J Mol Evol (1992) 34:219–230

Journal of Molecular Evolution

Methylation Sites in Angiosperm Genes

M. Gardiner-Garden,[1] J.A. Sved,[2] and M. Frommer[1,3]

[1] The Kanematsu Laboratories, Royal Prince Alfred Hospital, Missenden Road, Camperdown, NSW 2050, Australia
[2] School of Biological Sciences A12, University of Sydney, NSW 2006, Australia
[3] CSIRO Division of Biomolecular Engineering, PO Box 184, North Ryde, NSW 2113, Australia

Summary. The extent to which CpG dinucleotides were depleted in a large set of angiosperm genes was, on average, very similar to the extent of CpG depletion in total angiosperm genomic DNA and far less than the extent of CpG depletion in vertebrate genes. Gene sequences from *Arabidopsis thaliana,* a dicotyledonous species with relatively low levels of total 5-methylcytosine, were just as CpG depleted as the angiosperm genes in general. Furthermore, levels of TpG and CpA, the potential deamination mutation products of methylated CpG, were elevated in *A. thaliana* genes, supporting a high rate of deamination mutation as the cause of the CpG deficiency. Using a method that takes into account the dinucleotide frequencies within each sequence of interest, we calculated the expected frequencies of CpNpG trinucleotides, which are also highly methylated in angiosperm genomes. CpNpG trinucleotides were not extensively enriched or depleted in the angiosperm genes. Two hypotheses could account for our results. Differential depletion of CpG and CpNpG within angiosperm genes and differential depletion of CpG in angiosperm and vertebrate genes could arise from different efficiencies of mismatch repair or from different levels of cytosine methylation in the cell lineages that contribute to germ cells.

Key words: CpG dinucleotides — CpNpG trinucleotides — Angiosperm gene sequences — *Arabidopsis thaliana* gene sequences — Methylation sites — Deamination mutation — DNA sequence analysis

Offprint requests to: M. Gardiner-Garden

Introduction

DNA methylation has been widely implicated in gene regulation within both angiosperm and vertebrate nuclear genomes (for reviews, see Hepburn et al. 1987; Nelson 1988; Cedar and Razin 1990; Holliday et al. 1990), yet angiosperm and vertebrate genomes differ considerably in the frequency and type of methylatable sites. Angiosperm genomes are extensively methylated at cytosine residues in both CpG dinucleotides and CpNpG trinucleotides (Shapiro 1976; Bonen et al. 1980; Gruenbaum et al. 1981a), whereas in vertebrate genomes, only CpG is extensively methylated (Sinsheimer 1955; Gruenbaum et al. 1981b; Nyce et al. 1986).

In both angiosperm and vertebrate genomes, CpG dinucleotides are present at less than the frequency expected from the base composition (Swartz et al. 1962). The most probable cause of this CpG depletion is the high rate at which 5-methylcytosine (5^mC) can undergo deamination to thymine (Coulondre et al. 1978), so that methylated CpG dinucleotides will tend to mutate to TpG on the same strand and to CpA on the complementary strand (Bird 1980). The deamination theory predicts that TpG and CpA dinucleotides will be elevated where CpG dinucleotides are depleted, which is generally the case in both angiosperm and vertebrate genomes (Swartz et al. 1962; Russell et al. 1971; Bird 1980). However, although a similar proportion of total CpG is methylated in angiosperms (70–80%) and vertebrates (40–80%) (Gruenbaum et al. 1981a,b; Naveh-Many and Cedar 1982; Woodcock et al. 1987), angiosperm genomes are not as CpG depleted as are vertebrate genomes. In total wheat and cauliflower DNA, the frequency of CpG is 0.75 times the level expected from the base composition (Swartz et al. 1962; Rus-

sell et al. 1971), compared with a frequency of only 0.15–0.35 times the expected level in various vertebrate genomes (Josse et al. 1961; Swartz et al. 1962; Russell et al. 1976). Hence, angiosperm genomes contain a higher proportion of CpG dinucleotides (2.5–4%), relative to total dinucleotides, than do vertebrate genomes (0.5–1.5%) (Setlow 1976). The higher frequency of CpG methylatable sites and the additional CpNpG methylatable sites both contribute to the considerable difference in level of methylation found in angiosperms as compared with vertebrates (Gruenbaum et al. 1981a). In angiosperms, 6.5–37% of all cytosines are methylated, with the vast majority of reported assays yielding values >20% (Shapiro 1976; Wagner and Capesius 1981; Leutwiler et al. 1984); whereas in vertebrates, 3.5–15% of all cytosines are methylated, with almost all values <10% (Vanyushin et al. 1970; Shapiro 1976).

In the case of mammalian and avian DNA, numerous studies have shown that the gene sequences, like the genome as a whole, are generally highly depleted in CpG and contain elevated levels of TpG and CpA. However, many vertebrate gene sequences contain regions known as CpG islands, which are of the order of 1 kb in length and contain CpG dinucleotides at approximately the frequency expected from the base composition (McClelland and Ivarie 1982; Tykocinski and Max 1984; Bird et al. 1985; Bird 1986; Gardiner-Garden and Frommer 1987). CpG-depleted regions associated with mammalian and avian genes can vary considerably in their G+C composition, and the extent of CpG depletion is not strongly dependent on G+C content (Gardiner-Garden and Frommer 1987). CpG islands, on the other hand, are always G+C rich.

The genomes of most angiosperm species contain large amounts of repeated sequences (Flavell et al. 1974). If methylation patterns in genes and repeated sequences differ, then differences in the amount of CpG and/or CpNpG depletion may result. Hence, the established levels of methylatable sites in total genomic DNA may not reflect the levels in and around genes. Until recently, only a limited number of angiosperm gene sequences have been available for study, so that the extent to which angiosperm gene sequences are CpG depleted is not clear. McClelland (1983), in an analysis of 11 sequences representing 6 gene families, and Boudraa and Perrin (1987), in an analysis of 25 gene sequences representing 14 gene families, found considerable depletion of CpG dinucleotides and an elevation of TpG+CpA dinucleotides. Numerous angiosperm gene sequences are now available in GenBank. We have analyzed 111 genomic gene sequences, comprising 47 gene families from a total of 26 species, to determine the levels of the methylatable site, CpG, and the nonmethylatable site, GpC, and have compared these levels with those expected from the base composition of each sequence. We have analyzed the coding and noncoding regions of each protein-coding gene, the entire sequence of each gene coding for RNA functional products (RNA-coding gene), and have compared the results for genes from monocotyledonous (monocot) and dictoyledonous (dicot) plants. Because the *Arabidopsis* genome contains a low level of total 5^mC and the limited data available suggest that *Arabidopsis* genes may be strikingly hypomethylated (Leutwiler et al. 1984; Pruitt and Meyerowitz 1986), we have carried out a separate study of CpG distribution in *Arabidopsis* gene sequences.

Because CpNpG trinucleotides are heavily methylated in angiosperm genomes (Gruenbaum et al. 1981a), evidence of CpNpG depletion with TpNpG and CpNpA elevation might be found in sequence data (McClelland 1983). However, McClelland (1983) observed a small overall enrichment of CpNpG trinucleotides, relative to an expected level calculated using a combination of the base composition and observed CpG dinucleotide frequencies, suggesting that CpNpG depletion may not be a characteristic of angiosperm genes. In angiosperm and vertebrate genomes, several dinucleotides in addition to CpG do not occur at the frequency expected from the base composition (Swartz et al. 1962; Russell et al. 1971). We have therefore derived expected frequencies of trinucleotides from dinucleotide composition rather than base composition. We have compared the observed with the expected frequencies of the methylatable site, CpNpG, and the nonmethylatable site, GpNpC, and have similarly analyzed the frequencies of TpNpG and CpNpA trinucleotides, which are the deamination products of methylated CpNpG trinucleotides. We discuss our results in terms of the deamination theory and methylation in the germline.

Methods

Sequences Analyzed. The 111 genomic sequences of angiosperm genes included in the analysis were obtained from GenBank Release 60, June 1989, and are listed below according to their GenBank code. Where one representative only of each type of gene was required, those genes included in each analysis are marked as follows: monocot protein-coding genes, M; dicot protein-coding genes, D; *Arabidopsis thaliana* genes, A; and RNA-coding genes, R. Protein-coding genes derived from monocot plants were mzeact1g (M), mzeadh1s (M), mzeadh2n, blyalr, blyamyabd, whtamya (M), athlhcp2 (A, D), lgiab19 (M), whtcab, whtgir (M), whtgliabd, whtglna, whtglgb (M), mzeglut2e, ricglutg (M), whtglut1 (M), richis3, mzeh3c4, mzeh4c14, whth3, whth4 (M), blyhorb (M), mzea1g (M), astpht3a (M), mzesusysg (M), mzei19, and mzeze15g (M). Those derived from dicot plants were athact1a (A, D), soyact3g, hnng5alb2 (D), peaabn1, athadh (A, D), amachs, athchs (A, D), athlhcp2 (A, D), peacab80, petcab221, petcab37,

tobcabb, tomcab, tomcbpa, tomcbpe, tomcbpi, soybpsp (D), athepsps (A, D), tomgtoma (D), darext (D), alfglnag (D), petgcr1, tobgrpa (D), soyglycab (D), trthb (D), soyhsp, soyhsp175 (D), soyhsp179, athh3gb, athh4gb (A, D), pealeca, phvlect (D), soylea, vfaleb4, pealega (D), phvlba, soylba, soylbgi, potls1g (D), tobatp21 (D), bnanapa (D), soynod23g (D), soynod24g/soynod24h (D), soynod35g (D), potpatg1 (D), tobpr1ag (D), phvbcsp (D), phvdleca (D), phvdlecb, athpcg (A, D), spipcg, soyprp1 (D), potinhw1 (D), potpi2g, soyciipi, tomwipig, hnnrbcs, pearubps, petrbcs08 (D), petrbcs1a, soyrubpa, tobrbpco, tomrbcsa, tomrbcsc, cotspa (D), cotsph (D), athtuba (A), athtub1a, athtubb, soysb1tub (D), and soysb2tub. Genes encoding RNA products that were derived from monocot plants were ricrgsbha (R), mzerg17s, ricrge (R), and ricrgh. Those derived from dicot plants were rccricin, flxrga, luprga1 (R), minrga, luprgb, soyrge, clirrn26 (R), tobtrnyx, and soytgmi (R). For the separate analyses of *Arabidopsis* genes, we included 22 additional genomic *Arabidopsis* gene sequences representing eight additional gene families, provided by GenBank Release 64, June 1990, and EMBL Release 23, May 1990. Those marked "/E" are listed according to EMBL code: atacp/E (A), athat2s1 (A), athat2s2, athat2s3, athat2s4, atatp/E (A), atef1a23 (A), atef1aa4/E, athrpcb (A), x16077 (A), athrbcsb (A), atcra1/E, atcrb/E (A), athug22, athu23, athu24, athug25, athug27, athu5RNA (A), attrnami/E (A), athtrpb (A), and atubq4/E (A). We included only sequences that contained the complete protein- or RNA-coding region of a gene and that were >250 bp in length, with no more than 10 bases undetermined. Where the exon and intron sequences of two or more members of a multigene family in one species were almost identical, we included only the gene with the most sequence information. The name and species of each of the genes above are listed elsewhere (Gardiner-Garden and Frommer 1992).

Computer Generation of Random Sequences Using Observed Dinucleotide Frequencies. Sequences of defined length, with defined dinucleotide counts, were generated by successively adding dinucleotides from a dinucleotide pool of the same size and composition as that of the original sequence of interest. If, for example, the original sequence commenced with the nucleotide A, the first dinucleotide was chosen at random from the pool of ApA, ApC, ApG, and ApT. If the dinucleotide ApC was chosen, then the next dinucleotide was chosen from the pool of CpA, CpC, CpG, and CpT. Repeated use of this procedure led to a complete sequence in about 25% of cases. In about 75% of cases, the procedure failed to produce a complete sequence, when suitable dinucleotides had been exhausted from the pool, and the sequence was discarded.

Calculation of Expected Dinucleotide and Trinucleotide Frequencies. Values for the expected number of a given dinucleotide in each sequence were calculated as described previously (Gardiner-Garden and Frommer 1987). The expected number of a given trinucleotide was estimated as the average number of that trinucleotide occurring in 5000 randomly generated sequences with the same length and dinucleotide composition as the sequence of interest.

Comparison between Observed and Expected Frequencies of Dinucleotides and Trinucleotides. Observed/expected (O/E) ratios are a convenient method of comparison between observed and expected frequencies of nucleotide patterns. For statistical analysis, however, the disadvantage of O/E is that at low expected values (i.e., when %G+C is low and/or the length of the sequence is small) the variance of the statistic can be relatively high. To minimize this effect, instead of calculating the mean of the individual O/E values for a set of sequences, we calculated the ratio of the sum of observed frequencies to the sum of expected frequencies, that is, $[\Sigma\ (O_i/n_i)]/[\Sigma\ (E_i/n_i)]$, where O_i and E_i are the number of observed and expected motifs, respectively, and n_i is the length of the sequence (bp), with the summation carried out over N sequences. For convenience we will call this statistic the average O/E CpG of N sequences. The variance (V) of this statistic was calculated as

$$V\left(\frac{x}{y}\right) = N\left(\frac{x}{y}\right)^2 \left\{ \frac{V(O_i/n_i)}{x^2} + \frac{V(E_i/n_i)}{y^2} - \frac{2Cov[(O_i/n_i),(E_i/n_i)]}{xy} \right\}$$

where x and y are $\Sigma\ (O_i/n_i)$ and $\Sigma\ (E_i/n_i)$, respectively, and Cov is the covariance. Standard errors (S) were calculated as $\sqrt{V}$.

When testing whether, for a given set of sequences, the O/n values come from the same population as the E/n values, we have used the Wilcoxon nonparametric test for paired observations. The (O − E)/S statistic, which is testable by parametric methods and was developed by Boudraa and Perrin (1987) for a similar purpose, was not used here because it did not yield normal distributions with our data.

Calculations of Correlation Coefficients. Values for O/E were used in calculations of correlation coefficients between levels of CpG dinucleotides in coding versus noncoding regions and levels of CpG dinucleotides versus G+C content. The O/E statistic allows direct comparison of our results with previously published analyses of the extent of CpG depletion in vertebrate genes. We considered using the statistic (O/n) − (E/n) because of the high variance of O/E at low E values. However, if CpG is depleted, the statistic (O/n) − (E/n) may be inappropriate because it may introduce a negative bias in the relationship with G+C. When CpG is depleted, (O/n) − (E/n) will have a negative value. In a case where O/E CpG increases as %G+C increases, as we and others have found for vertebrate sequences (Adams and Eason 1984; Gardiner-Garden and Frommer 1987), then at a given value of n, as G+C increases, the (O/n) − (E/n) values will become less negative as a result of the relationship between O/E and G+C but will also tend to become more negative as a result of the larger numbers of both observed and expected dinucleotides. For example, where a direct positive linear relationship exists between O/E CpG and %G+C, the resulting relationship between (O/n) − (E/n) and %G+C has a negative slope at low G+C content and a positive slope at high G+C content.

Values for (O/n) − (E/n) were used in all analyses of the relationship between levels of CpG dinucleotides or CpNpG trinucleotides and their respective deamination or transition products because, if the depletion of a particular motif is due to deamination, the relationships should be linear. The dependence of (O/n) − (E/n) on G+C content and n does not cause problems in this case because the (O/n) − (E/n) values under comparison at each point are calculated from the identical DNA sequence.

Calculation of Moving Average O/E and %G+C Values. Moving average O/E dinucleotide and %G+C values were calculated for 100 nucleotide windows moving across individual sequences in one-nucleotide increments, as described previously (Gardiner-Garden and Frommer 1987). Data for individual gene sequences and all computer programs are available upon request.

Results

CpG-Depletion in Angiosperm and Vertebrate Genes

Average values of G+C content and O/E CpG and GpC were calculated for the angiosperm gene sequences listed in Methods. The genes analyzed contained on average 0.75 times the frequency of CpG dinucleotides expected from the base composition, a ratio very similar to that found in whole angio-

Table 1. Observed/expected (O/E) values for CpG and GpC, G+C content (%), and correlation between O/E and %G+C for protein- and RNA-coding genes of monocots and dicots

Genes	No. genes	No. bases	O/E[a]		% G+C	R^{2b}	
			CpG	GpC		CpG	GpC
Protein							
Coding regions							
Monocots							
All genes[c]	27	30,151	0.86 (0.05)[d]	1.12 (0.03)[d]	59 (2)[d]	0.72*	0.06
One gene[e]	15	19,817	0.79 (0.04)	1.07 (0.04)	57 (2)	0.61*	0.00
Dicots							
All genes	72	60,818	0.56 (0.03)	0.96 (0.02)	48 (1)	0.26*	0.02
One gene	35	33,249	0.60 (0.05)	0.95 (0.05)	48 (1)	0.39*	0.00
Noncoding regions[f]							
Monocots							
All genes	26	50,855	0.73 (0.04)	1.14 (0.03)	42 (1)	0.32*	0.06
One gene	15	29,417	0.71 (0.04)	1.19 (0.04)	41 (2)	0.04	0.01
Dicots							
All genes	71	107,448	0.59 (0.03)	0.98 (0.03)	31 (1)	0.02	0.00
One gene	35	65,631	0.61 (0.05)	0.99 (0.04)	31 (1)	0.00	0.02
RNA							
Monocots	4	7995	1.12 (0.05)	1.05 (0.03)	59 (4)	1.00*	0.48
Dicots	8	7531	1.02 (0.06)	1.00 (0.05)	49 (3)	0.66*	0.00

[a] $O/E = \Sigma (O_i/n_i)/\Sigma (E_i/n_i)$

[b] Correlation between O/E and %G+C in individual genes

[c] Calculations using the complete set of sequences

[d] $\bar{x}$ (SD)

[e] Calculations using only one gene of each type, selected at random

[f] richis3 and tomcbpa were not included because <250 bp of noncoding sequence was available for these genes

* $P < 0.01$

sperm genomes (Swartz et al. 1962; Russell et al. 1971) and somewhat higher than that found in the more limited set of angiosperm sequences analyzed by McClelland (1983). Both coding and noncoding regions of the protein-coding genes were CpG depleted to a similar degree, with an average O/E CpG ratio of 0.66. The RNA-coding genes were generally not CpG depleted, with the exception of the two tRNA gene sequences and the soybean 18S and yellow lupin 5S rRNA sequences, which had O/E CpG values of 0.8–0.9.

The same calculations were carried out separately on the sequences derived from monocot and dicot species (Table 1). Dicot and monocot genomes, as represented by cauliflower and wheat respectively, share the same overall O/E CpG as determined by nearest neighbor analysis (Swartz et al. 1962; Russell et al. 1971). The dicot genes we analyzed had a lower average O/E CpG than did the set of monocot genes, in both protein-coding and noncoding regions (Table 1). Because few clearly equivalent genes have been sequenced in both classes, however, we cannot be certain that this is a generalized difference. Our study included only seven genes for which at least one clearly equivalent sequence was available in each of the monocot and dicot classes: actin, alcohol dehydrogenase, chlorophyll a/b binding protein (photosystem II), histones H3 and H4, 17–18S rRNA, and 25–26S rRNA. O/E CpG values of the seven equivalent monocot and dicot genes were compared for the coding regions and, where available, for equal lengths of DNA either upstream of the initiation codon or downstream of the termination codon. Introns were not compared because they were generally of different lengths. The lower O/E CpG value in dicots was apparent only in the coding regions of the five protein-coding genes; all five dicot protein-coding sequences had a lower O/E CpG value than did the equivalent monocot sequence or sequences (data not shown).

The nonmethylatable GpC dinucleotide was present at the expected frequency, as measured by nearest neighbor analysis, in both cauliflower and wheat total genomes (Swartz et al. 1962; Russell et al. 1971). In the dicot sequences analyzed here, GpC was present at the expected frequency. In the less numerous monocot sequences, however, GpC was significantly enriched, both in protein-coding and noncoding regions (1% significance level, Wilcoxon nonparametric test for paired observations).

In a previous analysis of vertebrate gene sequences, we subdivided protein-coding sequences into those that were CpG depleted across their entire length and those that contained CpG islands (Gar-

diner-Garden and Frommer 1987). The vertebrate CpG-depleted genes had a mean O/E CpG of 0.22 (range 0–0.44), whereas those sequences containing CpG islands had a mean O/E CpG of 0.53 (range 0.25–0.95, the value depending on both the strength of the CpG island and the relative lengths of CpG island and total sequence). Therefore, the protein-coding angiosperm gene sequences analyzed in the present study had a higher average O/E CpG than did vertebrate gene sequences, including a subset of vertebrate sequences that contained CpG islands. Nevertheless, a small proportion of angiosperm genes was composed of DNA, or included stretches of DNA more than 1 kb in length, that was CpG depleted to an extent approaching that of vertebrate CpG-depleted genes. Only one of the angiosperm genes analyzed, petunia chlorophyll a/b binding protein Cab37, had a value for O/E CpG actually as low as that of the average vertebrate CpG-depleted gene (data not shown).

Because the sequences we analyzed contained several representatives of a number of genes, each from a different species, our results may have been biased by some genes with unusual sequence characteristics. We therefore calculated the average values of O/E CpG, O/E GpC, and %G+C using only one representative of each type of gene, selected at random, as listed in Methods. The results were almost identical to the results for the total set of genes (Table 1).

Correlation between O/E CpG and %G+C

Vertebrate sequences show a positive correlation between %G+C and O/E CpG values (Adams and Eason 1984; Gardiner-Garden and Frommer 1987). We found a similar relationship in the total set of angiosperm sequences, with a strong positive correlation between G+C content and O/E CpG and no correlation between G+C content and O/E GpC. The strength of the relationship between %G+C and O/E CpG, as measured by the square of the correlation coefficient, differed when the total sequences were either separated into monocot and dicot classes or subdivided into coding and noncoding regions. In general, the strength of the relationship between G+C content and O/E CpG appeared to increase with increasing G+C content of the subset of sequences under analysis (Table 1), as detailed below.

Protein-coding and RNA-coding regions were generally G+C rich relative to noncoding regions; protein-coding regions in monocots had a higher G+C content than did those in dicots. A strong positive correlation was found in the protein-coding regions and the RNA-coding gene sequences in both monocot and dicot classes. The relationship was stronger in monocot coding regions than in dicot coding regions.

Both monocot and dicot noncoding regions were A+T rich; monocot noncoding regions had a higher G+C content than did those in dicots. A significant correlation between G+C content and O/E CpG was apparent in the monocot noncoding sequences, whereas the relationship was absent in the dicot noncoding sequences. This absence of a relationship was evident even though the range of G+C values was almost as large and the range of O/E values was even larger in dicot noncoding regions than in monocot noncoding regions and despite the fact that many more dicot than monocot sequences were included in the analysis.

By separating vertebrate sequences into CpG-depleted DNA and CpG islands, which have a high O/E CpG and are G+C rich relative to CpG-depleted DNA, we previously showed that the relationship between O/E CpG and %G+C is mainly due to varying lengths of CpG island DNA in the vertebrate total sequences (Gardiner-Garden and Frommer 1987). We have investigated the location and G+C content of candidate CpG islands in coding and noncoding regions of angiosperm gene sequences (Gardiner-Garden and Frommer 1992). The differences in the relationships between O/E CpG and %G+C may relate to differences in the G+C content of candidate CpG islands in monocots and dicots.

Correlation between O/E CpG Values in Protein-Coding and Noncoding Segments of the Same Gene

We analyzed protein-coding genes for any relationship between the corresponding coding and noncoding regions of individual genes with respect to O/E CpG, O/E GpC, or %G+C (Table 2). We found a positive correlation between O/E CpG in coding and noncoding regions, whereas no significant relationship was apparent for O/E GpC. The linear relationship between coding and noncoding O/E CpG values was strong in the monocot sequences and quite weak (though significant at the 1% level) in dicot sequences.

Although the average G+C content differed considerably between coding and noncoding regions (Table 1), there was a strong positive correlation between the %G+C values for the corresponding coding and noncoding regions of individual genes. This relationship between %G+C in coding and noncoding regions was evident in both monocot and dicot genes (Table 2), in accordance with the findings of Matassi et al. (1989), who analyzed %G+C values in exons and introns versus combined flanking regions. The results were similar when only one rep-

resentative of each type of gene was included in the analysis.

Arabidopsis thaliana Gene Sequences

The small genome of *Arabidopsis* (7×10^7 bp) is notable for its relative lack of repetitive sequences (Pruitt and Meyerowitz 1986) and its relatively low level of cytosine methylation (6.3% of total cytosine) (Leutwiler et al. 1984). However, *Arabidopsis* gene sequences do share some of the features of gene sequences from larger, more highly methylated genomes, such as wheat (5×10^9 bp) (Flavell et al. 1974). The *Arabidopsis* gene sequences analyzed were CpG depleted; the 20 sequences, which included one sequence only from each gene family, had an average O/E CpG of 0.74. For the individual gene sequences, the values for (O/n) − (E/n) CpG were inversely proportional to the values for (O/n) − (E/n) TpG+CpA (Fig. 1). Such a relationship could arise both from mutations of ^{m}CpG and CpG to TpG and from reverse mutations of TpG to CpG. However, almost all values lie in the quadrant of negative (O/n) − (E/n) CpG and positive (O/n) − (E/n) TpG+CpA, suggesting that the relationship results almost entirely from 5^{m}CpG and/or CpG to TpG mutation.

Table 2. Correlation (R^2) between protein-coding and noncoding regions of individual monocot and dicot genes, calculated for observed/expected (O/E) values of CpG and GpC, and percentage of G+C content

Genes	No. genes	R^2 O/E CpG	R^2 O/E GpC	R^2 %G+C
Monocots				
All genes	26	0.50*	0.02	0.36*
One gene	15	0.46*	0.08	0.26
Dicots				
All genes	71	0.16*	0.00	0.31*
One gene	35	0.20*	0.00	0.40*

[a] Calculations using the complete set of sequences
[b] Calculations using only one gene of each type, selected at random
* $P < 0.01$

Depletion of CpA/TpG and GpA/TpC

CpNpG trinucleotides are important methylation sites in angiosperm genomes. If extensive CpNpG methylation is also present in angiosperm genes, we might expect evidence of CpNpG depletion in angiosperm gene sequences. We have calculated the average O/E ratio for CpNpG and GpNpC trinucleotides and for the products of 5^{m}C deamination or C transition mutations within these trinucleotides (Table 3) and have studied the relationship between

Table 3. Average observed/expected[a] values for the trinucleotides CpNpG and GpNpC and their respective ^{m}C and/or C to T mutation products in protein- and RNA-coding genes of monocots and dicots

Genes	CpNpG: CpApG	TpApG + CpTpA	CpTpG	TpTpG + CpApA	CpCpG	TpCpG + CpGpA	CpGpG	TpGpG + CpCpA
Protein								
Coding regions								
Monocots								
All genes[b]	0.91	0.99	0.93	1.36*	1.18*	0.99	0.96	1.12*
One gene[c]	0.98	0.94	0.93	1.32*	1.11*	1.03	0.97	1.10*
Dicots								
All genes	0.89*	0.95	0.77*	1.12*	1.19*	1.02	0.94	1.24*
One gene	0.90	0.93	0.75*	1.12*	1.11	0.99	0.85	1.26*
Noncoding regions								
Monocots								
All genes	0.93	1.03	0.93	0.98	1.02	1.00	1.14	1.13*
One gene	0.97	1.03	0.94	0.96	1.08	1.01	1.09	1.12
Dicots								
All genes	0.85*	0.97	0.73*	0.99	1.00	0.94	1.06	1.07*
One gene	0.83*	0.96	0.74*	0.98	0.99	0.91	1.07	1.05
RNA								
Monocots	0.96	1.03	1.02	1.06	1.05	0.96	1.10	0.97
Dicots	0.91	0.93	0.80	0.89	0.92	1.03	1.13	1.00

[a] $\Sigma (O_i/n_i)/\Sigma (E_i/n_i)$
[b] Calculations using the complete set of sequences
[c] Calculations using only one gene of each type, selected at random
* Observed (O/n) and expected (E/n) values from populations with different means, $P < 0.01$

each CpNpG and GpNpC trinucleotide and its 5^mC or C to T mutation products for individual gene sequences (Table 4). Several of the trinucleotides were not present at the expected frequency in the gene sequences analyzed, but the results could be accounted for by transition mutations in general, without introducing 5^mC to T mutations.

CpApG and its reverse complement, CpTpG, were significantly depleted (1% level) in both coding and noncoding regions of dicot protein-coding genes. The monocot protein-coding genes also had a small deficiency of CpApG and CpTpG, although the effect was not generally significant at the 1% level. The corresponding mutation products were generally not elevated over the expected levels in either monocots or dicots. The average O/E ratios for the nonmethylatable sites, GpApC and its reverse complement GpTpC, and their respective C transition mutation products paralleled the pattern observed for CpApG and CpTpG, except that GpApC was not significantly depleted in dicot noncoding regions.

For CpTpG in coding regions, the one case where a CpNpG trinucleotide was generally depleted and its respective 5^mC or C to T mutation products were elevated, there was no relationship between the (O/n) − (E/n) values for the trinucleotide and its mutation products in individual gene sequences (Table 4), suggesting that 5^mC deamination mutation was not the cause of the differential in O/E values.

Significant negative correlations were observed for CpApG and CpTpG versus their respective 5^mC or C to T mutation products in noncoding regions of individual gene sequences. These relationships could result from deamination mutations, with some other phenomenon tending to reduce the levels of the mutation products, which do not appear to be elevated (Table 3). However, the relationships were not strong (Table 4). In noncoding regions, a slight

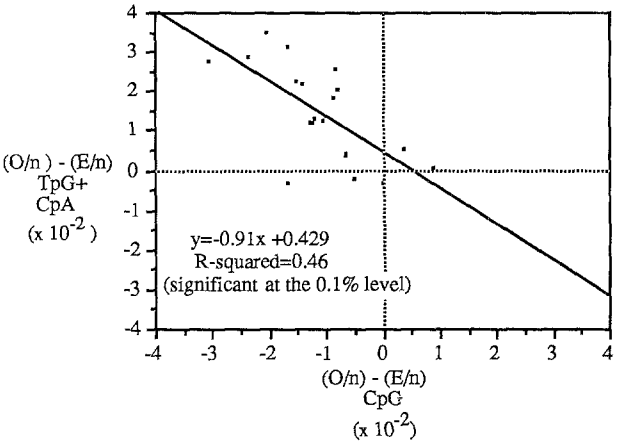

Fig. 1. Linear regression between values of (O/n) − (E/n) CpG and (O/n) − (E/n) TpG+CpA for *Arabidopsis thaliana* gene sequences. Each point represents an individual *A. thaliana* gene sequence. Only one example of each type of gene was included (see Methods).

Table 3. Extended

GpNpC							
GpApC	GpApT + ApTpC	GpTpC	GpTpT + ApApC	GpCpC	GpCpT + ApGpC	GpGpC	GpGpT + ApCpC
1.06	1.04	0.90*	1.27	1.03	0.99	1.15*	1.18*
1.03	1.01	0.94	1.28	1.01	1.03	1.12	1.13*
0.82*	0.99	0.73*	1.09*	1.05	1.13*	0.97	1.18*
0.80*	0.98	0.76*	1.11	1.07	1.11*	1.00	1.17*
0.87*	1.02	0.86*	1.04	1.00	1.02	1.04	0.98
0.86	1.00	0.89	1.06	1.00	1.05	1.03	0.94
1.00	0.95*	0.83*	0.97	1.04	1.00	0.98	1.00
1.00	0.91*	0.85*	0.95	1.09	0.95	1.01	1.00
0.92	1.01	1.03	1.10	1.09	0.90	1.18	1.02
1.00	0.97	0.88	0.85	0.94	0.96	0.96	1.00

Table 4. Correlation (R^2) for observed minus expected[a] values for CpNpG and GpNpC trinucleotides versus their respective mutation or transition products in protein- and RNA-coding genes of monocots and dicots

	CpNpG[b]				GpNpC[c]			
Genes	CpApG vs TpApG + CpTpA	CpTpG vs TpTpG + CpApA	CpCpG vs TpCpG + CpGpA	CpGpG vs TpGpG + CpCpA	GpApC vs GpApT + ApTpC	GpTpC vs GpTpT + ApApC	GpCpC vs GpCpT + ApGpC	GpGpC vs GpGpT + ApCpC
Protein								
Coding regions								
Monocots								
All genes[d]	0.06	0.04	0.27*	0.05	0.10	0.05	0.04[f]	0.01
One gene[e]	0.26	0.19	0.18	0.01	0.12	0.09	0.12[f]	0.07
Dicots								
All genes	0.03	0.09	0.13*	0.01[f]	0.00	0.06	0.00	0.03
One gene	0.03	0.05	0.04	0.00	0.00	0.04	0.02[f]	0.02
Noncoding regions								
Monocots								
All genes	0.23	0.36*	0.05	0.08[f]	0.02	0.10	0.00	0.03
One gene	0.45*	0.26	0.00	0.00	0.06	0.00	0.03[f]	0.04
Dicots								
All genes	0.24*	0.07	0.09*	0.02[f]	0.06	0.12*	0.05	0.04
One gene	0.30*	0.01[f]	0.10	0.00	0.16	0.13	0.05	0.06
RNA								
Monocots + dicots	0.46	0.04[f]	0.18	0.28	0.00	0.01[f]	0.16	0.18

[a] (O/n) − (E/n)
[b] CpNpG vs 5^mC and C to T mutation products
[c] GpNpC vs C to T transition products
[d] Calculations using the complete set of sequences
[e] Calculations using only one gene of each type, selected at random
[f] Positive slope
* $P < 0.01$

but statistically significant relationship was also found between the nonmethylatable sites, GpApC and GpTpC, and their respective cytosine transition products (Table 4).

Neither CpCpG nor its reverse complement, CpGpG, were depleted in protein- or RNA-coding gene sequences. In fact, CpCpG trinucleotides were significantly enriched in protein-coding regions of both the monocot and dicot subsets, whereas the deamination/transition mutation products of CpCpG were present at the expected frequency. The equivalent nonmethylatable trinucleotide, GpCpC, was not elevated. CpGpG was present at the expected frequency, and its deamination/transition mutation products were somewhat elevated, particularly in coding regions of protein-coding genes. The equivalent nonmethylatable site, GpGpC, and its transition mutation products showed similar results.

Significant (1% level) negative correlations between (O/n) − (E/n) values for a trinucleotide and its 5^mC or C to T mutation products were observed for CpCpG in protein-coding regions; the overall O/E values (Table 3) suggest a reverse relationship of TpCpG+CpGpA tending to mutate to CpCpG.

Discussion

CpG Depletion

We have shown that CpG dinucleotides in angiosperm gene sequences are depleted to much the same extent as the total genome. The dicot and monocot genes analyzed had an average O/E CpG of 0.68 and 0.79, respectively, calculated using one representative of each type of gene. These sequences comprise RNA-coding genes that contain CpG at the expected frequency and protein-coding genes that contain CpG at a slightly lower frequency than the genome as a whole. The protein-coding genes display a range of O/E CpG levels, with few sequences approaching the low O/E CpG levels of CpG-depleted vertebrate sequences.

The *Arabidopsis* genome has only about one-fourth the methylation level of most angiosperm

genomes analyzed (Leutwiler et al. 1984), a level similar to that found in most vertebrate genomes. Whether this is due to less CpG methylation, less CpNpG methylation, or lower levels of both CpG and CpNpG methylation is not known. A high level of methylation has been detected in some *Arabidopsis* satellite sequences (Martinez-Zapater et al. 1986) and some rDNA sequences (Pruitt and Meyerowitz 1986). However, few methylated HpaII/MspI sites were detected in eight unique sequences isolated at random from a genomic library (Pruitt and Meyerowitz 1986). Hence, *Arabidopsis* genes are considered hypomethylated, and we expected that *Arabidopsis* gene sequences would have a significantly higher O/E CpG value than would the total set of angiosperm sequences. CpG depletion of the *Arabidopsis* gene sequences included in this study ($\bar{x}$ O/E CpG = 0.74), however, was similar to that of gene sequences from other angiosperm species. Furthermore, the CpG depletion of *Arabidopsis* genes correlated with an elevation of the 5^{m}CpG deamination products TpG and CpA, suggesting that methylation of gene sequences may occur in *Arabidopsis* to a similar extent as in other angiosperm species. The apparent lack of methylation in randomly selected *Arabidopsis* unique sequence clones may result from selective cloning of hypomethylated sequences, at a time when only methylation-restricting bacterial host strains were available (Woodcock et al. 1989; Graham et al. 1990). Therefore, the small *Arabidopsis* genome could be a convenient dicot genome model for methylation studies.

CpNpG Depletion

We compared the levels of CpNpG trinucleotides to the level expected from the dinucleotide composition. To calculate expected levels of CpNpG trinucleotides, we used the observed dinucleotide counts for each sequence under consideration, so that O/E CpNpG values did not reflect elevated levels or deficiencies of component dinucleotides, such as CpG, TpG, CpA, and TpA. We found that the error in values generated by approximate formulae using observed dinucleotide counts, such as E (CpApG) = (CpA × ApG)/A, although low for long sequences, was considerable for short sequences such as some introns and exons. Expected trinucleotide frequency values were therefore calculated from randomly generated sequences with the same length and dinucleotide composition as the sequence in question. An exact formula for calculating the expected frequency of nucleotide patterns using dinucleotide counts has recently been derived (Cowan, in press). Ideally, although perhaps not realistically, when analyzing motifs more than three bases long, even more complicated models that take into account observed trinucleotide frequencies should be considered; several trinucleotides did not occur at the expected frequency in angiosperm sequences.

Significant depletion of CpApG and CpTpG trinucleotides occurred in the angiosperm sequences. The nonmethylatable trinucleotides GpApC and GpTpC were also depleted to an equivalent degree, and the depletion of CpApG and CpTpG was not generally accompanied by an elevation of the respective 5^{m}C deamination products. Therefore, the deficiency of CpApG and CpTpG trinucleotides probably results from a mechanism that does not involve methylation. Because the CpApG and CpTpG depletion occurred in noncoding DNA and coding DNA, it was not due to codon usage. Significant enrichment of CpCpG trinucleotides occurred in protein-coding regions.

Do Angiosperm and Vertebrate Genomes Differ in the Efficiency and Accuracy of T/G Mismatch Repair?

Much of the support for the deamination theory of CpG depletion comes from the observed positive correlation between the level of methylation and the extent of CpG depletion in various animal species (Bird 1980). However, we found no such relationship when published data for angiosperm and vertebrate total genomes were compared. Where examined, the level of methylation at CpG dinucleotides is at least as high in angiosperm genomes as in vertebrate genomes (Shapiro 1976), yet angiosperms have a frequency of CpG dinucleotides much closer to the expected value than do vertebrates (Swartz et al. 1962; Russell et al. 1971). Nevertheless, given the known high mutation rate of 5^{m}C to T, and the elevation of the products of 5^{m}CpG deamination (TpG and CpA) in angiosperm genomic DNA, the deamination theory is probably correct. A stable equilibrium level of CpG can be reached, with O/E CpG values equivalent to those found in CpG-depleted vertebrate sequences, using a mutation model with a relatively higher mutation rate for 5^{m}C to T than for other transition or transversion mutations (Bulmer 1986; Sved and Bird 1990). Lack of correlation between levels of methylation and CpG depletion in vertebrates and angiosperms could therefore result if the two taxa have different rates of fixation of 5^{m}C to T mutations relative to other mutations. In at least some vertebrate cells, T/G mismatches are repaired in favor of C/G at an efficiency of 90% (Brown and Jiricny 1987). A near perfect T/G to C/G mismatch repair in angiosperms

could account for the observed difference in O/E CpG in angiosperm and vertebrate whole genome data.

Do Vertebrates and Angiosperms Differ in the DNA Methylation Levels of Cells That Ultimately Contribute to the Germline?

Only deamination events occurring in the germline, that is, in the lineage of cells that will ultimately form germ cells, can become fixed (Cooper and Gerber-Huber 1985). The germline differs considerably between vertebrates and angiosperms. The vertebrate germline is delineated very early in development (Ginsburg et al. 1990). The expected frequency of CpG in a vertebrate DNA region can be maintained, even in the absence of direct selection against $5^{m}C$ to TpG mutation, if the region remains in an unmethylated state both in the embryo before the germline is set aside and in the germline cells throughout development. In angiosperms, cells throughout the plant's growth cycle can form part of an ill-defined germline. Hence, the maintenance of a particular CpG or CpNpG site in an angiosperm sequence may mean that the site remains unmethylated at many stages, including the apical initial cells within the embryo, the shoot apical meristem tissue (both at apex and axil positions), the floral apical meristem, and ultimately the germ cells. Shoots that eventually form floral parts may occasionally develop from adventitious buds, possibly requiring dedifferentiation of nonspecialized cells, such as cambium or parenchyma cells, to form apical or axillary meristematic tissue. In plant tissue that forms adventitious buds, the proportion of nonspecialized cells capable of contributing to the new apical meristem and the relative contribution of DNA mutations in such cells to the evolutionary process are not known.

The methylation levels reported for both angiosperm and vertebrate genomes were not measured in purely germline tissues, and general tissue and germline cells may differ with respect to DNA methylation levels. Hence, we propose another possible explanation for the lack of correlation between the level of methylation and the extent of depletion of CpG in angiosperm and vertebrate genomes—that the proportion of methylated CpG relative to total CpG in the cell lineages that normally contribute to angiosperm germ cells is less than the proportion of methylated CpG in vertebrate germline cells.

Methylation and CpG/CpNpG Frequency in Angiosperm Genes

Although the level of CpG and CpNpG methylation in whole genomes has been clearly established, the exact level in gene sequences is unknown. Cytosine methylation clearly occurs in the vicinity of some genes in angiosperm general tissues. Studies of several tissue-specific genes in angiosperms have found that the genes are methylated at an ill-defined proportion of CpG and CpNpG sites in nonexpressing cells and tend to become demethylated at certain sites in expressing cells (Bianchi and Viotti 1988; Ngernprasirtsiri et al. 1989; Riggs and Chrispeels 1990). The distribution of methylated and unmethylated sites relative to the transcription unit is not clear because studies where the methylation status of mapped CpG and CpNpG sites has been determined are rare (Nick et al. 1986; Riggs and Chrispeels 1990).

Angiosperm gene sequences analyzed here had a frequency of CpG dinucleotides much closer to the expected value than did vertebrate gene sequences. The differences may reflect a more efficient mismatch repair system or may indicate that the level of CpG methylation of angiosperm genes in those cells that normally contribute to the germ cells is lower than the level in vertebrate germlines.

There is no information about the extent of CpNpG depletion of the highly methylated total genomic DNA, so whether the high level of methylation is accompanied by extensive CpNpG depletion in the genome as a whole is unknown. Across a wide range of species and gene types, however, there is no indication of CpNpG depletion related to $5^{m}C$ deamination events within or in the immediate vicinity of genes. Our results could imply the existence of a "perfect" CpNpG mismatch repair system, even more accurate and efficient than for CpG (McClelland 1983). Alternatively, very little or no CpNpG methylation may occur within genes in the angiosperm germlines, and perhaps very little CpNpG methylation occurs within genes in any angiosperm tissues. If the lack of CpNpG methylation is confined to germlines, then the CpNpG methylation pattern will not be directly inherited from the previous generation but may be defined by or "copy" an inherited CpG methylation pattern. Thus, CpNpG hypomethylation may mark those cells within the meristem that normally contribute to angiosperm germlines.

Studies on rRNA-coding genes in *Lilium henryi* have been carried out using DNA from pollen mother cells at pachytene, and extensive CpG and CpNpG methylation was observed in rDNA repeats (von Kalm et al. 1986). Because genes encoding rRNA have, on average, close to the expected frequency of CpG dinucleotides, these *Lilium* data would seem to invalidate the hypothesis that CpG and CpNpG frequencies reflect methylation patterns in germ cell lineages. However, the sensitivity of the method for assaying methylation was such that the presence of

a small number of completely unmethylated rDNA repeat units in the germ cell DNA would not have been detected, and unmethylated genes may have been digested into small fragments that could not have been visualized on the gels used. Sequences of rDNA repeats are believed to be maintained constant (Flavell 1985) by genetic recombination and gene conversion events. A few unmethylated rRNA genes could potentially form the template sequences for gene conversion events, thus maintaining the high CpG and CpNpG frequency of the total set of rDNA repeats over evolutionary time.

This analysis defines methylation studies required to establish the mechanisms by which the observed CpG and CpNpG frequencies are maintained in angiosperm genes. Studies of the total level of genomic DNA methylation of CpG and CpNpG, by nearest neighbor analysis, in unequivocal germline cells such as premeiotic or meiotic pollen cells are required to establish whether the high levels of CpG and CpNpG methylation characteristic of angiosperm tissues are also present in angiosperm germ cells. Whole embryos (such as wheat germ) and pollen grains would not be suitable for these studies because only a few cells in the embryo are apical initials, and whole pollen contains many nongermline cells. The extent of CpG and particularly of CpNpG methylation within and around genes in differentiated tissues and in germ cells should be established by restriction analysis of numerous precisely mapped sites or ideally by genomic sequencing and related to the extent of CpG depletion and lack of CpNpG depletion of the sequence under analysis. The possibility of polymerase chain reaction amplification of restriction digests and of genomic sequencing protocols that can be used on small amounts of DNA should increase the feasibility of future methylation studies on germ cells.

Acknowledgments. We are extremely grateful to Simon Worthington, who wrote many programs that allowed us to screen sequences and plot data, and to Carolyn Bucholtz, who provided computer systems expertise and programs. Computer facilities were provided by the CSIRO Division of Biomolecular Engineering. We extend our thanks to Candy Briggs for helpful discussions and to Robin Holliday, Peter Molloy, and David Tremethick for critical reading of the manuscript.

References

Adams RLP, Eason R (1984) Increased G+C content of DNA stabilises methyl CpG dinucleotides. Nucleic Acids Res 12: 5869–5877

Bianchi MW, Viotti A (1988) DNA methylation and tissue-specific transcription of the storage protein genes of maize. Plant Mol Biol 11:203–214

Bird A, Taggart M, Frommer M, Miller OJ, Macleod D (1985) A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. Cell 40:91–99

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8:1499–1504

Bird AP (1986) CpG-rich islands and the function of DNA methylation. Nature 321:209–213

Bonen L, Huh TY, Gray MW (1980) Can partial methylation explain the complex fragment patterns observed when plant mitochondrial DNA is cleaved with restriction endonucleases. FEBS Lett 111:340–346

Boudraa M, Perrin P (1987) CpG and TpA frequencies in the plant system. Nucleic Acids Res 15:5729–5737

Brown TC, Jiricny J (1987) A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. Cell 50:945–950

Bulmer M (1986) Neighboring base effects on substitution rates in pseudogenes. Mol Biol Evol 3:322–329

Cedar H, Razin A (1990) DNA methylation and development. Biochim Biophys Acta 1049:1–8

Cooper DN, Gerber-Huber S (1985) DNA methylation and CpG suppression. Cell Differ 17:199–205

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in *Escherichia coli.* Nature 274:775–780

Cowan R (in press) J Appl Probab

Flavell RB (1985) Repeated sequences and genome change. In: Hohn B, Dennis ES (eds) Genetic flux in plants. Springer-Verlag, New York, pp 129–156

Flavell RB, Bennett MD, Smith JB, Smith DB (1974) Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem Genet 12:257–269

Gardiner-Garden M, Frommer M (1987) CpG islands in vertebrate genomes. J Mol Biol 196:261–282

Gardiner-Garden M, Frommer M (1992) Significant CpG-rich regions in angiosperm genes. J Mol Evol 34:231–245

Ginsburg M, Snow MHL, McLaren A (1990) Primordial germ cells in the mouse embryo during gastrulation. Development 110:521–528

Graham MW, Doherty JP, Woodcock DM (1990) Efficient construction of plant genomic libraries requires the use of *mcr*$^-$ host strains and packaging mixes. Plant Mol Biol Rep 8:18–27

Gruenbaum Y, Naveh-Many T, Cedar H, Razin A (1981a) Sequence specificity of methylation in higher plant DNA. Nature 292:860–862

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

Hepburn AG, Belanger FC, Mattheis JR (1987) DNA methylation in plants. Dev Genet 8:475–493

Holliday R, Monk M, Pugh JE (1990) DNA methylation and gene regulation. The Royal Society, London

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

Leutwiler LS, Hough-Evans BR, Meyerowitz EM (1984) The DNA of *Arabidopsis thaliana.* Mol Gen Genet 194:15–23

Martinez-Zapater JM, Estelle MA, Somerville CR (1986) A highly repeated DNA sequence in *Arabidopsis thaliana.* Mol Gen Genet 204:417–423

Matassi G, Montero LM, Salinas J, Bernardi G (1989) The isochore organization and the compositional distribution of homologous coding sequences in the nuclear genome of plants. Nucleic Acids Res 17:5273–5290

McClelland M (1983) The frequency and distribution of methylatable DNA sequences in leguminous plant protein coding genes. J Mol Evol 19:346–354

McClelland M, Ivarie R (1982) Asymmetrical distribution of CpG in an 'average' mammalian gene. Nucleic Acids Res 10: 7865–7877

Naveh-Many T, Cedar H (1982) Topographical distribution of

5-methylcytosine in animal and plant DNA. Mol Cell Biol 2: 758–762

Nelson O (1988) Plant transposable elements. Plenum, Madison WI

Ngernprasirtsiri J, Kobayashi H, Akazawa T (1989) Transcriptional regulation and DNA methylation of nuclear genes for photosynthesis in nongreen plant cells. Proc Natl Acad Sci USA 86:7919–7923

Nick H, Bowen B, Ferl RJ, Gilbert W (1986) Detection of cytosine methylation in the maize alcohol dehydrogenase gene by genomic sequencing. Nature 319:243–246

Nyce J, Liu L, Jones PA (1986) Variable effects of DNA-synthesis inhibitors upon DNA methylation in mammalian cells. Nucleic Acids Res 14:4353–4367

Pruitt RE, Meyerowitz EM (1986) Characterization of the genome of *Arabidopsis thaliana*. J Mol Biol 187:169–183

Riggs CD, Chrispeels MJ (1990) The expression of phytohemagglutinin genes in *Phaseolus vulgaris* is associated with organ-specific DNA methylation patterns. Plant Mol Biol 14:629–632

Russell GJ, Follett EAC, Subak-Sharpe JH (1971) The double-stranded DNA of cauliflower mosaic virus. J Gen Virol 11: 129–138

Russell GJ, Walker PMB, Elton RA, Subak-Sharpe JH (1976) Doublet frequency analysis of fractionated vertebrate nuclear DNA. J Mol Biol 108:1–23

Setlow P (1976) Nearest neighbor frequencies in deoxyribonucleic acids. In: Fasman GD (ed) Handbook of biochemistry and molecular biology. CRC Press, Cleveland OH, pp 312–318

Shapiro HS (1976) Distribution of purines and pyrimidines in deoxyribonucleic acids. In: Fasman GD (ed) Handbook of biochemistry and molecular biology. CRC Press, Cleveland OH, pp 241–275

Sinsheimer RL (1955) The action of pancreatic deoxyribonuclease: II. Isomeric dinucleotides. J Biol Chem 215:579–583

Sved J, Bird A (1990) The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. Proc Natl Acad Sci USA 87:4692–4696

Swartz MN, Trautner TA, Kornberg A (1962) Enzymatic synthesis of deoxyribonucleic acid: XI. Further studies on nearest neighbor base sequences in deoxyribonucleic acids. J Biol Chem 237:1961–1967

Tykocinski ML, Max EE (1984) CG dinucleotide clusters in MHC genes and in 5′ demethylated genes. Nucleic Acids Res 12:4385–4396

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

von Kalm L, Vize PD, Smyth DR (1986) An under-methylated region in the spacer of ribosomal RNA genes of *Lilium henryi*. Plant Mol Biol 6:33–39

Wagner I, Capesius I (1981) Determination of 5-methylcytosine from plant DNA by high-performance liquid chromatography. Biochim Biophys Acta 654:52–56

Woodcock DM, Crowther PJ, Diver WP (1987) The majority of methylated deoxycytidines in human DNA are not in the CpG dinucleotide. Biochem Biophys Res Commun 145:888–894

Woodcock DM, Crowther PJ, Doherty J, Jefferson S, DeCruz E, Noyer-Weidner M, Smith SS, Michael MZ, Graham MW (1989) Quantitative evaluation of *Escherichia coli* host strains for tolerance to cytosine methylation in plasmid and phage recombinants. Nucleic Acids Res 17:3469–3478

Received April 22, 1991/Revised and accepted September 24, 1991