The distribution of 5-methylcytosine in the nuclear genome of plants

L.M.Montero+, J.Filipski[1], P.Gil§, J.Capel, J.M.Martínez-Zapater and J.Salinas*
Dpto. de Protección Vegetal, Centro de Investigación y Tecnología, INIA, Ctra. de La Coruna Km 7, 28040 Madrid, Spain and [1]Institut J. Monod, CNRS, 2 Place Jussieu, Paris 75005, France

Received December 31, 1991; Revised and Accepted May 19, 1992

ABSTRACT

We have determined the 5-methylcytosine (5mC) content in high molecular weight DNA, from two dicot (tobacco and pea) and two monocot (wheat and maize) plant species, fractionated according to base composition. The results show that the proportion of 5mC in the genomic fractions increases linearly with their guanine+cytosine (G+C) content while the proportion of non-methylated cytosine remains almost constant. This can be interpreted as a consequence of a difference in mutation pressure related to spontaneous deamination of 5mC to thymine between the different compartments of plant genomes.

INTRODUCTION

In many organisms, nuclear DNA is methylated at cytosine (C) residues to give 5-methylcytosine (5mC). Evidence exists from previous studies which indicate that, in eukaryotes, DNA methylation is involved in regulation of gene expression (1–3). Levels of DNA methylation are variable in animal genomes, ranging from undetectable amounts in some insects to about 8% of total cytosine in vertebrates (4). In all cases, more than 95% of 5mC is found in the dinucleotide CpG (5). The nuclear genome of higher plants is, generally, more heavily methylated, the level of 5mC accounting for more than 30% of total cytosines in some species (4). The cytosine is methylated in plants not only at CpGs but also at a variety of other cytosine containing dinucleotides, all of which are part of the basic trinucleotide CpNpG where N can be any nucleotide (6).

Studies on the distribution of CpG dinucleotides in eucaryotic genomes, using restriction endonucleases (7, 8), revealed two basic patterns common for plants and vertebrates. In the first pattern, CpG dinucleotides are generally methylated and scattered along the DNA in both coding and noncoding sequences, at lower frequency than expected on the basis of G+C content. In the second pattern, CpGs are clustered in 1–2 Kb long segments of DNA, they are not methylated, and their frequency is close to expected. These clusters of CpGs have been called CpG islands, and they generally overlap regulatory sequences at the 5′ end of housekeeping genes and some tissue specific genes.

The two patterns mentioned above are related to the long-range structure of genomes. In fact, the nuclear genomes of angiosperms, as those of vertebrates, are compositionally compartmentalized; i. e. they are organized in mosaics of long (over 300 Kb), compositionally relatively homogeneous DNA segments, called isochores (9–11). In mammals, it has been shown that the CpG islands are located in G+C rich isochores, while G+C poor isochores contain few, if any, of them (12). A similar situation has been found when studying the distribution of CpG dinucleotides in several genes of angiosperms. CpGs are strongly avoided in G+C poor genes, which are located in G+C poor isochores, whereas they are only slightly avoided in G+C rich genes, located in G+C rich isochores (11). The pattern of distribution of CpG dinucleotides in DNA provides information relative to both the function and the evolutionary forces shaping DNA sequences. First, although the precise role of CpG islands is not entirely clear, general consensus is that they could be involved in regulating gene expression in neighboring genes. There are genes which lack this level of regulation. Second, the CpG dinucleotides mutate one order of magnitude faster than the others. Their accumulation in or their disappearance from a sequence indicates that some major selective forces, or mutational biases, were involved during the sequence evolution.

Here, we have determined the distribution of 5mC in DNA fractions from four plant genomes trying to shed light on the evolution of their compositional organization at the level of long DNA stretches. Because of the differences in genome organization between dicots and Gramineae (10,13), we selected for this study four angiosperm species representing both Gramineae (wheat and maize) and dicots (pea and tobaco). A similar study was carried out several years ago (14) when low molecular weight (around 1 Kb) DNAs from three dicots were fractionated into low, moderate and high G+C fractions, and the 5mC content of these fractions was determined. The authors found a constant level of non-methylated cytosine and a linear

* To whom correspondence should be addressed

Present addresses: +Instituto del Frio, CSIC, Ciudad Universitaria 28003 Madrid, Spain and §MSU-DOE Plant Research Laboratory, Michigan State University, East Lansing, MI 48824, USA

TABLE I. Composition of total unfractionated DNAs and their genomic fractions

MAIZE						WHEAT					
Fractions	%Genome	% G+C	% C	% 5mC	$r_{met/exp}$	Fractions	% Genome	% G+C	% C	% 5mC	$r_{met/exp}$
TOTAL	100.0	46.40	16.50	6.70	0.71	TOTAL	100.0	44.80	16.50	5.90	0.66
1	7.8	39.00	15.60	3.90	0.59	1	20.6	43.10	16.50	5.05	0.61
2	6.0	39.60	16.10	3.70	0.52	2	1.7	43.40	16.60	5.10	0.61
3	8.6	41.80	16.00	4.90	0.63	3	8.5	44.00	16.20	5.80	0.67
4	11.9	42.20	16.00	5.10	0.64	4	27.0	44.70	16.60	5.75	0.65
5	12.9	43.60	16.00	5.80	0.69	5	12.4	45.40	16.00	6.70	0.73
6	12.4	44.40	16.20	6.00	0.68	6	9.6	47.20	16.90	6.60	0.67
7	18.0	45.20	16.40	6.20	0.68	7	8.3	47.40	16.60	7.10	0.72
8	8.6	47.20	16.70	6.90	0.70	8	4.0	49.00	16.95	7.65	0.73
9	5.8	48.10	16.45	7.60	0.75	9	1.6	49.00	17.00	7.50	0.71
10	5.1	51.10	16.10	9.45	0.83	10	6.3	49.20	16.50	8.10	0.76
11	2.9	48.70	16.20	8.15	0.78						

TOBACCO						PEA					
Fractions	% Genome	% G+C	% C	% 5mC	$r_{met/exp}$	Fractions	% Genome	% G+C	% C	% 5mC	$r_{met/exp}$
TOTAL	100.0	37.70	12.35	6.50	0.96	TOTAL	100.0	37.40	13.00	5.70	0.90
1	11.9	35.20	12.20	5.40	0.96	1	16.5	31.40	12.40	3.30	0.76
2	1.3	35.20	12.70	4.90	0.87	2	12.0	34.40	12.60	4.60	0.85
3	4.0	35.80	12.70	5.20	0.98	3	15.3	34.90	12.50	4.95	0.89
4	30.8	36.60	12.40	5.90	0.97	4	13.7	36.60	12.75	5.55	0.91
5	19.2	38.00	12.30	6.70	1.02	5	12.2	37.60	12.80	6.00	0.94
6	19.8	38.00	12.10	6.90	1.05	6	5.5	38.20	13.10	6.00	0.91
7	5.8	39.20	12.50	7.10	1.02	7	6.5	38.40	13.10	6.10	0.92
8	3.0	39.60	12.80	7.00	0.92	8	6.9	38.60	13.10	6.20	0.92
9	1.3	40.20	12.60	7.50	1.03	9	4.5	40.40	12.60	7.60	1.04
10	2.9	41.40	12.70	8.00	1.04	10	6.9	39.00	12.95	6.55	0.95

This table provides the relative content of Guanine + Cytosine (% G+C), Cytosine (% C) and 5 methylcytosine (% 5mC) in total unfractionated DNAs and in DNA fractions. The values were determined by HPLC as outlined in Material and Methods. The relative amount of each DNA fractions within the genomes (%Genome) and the extend of methylation of expected targets (r met/exp) in total unfractionated DNAs and in DNA fractions are also indicated.

increase of 5mC with the G+C content. We extended these results to new plants using more refined fractionation techniques. In addition, we used for fractionation high molecular weight DNA (50–100 Kb) which enabled us to draw conclusions about the evolution of long range chromosomal structures. The results obtained here indicate that, in plant genomes, the differences in 5mC content account for most of the differences in G+C content between isochores, and allow us to suggest that the main cause for the compositional compartmentalization of the plant genomes is a different extent of mutational G+C and A+T pressures in different genomic compartments.

MATERIALS AND METHODS

Isolation and fractionation of nuclear DNA

Etiolated seedlings from wheat (*Triticum aestivum* L., cv. Ribereño) and pea (*Pisum sativum* L., cv. Desso) and leaves from tobacco (*Nicotiana tabacum* L., cv. White Burley) and maize (*Zea mays* L., inbred line 82-2017-1/3012-4), were used to isolate nuclear DNA as previously described (10). The size of the DNA fragments obtained in these preparations was in the 50–100 kb range, as determined by rotational electrophoresis with appropriate size markers.

Fractionation of nuclear DNAs by preparative centrifugation in Cs_2SO_4 density gradients, in the presence of the DNA ligand BAMD [3,6 bis (acetate-mercuri-methyl) dioxane], was carried out as described elsewhere (10). Gradients were fractionated, and aliquots were pooled into 10 or 11 fractions, the pellet being considered as the first fraction. In the case of the pea seedlings, we used the fractions described in the Figure 1 of reference 10.

HPLC base composition analysis

Two micrograms of DNA from each fraction, from total unfractionated nuclear DNA, and from pBR322, used as standard, were hydrolized to individual bases as previously described (15). The separation of the bases was performed on a Beckman Ultrasphere IP C18 (4.6×150 mm) column equilibrated with 10 mM potassium dihydrogen phosphate pH 3.1, 3 mM hexanesulfonic acid, and 2% (v/v) acetonitrile at a flow rate of 1 ml/min. The peaks corresponding to each base were detected at 280 nm. The mol percentage of each base was determined from their peak areas. Each hydrolizate was analyzed twice, being both values always less than 0.5% different in G+C. The results presented are the average of the two analyses.

Estimation of expected methylation targets

The amount of theoretically expected methylation targets was estimated assuming that (i) there are two types of targets in plant genomes (CpG and CpNpG) and (ii) there is a random distribution of nucleotides along the DNA. The density of targets was calculated assuming that the probability of encountering the dinucleotide CpG in a nucleotide sequence equals $p_1 = r_C.r_G$ (where r_C and r_G correspond to the fractions of total cytidine, methylated and non-methylated, and guanidine nucleotides respectively in the DNA), and the probability of trinucleotide CpNpG equals $p_2 = r_C.r_N.r_G$. The CpG-containing trinucleotide CpCpG has two methylation targets both of which should be taken into account, while the trinucleotide CpGpG, with probability $p_3 = r_C.r_G^2$, should not be taken into account because its cytosine was already accounted for as a methylation target in CpG. Therefore, the formula for the expected density is:

$$p = 2r_C.r_G - r_C.r_G^2 = r_C.r_G(2 - r_G) \quad \text{(equation 1)}$$

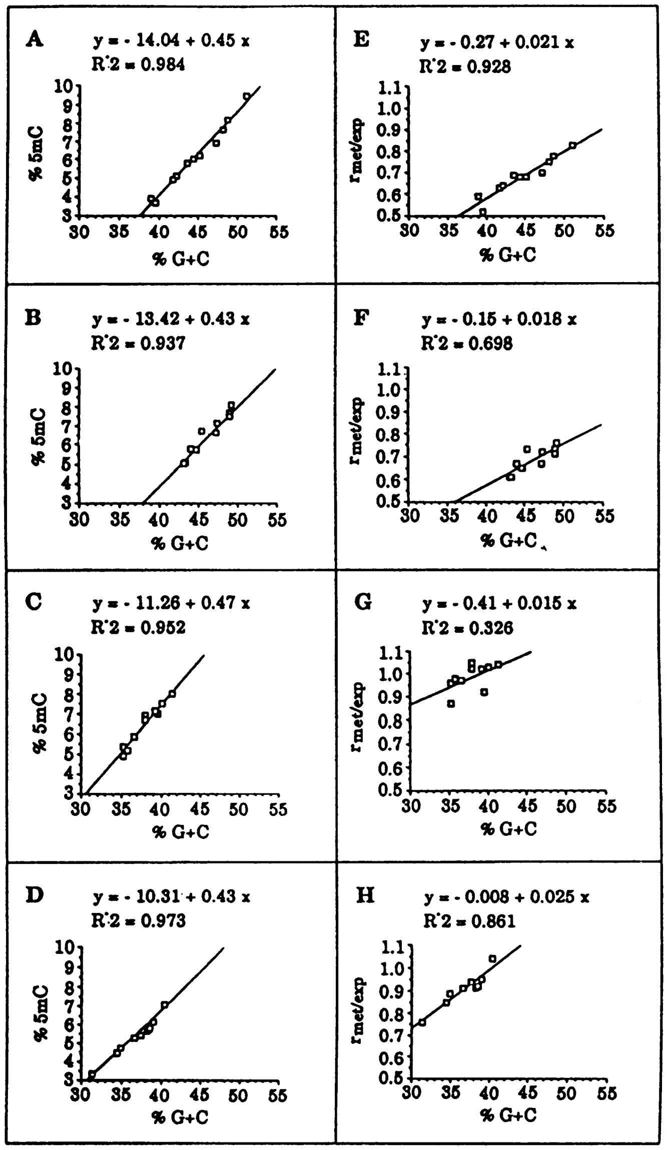

Figure 1. 5mC content (A–D) and relative methylation of calculated targets (E–H) as a function of G+C content in DNA fractions from maize (A and E), wheat (B and F), tobacco (C and G) and pea (D and H).

RESULTS AND DISCUSSION

The proportion of 5mC in genomic fractions is linearly related to their G+C contents

Plant genomes have been shown to be organized in mosaics of long, compositionally homogeneous DNA segments called isochores (10, 11). Here, we investigated the distribution of 5mC in DNA fractions containing different families of isochores which compose the genomes of two *Gramineae* and two dicot species. Table I presents the total G+C, cytosine and 5mC contents in unfractionated DNA and in DNA fractions. The relative amount of each DNA fraction within the genomes is also indicated. The 5mC contents of total unfractionated DNAs are very close to those previously reported (4, 16).

The results reveal that the level of methylated cytosine in the genomic fractions strongly depend on their G+C content (Fig. 1). On the other hand, the proportion of unmethylated cytosine varies very slightly within the genomes, in spite of considerable differences in the total G+C content between fractions (Table I). These results indicate that the differences in composition between genomic compartments are mainly due to different levels of 5mC in different chromosomal regions.

The extent of methylation of expected targets in genomic fractions is linearly related to their G+C content

Table I also presents the extent of methylation of expected targets ($r_{met/exp}$), i.e. the ratio between the content of 5mC found in fractions (and in total unfractionated DNAs), and the number of theoretically calculated methylation targets. The extent of methylation of expected targets in a DNA sequence is the product of two parameters:

$$r_{met/exp} = r_{ava/exp} \cdot r_{met/ava} \qquad \text{(equation 2)}$$

$r_{ava/exp}$ is the ratio of available to expected targets. Usually, it is lower than 1 due to the underrepresentation of methylation targets in genomes. It depends on the type of target and on the type of nucleotide sequence. The CpNpG targets are apparently not underrepresented in plants ($r_{ava/exp}=1$) (17), unlike the CpG targets in which case this parameter varies widely (7).

$r_{met/ava}$ represents the methylation of available targets (for example, 0.7–0.9 on the average in wheat germ, 6).

The $r_{met/exp}$ could be calculated by using the equation 1:

$$r_{met/exp} = \%5mC/100p \qquad \text{(equation 3)}$$

In all four plants studied, the extent of methylation of expected targets ($r_{met/exp}$) increases in succeeding fractions as the G+C content increases (Table 1; Fig. 1). This was unexpected since one might have anticipated a lower $r_{met/exp}$ in G+C rich fractions than in A+T rich ones due to the high proportion of CpG dinucleotides found in genes located in G+C rich fractions (11). Thus, the CpG islands probably constitute a too small portion of the G+C rich fractions to influence their average base composition.

The observed increase of $r_{met/exp}$ with G+C content (Table 1; Fig. 1) is probably a result of the increase of $r_{ava/exp}$ in succeeding fractions. In other words, methylation targets are probably less underrepresented in G+C rich compartments than in the A+T rich ones. The parameter $r_{met/ava}$, representing the extent of methylation of available targets, would influence the $r_{met/exp}$ to a lower extent. The alternative possibility, i.e. that the differences in $r_{met/exp}$ are mainly due to lower methylation of available targets ($r_{met/ava}$) in A+T rich fractions than in G+C rich ones, is not very plausible because a high frequency of CpG doublets, typical of non-methylated islands, have been found in G+C rich compartments (11), and outside of these islands the plant genomes are apparently uniformly and highly methylated (7).

Compositional compartmentalization could be caused by different mutational pressures in different genomic regions

The persistence of higher levels of CpG dinucleotides in G+C rich DNA fractions, and even the existence of G+C rich compartments in spite of this mutational pressure towards A+T is probably due to two different phenomena. On one hand, CpG dinucleotides present at CpG islands are generally not methylated and, therefore, they do not constitute mutational hot spots. On the other hand, DNA repair mechanisms likely exist which protect 5mC from disapearance by mutation (18). McClelland (17) postulated that there is a system of mismatch repair, which preferentially repairs G:T back to G:C, if these mismatches arise by deamination of 5mC located in CpNpG trinucleotides. This would explain why these trinucleotides are not underrepresented in plants. A similar system might also protect the CpG from disappearance. Evidence indicates that such systems may exist in mammals (19) and in Xenopus (20). Both of these mechanisms would not only maintain the existing methylation targets, but would also occasionally convert to G:C the G:T mismatches resulting from errors of replication of A:T base pairs, exerting a mutational pressure towards G+C. In the compartmentalized genomes of plants, the balance between mutational pressures towards A+T and G+C would had been set at different levels in different compartments, because of different degrees of involvement of the protection mechanisms of methylation targets in the process of DNA repair (20).

The constant level of non-methylated cytidine in DNA fractions differing in composition suggests that the genomic compartments in plants derived from compositionally more uniform DNA mainly by 5mC:G to T:A mutations, occurring with different frequencies in different chromosomal regions. In fact, as a result of these mutations, the proportion of A and T increases, the proportion of 5mC and G diminishes, while the proportion of non-methylated C remains constant.

In addition, we have found that the G+C rich genomes (wheat and maize) are relatively less methylated (about 70% of calculated targets) than the A+T rich genomes (pea and tobacco, 90 and 96% of calculated targets respectively, see Table 1). This finding could exemplify a more general trend. On this way, preliminary calculations for 66 plants, performed by using data from the literature (4) and assuming random nucleotide distribution, showed that the higher G+C content, the lower the occupation of expected methylation targets (Filipski et al. unpublished). This also suggests that 5mC:G to T:A mutation pressure weights heavily on the overall balance of biasses and selection forces operating in plant genomes.

ACKNOWLEDGEMENTS

We thank J.L.Tadeo for technical advice. L.M.M. was a recipient of a fellowship from the Spanish Ministry of Science and Education. This work was supported by grant INIA7556 from the Spanish Ministry of Agriculture.

REFERENCES

1. Cedar, H. (1988). Cell **53**, 3–4.
2. Adams, R.L.P. (1990). Biochem. J., **265**, 309–320.
3. Lewis, J. and Bird, A. (1991). FEBS Lett., **285**, 155–159.
4. Shapiro, H.S. (1976). Handbook of Biochemistry and Molecular Biology, CRC Press, pp. 258–262.
5. Bonen, L., Huh, T.Y., and Gray, T.W. (1980). FEBS Lett., **111**, 340–346.
6. Gruenbaum, Y., Naveh-Many, T., Cedar, H., and Razin, A. (1981). Nature **292**, 860–862.
7. Antequera, F., and Bird, A.P. (1988). EMBO J., **7**, 2295–2299.
8. Bird, A.P. (1986) Nature **321**, 209–213.
9. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M. and Rodier, F. (1985). Science **228**, 953–958.
10. Salinas, J., Matassi, G., Montero, L.M., and Bernardi, G. (1988). Nucleic Acids Res., **17**, 5273–5290.
11. Montero, L.M., Salinas, J., Matassi, G., and Bernardi, G. (1990). Nucleic Acids Res., **18**, 1859–1867.
12. Bickmore, W.A. and Sumner, A.T. (1989). Trends Genet., **5**, 144–148.
13. Matassi, G., Montero, L.M., Salinas, J. and Bernardi, G. (1989). Nucleic Acids Res., **17**, 5273–5290.
14. Sulimova, G.E., Mazin, A.L., Vanyushin, B.F. and Belozerskii, A.N. (1970). Dokl. Acad. Nauk. SSSR **193**, 1422–1425.
15. Klaas, M. and Amasino, R.M. (1989). Plant Physiol., **91**, 451–454.
16. Wagner, I., and Capesius, I. (1981). Biochim. Biophys. Acta **654**, 52–56.
17. McClelland, M. (1983). J. Mol. Evol., **19**, 346–354.
18. Filipski, J. (1987). FEBS Lett., **217**, 184–186.
19. Brown, T.C. and Jiricny, J. (1987) Cell **50**, 1571–1575.
20. Filipski, J. (1990). Adv. Mutagen. Res., **2**, 1–54.