CpG and TpA frequencies in the plant system

Mustapha Boudraa and Pascale Perrin

Institut d'Evolution Moléculaire, Université Claude Bernard, Lyon I, 43 Boulevard du 11 Novembre 1918, 69622 Villeurbanne, France

Received May 26, 1987; Accepted June 12, 1987

ABSTRACT

Higher plant nuclear sequences reveal avoidance of CpG and TpA doublets. Chloroplast sequences avoid the TpA doublet in all codon positions. The chloroplast genome is not methylated but codon positions II-III and untranslated regions avoid CpG. The mitochondrial genome, also unmethylated, avoids CpG in all codon positions. We therefore deduce that methylation is not sufficient to explain CpG avoidance in the higher plant systems. Other factors must be taken into account such as amino acid composition, codon choices and perhaps stability of the DNA helix.

INTRODUCTION

Uncertainty continues over the reasons for the rarity of the CpG dinucleotide in DNA (1-6), obscuring the relationship of this aspect of molecular evolution to the frequencies of other dinucleotides, codon usage, genomic base composition, variable G+C content along sequences, and runs of one kind of base (7-15). In addition, the TpA doublet is even more generally rare than is CpG, although TpA avoidance is often less intense (9-11). Recognizing the frequent neglect of plant system sequences in studies of molecular evolution, we have decided to analyze those published for CpG and TpA occurrences in order to gain a better perspective on the avoidance of these doublets. CpG and TpA are the most generally avoided of the 16 doublets in double-stranded DNA genomes (9). CpG is not avoided in prokaryotes but TpA is rare in both prokaryotes and eukaryotes, while in vertebrate mitochondria it is not avoided at all (6,9,14,15). Unfortunately insufficient sequences from postulated mitochondrial ancestors are available to see whether they avoid TpA. To rationalize these observations, two hypotheses have been advanced: methylation of CpG followed by deamination mutation to produce TpG (1,2), and absence of stop codons TAA and TAG, removing TpA from codon position I-II except for coding tyrosine. mCpG is therefore replaced by TpG, and CpA on the complementary

strand, resulting in a CpG deficiency in favor of TpG and CpA increase (2).

These hypotheses seem unable to explain all observed phenomena. For example, our results reveal that CpG deficiency appears even in non-methylated organelles. Studies have already shown that several unmethylated species avoid CpG (5,10,16). Likewise, from the absence of in-frame stop codons, low TpA frequency is expected only in codon position I-II. Our study reveals, however, that TpA is rare in non-coding sequences (rRNA and introns) as well as in codons of the plant system. The same result has been noted in other systems (6).

MATERIALS AND METHODS

Sample

Taking the largest possible sample of available sequences in GenBank (release 48), we extract and analyze sequences with the ACNUC retrieval system and ANALSEQ programs (9,17-19). For some families of homologous genes, sequences are known in several species. This could strongly bias the analysis hence we arbitrarily select only one sequence in each gene family. In fact we have found similar values to those reported here by simply analyzing all the sequences, without elimination of homologous ones. The number of sequences of each type is distributed as follows:

Nuclear genes	Chloroplast genes	Mitochondrial genes
25	38	3
(26,850)	(24,426)	(3,561)

where the base number in each category appears in parenthesis

Table 1 gives the complete list of sequences studied with species, and category of protein.

Methods

The method generally used to study dinucleotide frequencies is the ratio between observed and theoretical (expected) values calculated from base composition (4,9,14,15,20).

In non-coding regions we count S, the number of occurrences of each dinucleotide. Then we do all possible base permutations, conserving the base composition of the sequence, and then calculate the mean E(S) and the variance V(S) of the distribution of S. In this case the formulas of

Gautier et al. (21) become those of Knight (22):

$$E(S) = \frac{nA \cdot nB}{n} \quad \begin{array}{l} n, \text{ sequence length in bases} \\ nA, \text{ number of base A} \\ nB, \text{ number of base B} \end{array}$$

$$V(S) = \frac{nA \cdot (nA-1) \cdot nB \cdot (nB-1)}{n(n-1)} + E(S) - E(S)^2$$

In coding sequences, we respect triplet organization by considering dinucleotide frequency in each position: I-II, II-III and III-I; the last corresponds to the third base of one codon and the first base in the following codon.

Next Z, the number of occurrences of a dinucleotide in a particular codon position in the studied sequence, or in the reference set, is computed. The reference set is all possible sequences having the same base content in each codon position as the real sequence. We then calculate the mean $E(Z)$ and variance $V(Z)$ of the distribution of Z and the expression:

$$\frac{Z - E(Z)}{\sqrt{V(Z)}} \quad (21)$$

This variable approximates a normal law. So the absolute value of this index must be compared to 1.96 or 2.57, which represent the 5% and 1% probability level, respectively. For a group of sequences, we cumulate as follows:

$$\frac{1}{\sqrt{N}} \sum \frac{Z_i - E(Z_i)}{\sqrt{V(Z_i)}}, \quad \begin{array}{l} \text{where } N \text{ is the number of sequences, and} \\ Z_i \text{ the number of occurrences in the } i\text{th} \\ \text{sequence.} \end{array}$$

This measure express the degree of certainty that a doublet is avoided in a group (workfile) of sequences. Therefore, to evaluate the average intensity of the avoidance or preference of a doublet in a workfile, we will use the mean value i.e. the mean of normalized frequencies for the sequences in a workfile (6). This value can be used to compare doublet usage between different workfiles (when sequence lengths are similar).

RESULTS

CpG avoidance and the methylation problem

Table 2 gives normal statistics for dinucleotides TpA, CpG, CpA and TpG. Results for CpG, TpG and CpA in plant nuclear genes appear compatible with the methylation scheme; as in vertebrates, CpG is avoided in all codon positions. This avoidance is counterbalanced by elevated TpG and CpA

Table 1. Higher plant sequence portfolio. These sequences are all in GenBank, release 48. The symbols comprise two parts: the first one, before the point, is the sequence name, and the second one concerns the sequence type (PE for peptide)(19). Chl, chloroplast; Nuc, nuclear; Mit, mitochondrion

Organelle, Species, gene	Symbol	No. bases
Chl, <i>Hordeum vulgare</i> , ATPase subunit B	BLYCPATPB.PE1	1497
Chl, <i>Hordeum vulgare</i> , " E	BLYCPATPB.PE2	414
Chl, <i>Zea mays</i> , ribosomal protein S4	MZECPRPS4.PE1	606
Chl, <i>Zea mays</i> , ribulose biphosphate carboxylase large subunit	MZECPRUBP.PE1	1428
Chl, <i>Oenothera hookeri</i> , pre-apocytochrome f	OHOCPCYTF.PE1	954
Chl, <i>Pisum sativum</i> , cytochrome f	PEACPCYF.PE1	102
Chl, <i>Sinapsis alba</i> , photosystem III membrane protein	SALCPPSII.PE1	1062
Chl, <i>Spinacia oleracea</i> , p-680 chlorophyll alpha apoprotein	SPICAAP.PE1	1527
Chl, <i>Spinacia oleracea</i> , rps 19' gene	SPICPTGH.PE1	345
Chl, <i>Spinacia oleracea</i> , ribosomal protein L2	SPICPTGI.PE1	861
Chl, <i>Spinacia oleracea</i> , ribosomal protein S19	SPICPTGI.PE2	279
Chl, <i>Nicotiana tabacum</i> , ribosomal protein S16	TOBCPCG.PE2	258
Chl, <i>Nicotiana tabacum</i> , ATPase III subunit	TOBCPCG.PE5	246
Chl, <i>Nicotiana tabacum</i> , ribosomal protein S2	TOBCGCG.PE7	711
Chl, <i>Nicotiana tabacum</i> , RNA polymerase beta subunit	TOBCPCG.PE8	3213
Chl, <i>Nicotiana tabacum</i> , ribosomal protein S14	TOBCPCG.PE11	303
Chl, <i>Nicotiana tabacum</i> , psI p700 apoprotein A2	TOBCPCG.PE12	2205
Chl, <i>Nicotiana tabacum</i> , ribosomal protein S4	TOBCPCG.PE14	606
Chl, <i>Nicotiana tabacum</i> , ribosomal protein L33	TOBCPCG.PE21	201
Chl, <i>Nicotiana tabacum</i> , ribosomal protein S18	TOBCPCG.PE22	306
Chl, <i>Nicotiana tabacum</i> , ribosomal protein L20	TOBCPCG.PE23	387
Chl, <i>Nicotiana tabacum</i> , ribosomal protein S12	TOBCPCG.PE24	372
Chl, <i>Nicotiana tabacum</i> , ribosomal protein S11	TOBCPCG.PE31	417
Chl, <i>Nicotiana tabacum</i> , RF 96	TOBCPCG.PE32	291
Chl, <i>Nicotiana tabacum</i> , ribosomal protein S8	TOBCPCG.PE33	405
Chl, <i>Nicotiana tabacum</i> , ribosomal protein L14	TOBCPCG.PE34	372
Chl, <i>Nicotiana tabacum</i> , ribosomal protein L16	TOBCPCG.PE35	405
Chl, <i>Nicotiana tabacum</i> , ribosomal protein S3	TOBCPCG.PE36	657
Chl, <i>Nicotiana tabacum</i> , ribosomal protein L22	TOBCPCG.PE37	468
Chl, <i>Nicotiana tabacum</i> , ribosomal protein S7	TOBCPCG.PE41	468
Chl, <i>Nicotiana tabacum</i> , ribosomal protein S15	TOBCPCG.PE42	264
Chl, <i>Nicotiana tabacum</i> , ribosomal protein L23	TOBCPCG.PE44	282
Chl, <i>Nicotiana tabacum</i> , ribosomal protein CS19	TOBCPRPS.PE1	279
Chl, <i>Nicotiana tabacum</i> , proton-translocating ATPase alpha subunit	TOBCPTGG.PE1	138
Chl, <i>Triticum aestivum</i> , ATP synthase proton-translocating subunit	WHTCPATP.PE1	246
Chl, <i>Triticum aestivum</i> , CF-O subunit I	WHTCPATPS.PE2	552
Chl, <i>Triticum aestivum</i> , cytochrome b559	WHTCPCYTB.PE1	252
Chl, <i>Triticum aestivum</i> , unidentified reading frame	WHTCPCYTB.PE2	120
Nuc, <i>Avena sativa</i> , Phytochrome type 3	ASTAP3R.PE1	3390
Nuc, <i>Avena sativa</i> , Phytochrome type 4	ASTAP4R.PE1	3390
Nuc, <i>Hordeum vulgare</i> , alpha-amylase type A	BLYAMYAA.PE1	1317

Nuc, <i>Hordeum vulgare</i> , prealpha-amylase type B	BLYAMYABD.PE1	1284
Nuc, <i>Hordeum vulgare</i> , B1 hordein precursor	BLYHORB.PE1	882
Nuc, <i>Brassica napus</i> , napin prepropeptide	BNASSPB.PE1	537
Nuc, <i>Canavalia ensiformis</i> , concanavalin A	CENCONA.PE1	873
Nuc, <i>Daucus carota</i> , extensin protein	DAREXT.PE1	921
Nuc, <i>Lemna gibba</i> , chlorophyll a/b apoprotein	LGIAB19.PE1	795
Nuc, <i>Zea mays</i> , actin	MZEACT1G.PE1	1128
Nuc, <i>Zea mays</i> , alcohol dehydrogenase	MZEADH1F.PE1	1140
Nuc, <i>Zea mays</i> , ATP/ADP translocator	MZEANT.PE1	957
Nuc, <i>Zea mays</i> , glutelin-2-precursor	MZEEG2R.PE1	672
Nuc, <i>Zea mays</i> , histone H3	MZEH3C2.PE1	411
Nuc, <i>Zea mays</i> , histone H4	MZEH4C14.PE1	312
Nuc, <i>Zea mays</i> , sucrose synthase	MZESUSYSG.PE1	2409
Nuc, <i>Zea mays</i> , zein 19kd protein	MZEZE19A.PE1	708
Nuc, <i>Zea mays</i> , zein 22kd protein	MZEZE22A.PE1	792
Nuc, <i>Zea mays</i> , zein 26.99kd protein	MZEZE22B.PE1	801
Nuc, <i>Zea mays</i> , zein	MZEZEA20M.PE1	723
Nuc, <i>Zea mays</i> , zein	MZEZEA30M.PE1	705
Nuc, <i>Zea mays</i> , zein	MZEZEZ4G.PE1	804
Nuc, <i>Pisum sativum</i> , albumin 1	PEAABN1.PE1	393
Nuc, <i>Pisum sativum</i> , albumin 2	PEAABN2.PE1	696
Nuc, <i>Pisum sativum</i> , major light-harvesting chlorophyll a/b-binding protein	PEACAB80.PE1	810
Mit, <i>Zea mays</i> , cytochrome c oxydase subunit I	MZEMTCOX1.PE1	1587
Mit, <i>Oenothera berteriana</i> , cytochrome oxydase subunit II	OBEMTCYO2.PE1	777
Mit, <i>Triticum aestivum</i> , apocytochrome b	WHTMTCYBB.PE1	1197

frequencies except in positions I-II where TpG is not elevated.

Organelles are never methylated, except in the *Chlamydomonas reinhardtii* chloroplast during gametogenesis (23-25). Nevertheless, we consistently find CpG avoidance in mitochondrial and chloroplast genes of higher plants. This avoidance is not generally compensated by TpG and CpA frequencies, however, as it is in plant nuclear genes (table 2) or vertebrate nuclear genes (1-3;6,9,10). It suggests other mechanisms for CpG suppression.

In chloroplasts, CpG is statistically favored in position I-II; in this position CpG corresponds to arginine codons CGN (N=A,C,G or U). Low CpG frequency in this position indicates either rarity of arginine or preference for its codons AGR (R= A ou G). In fact, the chloroplast sample is richer in arginine than is the nuclear one (7.6% and 3.8% respectively) and theoretical values for AGR and CGN are approximately respected in chloroplasts. In nuclear genes, nearly 47% of Arg residues are coded by AGR codons, the expected amount is 33.3% (2/6 for AGR and 4/6 for CGN codons). This abnormal high frequency of CpG in the third codon position

Table 2. CpG, TpA, TpG and CpA doublet normalized frequencies by codon position in nuclear, chloroplast and mitochondrial genes of plants for the whole workfile. Values in parenthesis are mean values and top values are cumulated test values on each type of sequence (see methods). Positive values indicate doublets of higher than expected frequency (obtained from permutations conserving base composition in each position). Negative values reveal avoided doublets. ns, not significant. Absolute values of cumulated test values >1.96 are statistically significant at 5%; absolute values >2.57 are significant at 1% (see reference 21).

	TpA	CpG	TpG	CpA	
Nuc	-9.82 (-1.96)	-9.83 (-1.97)	ns (0.08)	6.35 (1.27)	I-II
Chl	-9.56 (-1.55)	7.13 (1.16)	-8.81 (-1.43)	ns (0.26)	
Mit	ns (-1.12)	-2.10 (-1.21)	ns (-0.83)	3.34 (1.93)	
Nuc	-12.22 (-2.44)	-12.79 (-2.56)	6.09 (1.22)	9.17 (1.83)	II-III
Chl	-4.22 (-0.68)	-5.19 (-0.84)	5.10 (0.83)	-5.99 (-0.97)	
Mit	ns (-0.24)	2.56 (-1.48)	ns (0.82)	2.71 (1.56)	
Nuc	-9.86 (-1.97)	-12.43 (-2.49)	16.40 (3.28)	10.69 (2.14)	III-I
Chl	-3.72 (-0.60)	ns (-0.04)	-5.43 (-0.88)	ns (0.10)	
Mit	-3.39 (-1.96)	-2.97 (-1.71)	ns (-0.71)	3.85 (2.23)	
Nuc	-18.67 (-3.73)	-20.50 (-4.10)	14.54 (2.91)	15.51 (3.10)	Total
Chl	-9.73 (-1.58)	ns (0.18)	-5.43 (-0.88)	-2.55 (-0.41)	
Mit	-3.30 (-1.91)	-4.46 (-2.57)	ns (-0.44)	5.66 (3.27)	

Nuc, nuclear; Chl, chloroplast; Mit, plant mitochondrial; See Table 1 for names of sequences analyzed.

can be explained by amino acid composition and codon usage (8). In our chloroplast sequence sample 76% of the overall C+G content is in the first and second position of codons. In chloroplasts, CpG is avoided in position II-III (Table 2). In mitochondria we observe CpG avoidance in all positions, correlated with high usage of CpA doublets. Comparison of the mean values in each workfile reveals a higher avoidance of CpG in plant nuclear sequences than in those of chloroplasts.

Table 3. Normal statistics in intron and rRNA sequences of plants and chloroplasts; frequencies were normalized by reference to permutations conserving overall base composition of each real sequence. The value in parenthesis is the mean and the top value the cumulated test value for the sequences. The mean permits a comparison between different types of sequences that is not influenced by the number of sequences. All values are significant at 1% except that for CpG in the smallest sample, rRNA sequences of plants. See table 2 for abbreviations.

	Introns		rRNA	
	Nuc	Chl	Nuc	Chl
No. sequences	77	16	4	9
No. bases	17,509	109,210	6,903	17,537
TpA	-9.5 (-1.1)	-19.9 (-5.0)	-3.4 (-1.7)	-4.2 (-1.4)
CpG	-12.6 (-1.4)	-5.5 (-1.4)	ns (0.0)	-6.0 (-2.0)

TpA avoidance

TpA is highly avoided in most species (6,9-11). We find this avoidance in all codon positions in higher plants, but only in position I-II of chloroplasts, and in positions I-II and III-I of plant mitochondria. Usually, rarity of this doublet is attributed to absence of stop codons TAA and TAG; in fact, TpA occurrence is low in position I-II also because tyrosine (coded by TAY, where Y= C or T) is often of relatively low frequency in chloroplast proteins (3.0% while 3.3% is expected from the number of codons).

TpA in position II-III (NTA triplets) involves 4 codons and 3 amino acids: Leu, Val and Ile. These aminoacids make up 23.9% and 22.0% of chloroplast and nuclear gene proteins, respectively. NTA codons represent for nuclear genes, 15.1%, and for chloroplast genes, 38.2%, generally agreeing with degenerate base usage in the 2 types of genes. TpA is not avoided in position II-III of chloroplast genes. TpA in position III-I involves codon couples NNT-ANN. NNT codons are often used and ANN codons correspond to abundant amino acids in chloroplasts (8).

These results help us to better understand dinucleotide frequencies, but cannot resolve the problem of TpA and CpG rarity in non-coding sequences. As shown in table 3, both dinucleotides are strongly avoided in untranslated regions of plant nuclear and chloroplast sequences.

DISCUSSION

This study has revealed that difficult problems remain in rationalizing dinucleotide frequencies in the plant system. Plant nuclear DNA, like that in vertebrates, is methylated and mainly on cytosine residues in CpG doublets (26). This methylation is accepted as producing mutationally unstable mCpG, which deaminates to TpG, resulting in a negative correlation between CpG and TpG (or CpA) frequencies, in plant nuclear sequences. Our results reveal that avoidance of CpG, though less strong, exists in unmethylated sequences of chloroplasts and mitochondria. It seems, then, that other factors are involved, particularly in the CpG suppression seen in organelles.

As seen above, TpA has a curious pattern of avoidance in most species, being as rare in plant nuclear genes as is CpG; this is observed in all codon positions and in rRNA and intron sequences. The same is true of chloroplast untranslated regions.

We find that coding strategy and amino acid composition influence dinucleotide frequencies. However, we note that these two doublets correspond to the highest and lowest values of free energy of helix disruption according to Breslauer et al. (27); This energy being only 0.9 kcal/mol for TpA and 3.6 kcal/mol for CpG, while the other 14 doublets all fall between 1.3 and 3.1 kcal/mol. High frequencies of these two doublets could be disadvantageous for the stability of DNA and their avoidance could represent a mechanism for thermodynamically optimizing sequences.

ACKNOWLEDGEMENTS

We thank C. Gautier, T. Greenland and R. Grantham for help. One of us (P.P.) thanks the Fondation pour la recherche médicale for financial support.

REFERENCES

1. Salser, W., (1977) Cold Spring Harbor Symp. Quant. Biol. 42, 985-1002.
2. Bird, A.P. (1980) Nucleic Acids Res. 8, 1499-1504.
3. Cooper, D.N. and Gerber-Huber, S. (1985) Cell Differentiation 17, 199-205.
4. Grantham, R. (1978) FEBS letters 95, 1-11.
5. Grantham, R. (1985) Nature 313, 437.
6. Gautier, C. (1987) Thèse d'état Lyon I, 23 Feb. 1987.
7. Bernardi, G., Olofsson B., Filipinski J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., and Rodier, F., (1985) Science 228, 953-958.
8. Boudraa, M. (1987) Genet. Selec. Evol. 2 (in press).
9. Grantham, R., Greenland, T., Louail, S., Mouchiroud, D., Prato, J.L., Gouy, M., and Gautier, C. (1985) Bull. Inst. Pasteur 83, 95-148.

-
10. Grantham, R., Perrin P., and Mouchiroud D. (1986) Oxford Surveys in Evolutionary Biology, 3, 48-81.
 11. Grantham, R., (1980) FEBS Letters 121, 193-199.
 12. Max, E.E. (1984) Nature 310, 100.
 13. McClelland, M. and Ivarie, R. (1982) Nucleic Acids Res. 10, 7865-7877.
 14. Nussinov, R. (1984) Nucleic Acids Res. 12, 1749-1763.
 15. Smith, T.F., Waterman, M.S. and Sadler, J.R. (1983) Nucleic Acids Res. 11, 2205-2220.
 16. Min Jou, W., Verhoeyen, M., Devos, R., Saman, E., Fang, R., Huylebroeck, D., Fiers, W., Threlfall, G., Barber, C., Carey, N., Emtage, S. (1980) Cell 19, 683-696.
 17. Gouy, M., Milleret, F., Mignier, C., Jacobzone, M. and Gautier, C. (1984) Nucleic Acids Res. 12, 121-127.
 18. Gouy, M., Gautier, C. and Milleret, F. (1985) Biochimie 67, 433-436.
 19. Gouy, M., Gautier, C., Attimonelli, M., Lanave, C. and di Paola, G. (1985) CABIOS 1, 167-172.
 20. Tykocinski, M.L. and Max, E.E. (1984) Nucleic Acids Res. 12, 4385-4396.
 21. Gautier, C., Gouy, M. and Louail, S. (1985) Biochimie 67, 449-543.
 22. Knight, W. (1974) Biometrics 30, 553-555.
 23. Lindahl, T. (1981) Nature 290, 363-364.
 24. Dyer, T.A. (1982) Nature 298, 422-423.
 25. Sano, H., Grabowy, C. and Sager, R. (1984) Mol. Cell. Biol. 40, 2103-2108.
 26. Bird, A. (1986) Nature 321, 209-213.
 27. Breslauer, K.J., Ronald, F., Blocker, H. and Marky L. A. (1986) Proc. Natl. Acad. Sci. USA 83, 3746-3750.