biol. Chem.* **257,** 3026–3031.
32. Grosjean, H., Sankoff, D., Min Jou, W., Fiers, W. & Cedergren, R. (1978) *J. Mol. Evol.* **12,** 113–119.
33. Grosjean, H. & Fiers, W. (1982) *Gene* **18,** 199–209.
34. Grantham, R., Perrin, P. & Mouchiroud, D. (1986) *Oxford Surv. Evol. Biol.* **3,** 48–82.
35. Hoekema, A., Kastelein, R. A., Vasser, M. & deBoer, H. A. (1987) *Mol. Cell. Biol.* **7,** 2914–2924.
36. Collins, M. & Myers, R. M. (1987) *J. Mol. Biol.* **198,** 737–744.
37. Schorderet, D. F. & Gartler, S. M. (1990) *Nucleic Acids Res.* **18,** 6965–6969.
38. Max, E. E. (1984) *Nature (London)* **310,** 100.
39. Bird, A., Taggart, M., Frommer, M., Miller, O. J. & Macleod, D. (1985) *Cell* **40,** 91–99.
40. Pfeifer, G. P., Tanguay, R. L., Steigerwald, S. D. & Riggs, A. D. (1990) *Genes Dev.* **4,** 1277–1287.
41. Driscoll, D. J. & Migeon, B. R. (1990) *Somatic Cell Mol. Genet.* **16,** 267–282.
42. Ikemura, T. (1985) *Mol. Biol. Evol.* **2,** 13–34.
43. McCarrey, J. R. (1990) *Nucleic Acids Res.* **18,** 949–955.
44. Selker, E. U. (1990) *Trends Biochem. Sci.* **15,** 103–107.
45. Gartler, S. M. & Riggs, A. D. (1983) *Annu. Rev. Genet.* **17,** 155–190.
46. Furano, A. V., Robb, S. M. & Robb, F. T. (1988) *Nucleic Acids Res.* **16,** 9215–9231.
47. Scott, A. F., Schmeckpeper, B. J., Abdelrazik, M., Comey, C. T., O'Hara, B., Rossiter, J.-P., Cooley, T., Heath, P., Smith, K. D. & Margolet, L. (1987) *Genomics* **1,** 113–125.