The EMBO Journal vol.7 no.8 pp.2295–2299, 1988

Unmethylated CpG islands associated with genes in higher plant DNA

Francisco Antequera and Adrian P.Bird

MRC Clinical and Population Cytogenetics Unit, Western General Hospital, Crewe Road, Edinburgh EH4 2XU, UK

Present address: Research Institute for Molecular Pathology, Dr Bohr-Gasse 7, 1030 Vienna, Austria

Communicated by A.P.Bird

The genomes of many higher plant species are the most highly methylated among eukaryotes. We report here that in spite of their heavy methylation, genomic DNAs from four plant species contain a fraction that is very rich in non-methylated sites. The fraction was characterized in maize where it represents about 2.5% of the total nuclear genome. In order to establish the genomic origin of the fraction, three maize genes containing clustered CpG were tested for methylation and were found to be non-methylated in the CpG-rich regions. By contrast, tested CpGs were methylated in a gene whose sequence showed no clustering of CpG. These observations suggest that the CpG-rich fraction of plants is at least partially derived from non-methylated regions that are associated with genes. A similar phenomenon has been described in vertebrate genomes. We discuss the evolution of CpG islands in both groups of organisms, and their possible uses in mapping and gene isolation in plants.

Key words: CpG islands/DNA methylation/*Zea mays* DNA

Introduction

The nuclear DNA of higher plants is heavily methylated at cytosine residues (Shapiro, 1976). In some species the level of 5-methyl cytosine (m^5C) amounts to ~30% of total cytosines, distributed between the sequences m^5CpG and m^5CpXpG (Gruenbaum *et al.*, 1981). In animal genomes methylation appears to be confined to the sequence m^5CpG, and levels of DNA methylation are generally much lower than in plants, ranging from indetectable amounts in some insects to ~8% of total cytosine in vertebrates (Shapiro, 1976). Although vertebrate genomes are relatively highly methylated, they contain clusters of non-methylated CpG known as 'CpG islands', many of which are associated with genes (Bird, 1986; Bird *et al.*, 1985; Gardiner-Garden and Frommer, 1987). Most animal genomes, however, appear to be only fractionally methylated and lack discrete CpG islands (Bird, 1987).

As in animals, the level of modification at methylatable sequences in plants is <100%, and this raises the question of the relative location of methylated and non-methylated CpG and CpXpG in the genome. We were particularly interested to know if the CpG island phenomenon was present in plants, which are evolutionary distant from vertebrates, but which have high levels of genomic methylation. In this study we have used methyl-sensitive restriction endonucleases to look for genome-wide clustering of non-methylated sites. We report the finding of a DNA fraction in several plants that lacks methylation and contains closely spaced sites for CpG enzymes. Analysis of four maize genes shows that the fraction is derived, in part at least, from CpG clusters that are associated with genes.

Results

An HTF fraction in plant DNA

The methylated sequences in plant DNA are m^5CpG or m^5CpXpG, where X can be any nucleotide (Gruenbaum *et al.*, 1981). Thus when maize (*Zea mays*) DNA was digested with the methyl sensitive restriction endonucleases *Hpa*II (CCGG), *Hin*PI (GCGC) and *Eco*RII (CCA/TGG) very little cleavage was evident on ethidium bromide-stained agarose gels (Figure 1a). The DNA was more extensively hydrolized, however, by *Msp*I (CCGG), which is insensitive to m^5CpG, and *Bst*NI (CCA/TGG), which is insensitive to m^5CpXpG. Relatively weak cleavage by *Msp*I is probably due to methylation of the first C in its recognition sequence, since this residue is part of a CpXpG motif (Gruenbaum *et al.*, 1981). Similar results to those of Figure 1a were obtained with DNA from rye (*Secale cereale*), tobacco (*Nicotina tabacum*), and wheat (*Triticum aestivum*) (data not shown).

Staining with ethidium bromide visualizes the weight distribution of DNA fragments on an agarose gel, whereas end-labelling gives a number distribution. End-labelling has been used previously to detect a non-methylated *Hpa*II tiny fragments (HTF) fraction of vertebrate DNA (Cooper *et al.*, 1983). When the *Hpa*II and *Hin*PI fragments of maize DNA were end-labelled with ^{32}P, the autoradiograph showed a prominent low mol. wt fraction resembling the HTF fraction seen in vertebrate genomes (Figure 1b). More than half of the labelled fragments were smaller than 0.5 kb, while most of the remainder were too large to be resolved by the 1.2% agarose gel. This indicated that a fraction of maize DNA is very rich in non-methylated *Hpa*II and *Hin*PI sites. End-labelled digests of rye, tobacco and wheat DNA gave the same pattern of non-methylated fragments with *Hpa*II (Figure 1c) and with *Hin*PI (data not shown). The low mol. wt fraction of maize DNA was more accurately sized in 12% polyacrylamide gels. It ranged from 25 to 250 bp with an average size of 120 bp and represented ~2.5% of the total nuclear maize DNA (not shown). When these end-labelling experiments were repeated using *Eco*RII instead of *Hpa*II or *Hin*PI, the non-methylated fraction in maize DNA was much less prominent suggesting that non-methylated CpXpG sites are less clustered than unmethylated CpG sites (not shown).

CpG clusters at genes

Since the low mol. wt fragments of plant DNA closely resemble the HTF fraction found previously in vertebrate

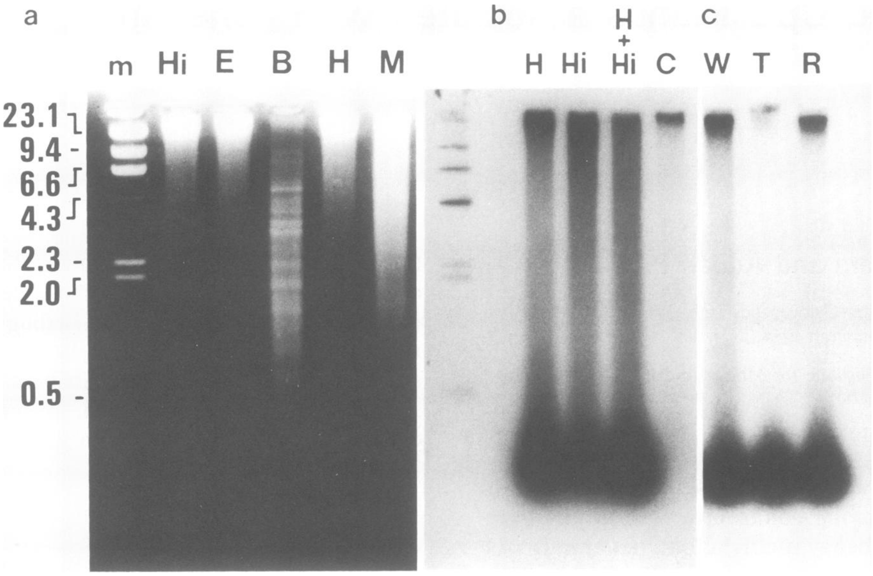

Fig. 1. Restriction pattern of plant DNA with methyl sensitive and insensitive endonucleases. (**a**) Ethidium bromide pattern of maize DNA digested with *Hin*PI (Hi), *Eco*RII (E), *Bst*NI (B), *Hpa*II (H) and *Msp*I (M). m, lambda DNA digested with *Hin*dIII. Numbers on the left indicate kilobase pairs (kb). (**b**) Maize DNA undigested (C) and digested with *Hpa*II (H), *Hin*PI (Hi) or *Hpa*II + *Hin*PI and end-labelled with ^{32}P before electrophoresis. Size markers as in (a). (**c**) End-labelling of *Hpa*II-digested DNA from wheat (W), tobacco (T) and rye (R).

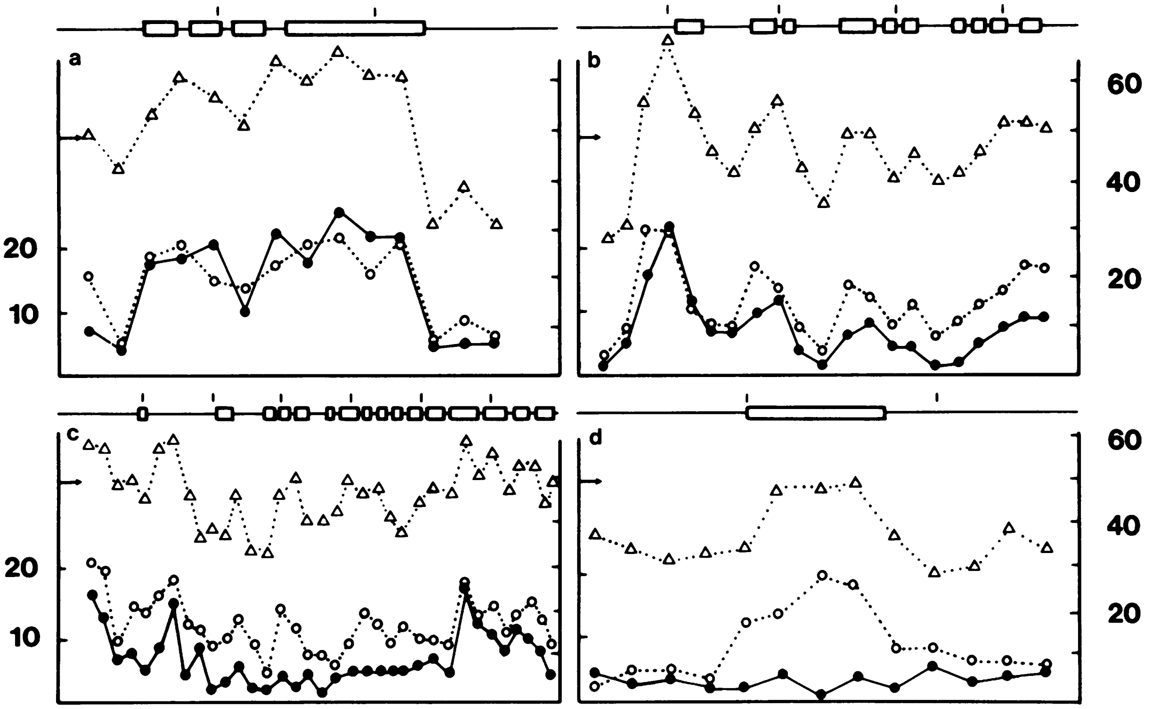

Fig. 2. CpG and GpC frequencies, and GC content across four maize gene sequences. (**a**) Anthocyanin 1 (*a1*); (**b**) Alcohol dehydrogenase-1 (*Adh1*); (**c**) sucrose synthase (*Sh1*); and (**d**) zein. The left hand ordinate refers to the number of CpGs (●) and GpCs (○) in 200 bp steps across the genes. Each gene is shown diagrammatically above each panel. Open boxes represent exons. The right-hand ordinate denotes the percentage of GC in the same 200 bp steps (△). The arrows on the left of each panel represent the average GC content of maize DNA (49%) (Shapiro, 1976). Vertical lines above the genes represent a scale in kb.

animals, we wished to know whether they, like the vertebrate fragments, were derived from CpG islands associated with genes. One of the characteristics of vertebrate CpG islands is that they show no CpG suppression, as CpG and GpC frequencies are roughly equal and are approximately as predicted by the base composition of the DNA. Bulk genomic DNA of both vertebrates and higher plants, however, contains significantly less CpG than would be expected from base composition (Setlow, 1976). The distinctively high CpG frequency of CpG islands can be used to identify CpG islands in regions of the genome whose sequence is known. In order to look for potential islands at plant genes, we examined four maize genes whose DNA sequences had been determined. Three of the genes indeed showed obvious clusters of CpG

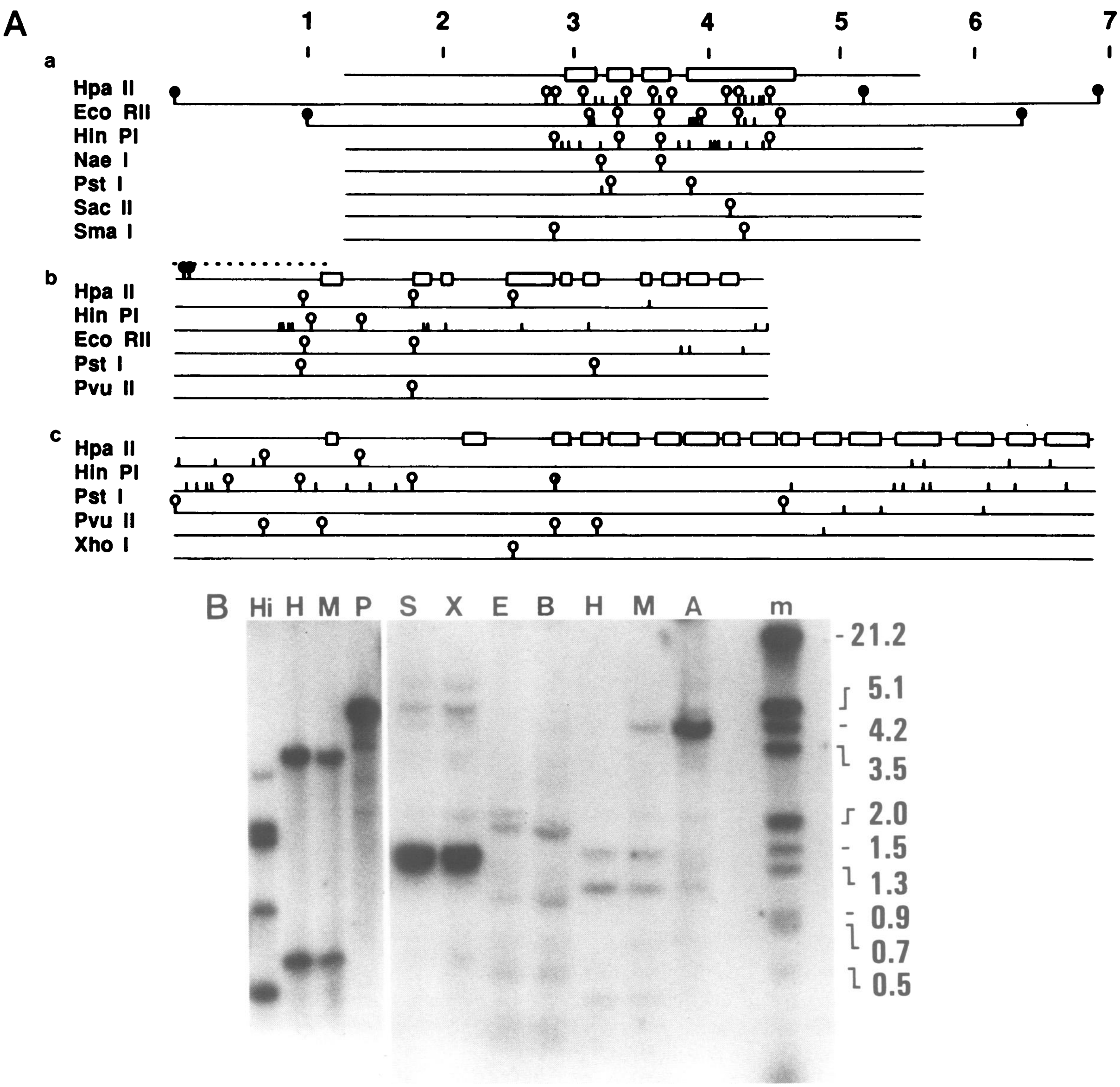

Fig. 3. Methylation of three maize genes containing clustered CpG sites. (**A**) Maps of methylated and unmethylated restriction sites for the *a1* (**a**); *Adh1* (**b**); and *Sh1* (**c**) genes. Methylated, unmethylated and partially methylated sites are represented by solid, open and partially solid circles, respectively. The same scale in kb on the top of the figure applies to the three genes. Boxes represent exons. The broken line in (b) indicates the region that was analysed by genomic sequencing (See Discussion) (Nick *et al.*, 1986). (**B**) Examples of data used to construct maps above. Right panel, maize DNA digested with *Hin*dIII + *Eco*RI (A) and with these enzymes plus *Sma*I (S), *Xma*I (X), *Eco*RII (E), *Bst*NI (B), *Hpa*II (H) and *Msp*I (M). After blotting the filter was hybridized with the 4.3 kb *Hin*dIII–*Eco*RI genomic fragment including the *a1* gene represented on the map above. The methyl sensitive and insensitive isoschizomer pairs *Sma*I–*Xma*I, *Eco*RII–*Bst*NI and *Hpa*II–*Msp*I show several identical bands indicating lack of methylation in the region corresponding to the probe. The 4.3 kb band in right panel line M is due to contamination from adjacent line A in this experiment due to a loading error. This band was not present when the blot was repeated. m, bacteriophage lambda DNA digested with *Hin*dIII + *Eco*RI. Numbers on the right indicate kb. These and other blots were used to generate the maps shown in (A). For example, to map flanking sites outside the *a1* cluster, the DNA was digested with *Msp*I, *Hpa*II, *Bst*NI and *Eco*RII. After blotting, the filter was probed with the 1.6 kb fragment spanning from the *Hin*dIII site to the left-most *Hin*PI site. After that, the filter was stripped and probed with the 1.2 kb fragment spanning from the right-most *Hin*PI site to the *Eco*RI site (see map). For the *Adh1* gene, the DNA was digested with the indicated enzymes. The *Adh1* cDNA was used as a probe. For the *Sh1* gene [map (c) and left autoradiograph], maize DNA digested with *Pst*I (P) or with *Pst*I + *Msp*I (M), *Hpa*II (H) or *Hin*PI (Hi) was probed with a 4.4 kb *Pst*I genomic fragment of the sucrose synthase gene (see map above). Sites in the sucrose synthase gene are equally cut by *Msp*I and *Hpa*II indicating lack of methylation at these sites. The region is also sensitive to *Hin*PI, although a reproducible 2.7 kb band shows the existence of a partially methylated site (see map). *Eco*RII, *Pst*I and *Pvu*II are inhibited by 5mCp(A/T)pG within their recognition sequence (Gruenbaum *et al.*, 1981). *Nae*I, *Sac*II, *Sma*I and *Xho*I are blocked by m^{5}CpG.

flanked by sequences in which CpG was suppressed (Figure 2). The anthocyanin 1 gene (*a1*) is involved in anthocyanin biosynthesis and it is thought to be expressed in every tissue of maize (Schwarz-Sommer *et al.*, 1987). The gene is within a region of ~2 kb that is GC rich in base composition and shows no CpG suppression (Figure 2a). The alcohol dehydrogenase 1 (*Adh1*) is induced under anaerobic conditions in several tissues (Dennis *et al.*, 1984). A region of several hundred base pairs at the 5′ end of the gene is both GC rich and shows no CpG suppression (Figure 2b). The sucrose synthase gene (*Sh1*) is involved in starch metabolism in endosperm (Werr *et al.*, 1985). The gene has GC-rich regions at both its 3′ and 5′ ends (Figure 2c). The CpG frequency within the 5′ GC rich region is on average about half of the equivalent GpC frequency and less than predicted from base composition. This cluster is therefore a less obvious candidate for a CpG island than the others. A fourth gene, that for the major seed protein zein, showed no clustering of CpG (Figure 2d). The zein gene family is expressed exclusively in endosperm and comprises multiple gene copies with some sequence heterogeneity between them (Pedersen *et al.*, 1982). CpGs are significantly suppressed across the entire gene copy shown in Figure 2d.

Lack of methylation at CpG clusters

The presence of clustered CpGs associated with genes *a1*, *Sh1* and *Adh1* is reflected in an increased local frequency of *Hpa*II and *Hin*PI sites (Figure 3). The close spacing of the sites means that DNA fragments from these regions could contribute to the HTF fraction of maize DNA, but only if the sites for *Hpa*II and *Hin*PI are non-methylated in the genome. We tested this by using restriction endonucleases to determine the methylation status of the sites in DNA from maize shoots. Gene-specific probes were used to detect the relevant sequences in blots of total maize DNA. In the case of the *a1* gene, all 26 testable sites containing either CpGs or CpXpGs across the coding region turned out to be non-methylated, while the nearest *Hpa*II and *Eco*RII sites 5′ and 3′ were completely methylated (Figure 3). The 5′ CpG-rich regions of the *Adh1* and *Sh1* genes were also non-methylated, as were neighbouring sequences within the genes. One partially methylated site was found between the two CpG clusters of the *Sh1* gene (Figure 3). In cases where restriction sites are very close together (as for *Hpa*II, *Hin*PI or *Eco*RII) it is not possible to resolve all the resulting restriction fragments in agarose gels because of their tiny size. However, it is probable that they are unmethylated also because the subset we have tested is unbiased and dependent only on the chance of two sites being separated enough from each other to produce an identifiable single band on the autoradiograph. From these results we conclude that the HTF fraction of DNA from maize shoots (Figure 1) contains DNA fragments derived from the CpG clusters at the *a1, Adh1* and *Sh1* genes.

The zein gene, on the other hand, almost certainly does not contribute fragments to the HTF fraction. This gene is CpG-deficient (Figure 2d), with only one *Hpa*II site and one *Hin*PI site present in a cloned representative gene (Pedersen *et al.*, 1982). Moreover no low mol. wt fragments were seen when the zein gene family was probed in maize DNA that had been digested with methyl-sensitive restriction enzymes (Figure 4). The rarity of testable sites, and the sequence heterogeneity map for the zein gene family, makes it impossible to deduce a methylation map for the zein locus. The sequence data and the blots suggest, however, that most zein genes are not associated with clusters of non-methylated CpG sites.

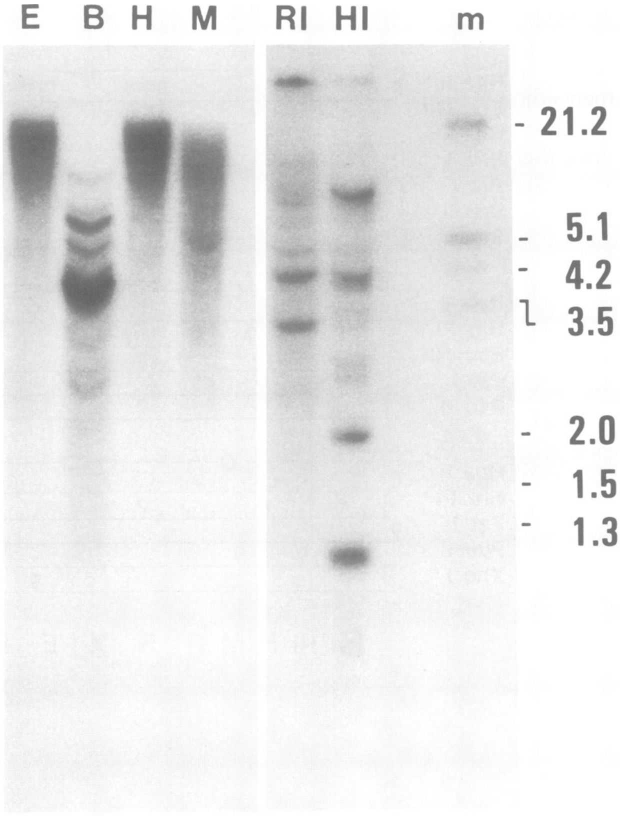

Fig. 4. Presence of m^5CpG and $m^5Cp(A/T)pG$ sites at the zein cluster. Right panel, nuclear maize DNA was digested with *Eco*RI (RI) and *Bam*HI (HI). After blotting, the filter was probed with the zein cDNA. Both enzymes cut outside the zein genes. The numerous bands give an indication of the high number of copies per genome. Left panel, blot of maize nuclear DNA after digestion with *Eco*RII (E), *Bst*NI (B), *Hpa*II (H) and *Msp*I (M) and probed with zein cDNA. The zein sequences are resistant to the m^5C sensitive endonucleases *Eco*RII and *Hpa*II, but are more extensively cut by *Bst*NI and *Msp*I (see text). Marker DNA as in Figure 3.

Discussion

Our experiments have shown that four plant genomes contain an HTF fraction in which sites for CpG enzymes are non-methylated and closely spaced. Three maize genes, *Adh1, a1* and *Sh1,* possess CpG enzyme sites that are sufficiently frequent to contribute to this fraction, and we have found that these sites are indeed non-methylated in the genome. The results are surprising as plant DNA is known to be very highly methylated, and they recall parallel findings in the genomes of vertebrate animals.

How close is the similarity between the non-methylated CpG regions seen here and the CpG islands of vertebrates? The CpG-rich region associated with the *a1* gene is indistinguishable from a typical vertebrate island. It is almost 2 kb in length, and covers the entire transcribed region plus a shorter sequence upstream of the 5′ end of the gene (Figure 3A). Normally CpG islands are 1−2 kb long and cover only part of the transcribed portion of a gene (Gardiner-Garden and Frommer, 1987; Bird, 1987), but when the gene is short, it may be entirely within the CpG island (for example the alpha globin gene of man; see Bird *et al.*, 1987). The CpG clusters associated with *Sh1* and *Adh1* cover only part of the transcribed portion of the gene, although mapping with methyl-sensitive endonucleases did not accurately define the boundaries of the non-methylated region. A few non-methylated sites in DNA from maize shoots were found outside the CpG-rich region (Figure 3A). It is clear, however, that the CpG-rich regions at these genes are non-methylated in shoot DNA.

Another diagnostic feature of CpG islands is that they lack methylation in a wide range of tissues, whether or not the associated gene is expressed (Bird, 1987). We therefore expect that the islands at the *a1*, *Adh1* and *Sh1* genes should be methylation-free regardless of expression. Our experiments did not address this question for the maize genes, as we have analysed DNA from shoots which can express all three (Schwarz-Sommer *et al.*, 1987; Freeling and Bennett, 1985; Springer *et al.*, 1986). However, Nick *et al.* (1986) have examined in detail the methylation status of the CpG cluster at the maize *Adh1* gene in leaf, a tissue in which the gene is repressed. Using the technique of genomic sequencing, they analysed 61 CpG and CpXpG sites within a 1400 bp sequence at the 5′ end of the gene. None of the sites within 900 bp upstream of the gene were methylated (see broken line in Figure 3A(b)), indicating that methylation is absent even when the gene is silent. This property, together with its CpG-richness (Figure 2), confirms this region as a CpG island. We propose that the CpG clusters associated with the *a1* and *Sh1* genes are likewise non-methylated in a tissue-independent manner, and therefore are also analogous to the CpG islands in vertebrate genomes. Other plant genes that could be associated with the unmethylated fraction are the *Waxy* locus in maize (Klosgen *et al.*, 1986) and the lectin 1 gene in soybean (Vodkin *et al.*, 1983). Remarkably, both show the same CpG clustering that is present in the *a1*, *Adh1* and *Sh1* genes. Moreover, in the *Waxy* locus, at least 10 different CpG and CpXpG sites have been described as unmethylated, suggesting that a CpG island exists at this gene (Schwarz-Sommer *et al.*, 1984; Wessler *et al.*, 1986; Chomet *et al.*, 1987).

In mammals, the existence of islands has greatly facilitated genome mapping and has helped in locating genes. The reason for this is that CpG islands are preferentially cut by enzymes containing CG rich recognition sequences that are blocked by m^{5}C (Brown and Bird, 1986). If many or all CpG islands in plants turn out to be associated with genes it should be possible to pinpoint genes or their 5′ regions by finding sequences sensitive to CpG enzymes. Furthermore, given the enormous genome size of many important plants, non-methylated CpG islands might be used as landmarks for long-range mapping of the chromosomal DNA.

The presence of CpG islands in the genome of higher plants is striking because among animals only vertebrates have been found to exhibit this feature. This means that >95% of the animal species may not have CpG islands. Since there is no obvious common ancestor with CpG islands linking plants and vertebrates, it is likely that this phenomenon has arisen independently in each group. The implication is that both types of organisms experienced a similar selective pressure which led to the appearance of islands. It is possible that the genome size was involved in this. Vertebrates and higher plants are alike in having comparatively large genomes most of which are never transcribed. CpG islands may reduce the problems associated with large genome size by highlighting genetically active regions of the chromosomes.

Materials and methods

DNAs were isolated from leaves of wheat, tobacco and rye as in Graham (1978) and from 3 day old maize shoots as in Luthe and Quatrano (1980). Restriction digests were performed according to the manufacturer's instructions. Complete digestion was monitored in all cases by addition of bacteriophage lambda DNA to an aliquot of the digestion mixes.

DNA fragments were end-labelled with [^{32}P]dCTP and the Klenow fragment of DNA polymerase I before electrophoresis on 1.2% agarose gels as in Cooper *et al.* (1983).

For Southern analysis, restriction fragments were separated in 1.2% agarose gels and then alkaline blotted onto Zeta-probe membranes following the suppliers indications (Bio-Rad). Appropriate probes (see legend for Figure 3) were oligo-labelled according to Feinberg and Vogelstein (1984).

Acknowledgements

We are very grateful to Brian Bathgate, Laura Smith and Chris Leaver for the gift of some DNAs and advice during maize DNA isolation, and María José F.Sarabia, Susan Lindsay and Patrizia Lavia for comments on the manuscript. We also would like to thank Z.Schwarz-Sommer and H.Saedler for the *a1*, P.Starlinger and W.Werr for the *Sh1*, L.Dennis and M.Freeling for the *Adh1* and J.Brown for the zein probes. F.A. is supported by a long term EMBO fellowship. This work was supported by the MRC, London.

References

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

Bird,A.P. (1987) *Trends Genet.*, **3**, 342–347.

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

Bird,A.P., Taggart,M.H., Nicholls,R.D. and Higgs,D.R. (1987) *EMBO J.*, **6**, 999–1004.

Brown,W.R.A. and Bird,A.P. (1986) *Nature*, **322**, 477–481.

Chomet,P.S., Wessler,S.R. and Dellaporta,S.L. (1987) *EMBO J.*, **6**, 295–302.

Cooper,D.N., Taggart,M.H. and Bird,A.P. (1983) *Nucleic Acids Res.*, **11**, 647–658.

Dennis,E.S., Gerlach,W.L., Pryor,A.J., Bennetzen,J.L., Inglis,A., Llewellyn,D., Sachs,M.M., Ferl,R.J. and Peacock,W.J. (1984) *Nucleic Acids Res.*, **12**, 3983–4000.

Feinberg,A.P. and Vogelstein,B. (1984) *Anal. Biochem.*, **137**, 266–267.

Freeling,M. and Bennett,D.C. (1985) *Annu. Rev. Genet.*, **19**, 297–323.

Gardiner-Garden,M. and Frommer,M. (1987) *J. Mol. Biol.*, **196**, 261–282.

Graham,D.E. (1978) *Anal. Biochem.*, **86**, 609–613.

Gruenbaum,Y., Naveh-Many,T., Cedar,H. and Razin,A. (1981) *Nature*, **297**, 860–862.

Klosgen,R.B., Gierl,A., Schwartz-Sommer,Z. and Saedler,H. (1986) *Mol. Gen. Genet.*, **203**, 237–244.

Luthe,D.S. and Quatrano,R.S. (1980) *Plant Physiol.*, **65**, 305–308.

Nick,H., Bowen,B., Ferl,R.J. and Gilbert,W. (1986) *Nature*, **319**, 243–246.

Pedersen,K., Devereux,J., Wilson,D.R., Sheldon,E. and Larkins,B.A. (1982) *Cell*, **29**, 1015–1026.

Schwarz-Sommer,Z., Gierl,A., Klosgen,R.B., Wienand,U., Peterson,P.A. and Saedler,H. (1984) *EMBO J.*, **3**, 1021–1028.

Schwarz-Sommer,Z., Shepherd,N., Tacke,E., Gierl,A., Rodhe,W., Leclerq,L.M., Mattes,M., Berndtgen,R., Peterson,P.A. and Saedler,H. (1987) *EMBO J.*, **6**, 287–294.

Setlow,P. (1976) *Handbook of Biochemistry and Molecular Biology.* CRC Press, pp. 313–318.

Shapiro,H.S. (1976) *Handbook of Biochemistry and Molecular Biology.* CRC Press, pp. 258–262.

Springer,B., Werr,W., Starlinger,P., Bennett,D.C., Zokolica,M. and Freeling,M. (1986) *Mol. Gen. Genet.*, **205**, 461–468.

Vodkin,L.O., Rhodes,P.R. and Goldberg,R.B. (1983) *Cell*, **34**, 1023–1031.

Werr,W., Frommer,W.B., Maas,C. and Starlinger,P. (1985) *EMBO J.*, **4**, 1373–1380.

Wessler,S.R., Baran,G., Varagona,M. and Dellaporta,S.L. (1986) *EMBO J.*, **5**, 2427–2432.

Received on January 25, 1988; revised on March 28, 1988