Volume 8 Number 7 1980 Nucleic Acids Research

DNA methylation and the frequency of CpG in animal DNA

Adrian P.Bird

MRC Mammalian Genome Unit, Kings Buildings, West Mains Road, Edinburgh EH9 3JT, UK

Received 5 February 1980

ABSTRACT

An analysis of nearest neighbour dinucleotide frequencies and the level of DNA methylation in animals strongly supports the suggestion that 5-methylcytosine (5mC) tends to mutate abnormally frequently to T. This tendency is the likely cause of the CpG deficiency in heavily methylated genomes.

INTRODUCTION

Since the early nearest neighbour sequencing studies of Josse et al (1) and Swartz et al (2), it has been recognised that in many animals the dinucleotide sequence CpG is present in the DNA less frequently than would be expected from base composition. Thus in human DNA, where the fraction of (G+C) is 0.4, we would expect CpG to occur with a frequency of $0.2 \times 0.2 = 0.04$, whereas the observed frequency is about 0.008. Salser (3) has suggested that the CpG deficiency is related to DNA methylation, since mCpG is the major methylated sequence in animals. The present work provides support for this suggestion by relating known CpG frequencies to recent comparative data on levels of DNA methylation (4,5). Furthermore, by an analysis of nearest neighbour frequencies alone, it emerges that the basis for the relationship is the tendency for mCpG to mutate to TpG.

METHODS

Levels of CCGG methylation were estimated by comparing Hpa II and Msp I digests of the DNA under study (6,7). The experimental procedures used, and examples of the gels obtained, have been presented previously (4). For vertebrate DNAs, the average molecular weights of the digests were compared in order to estimate what fraction of available CCGG was left uncut by Hpa II. This approach was not applicable to non-arthropod invertebrates, however, since methylated and unmethylated fractions each comprised a significant

fraction of the total DNA. In these cases the proportion of DNA remaining at high molecular weight after Hpa II digestion was considered equal to the fraction of fully methylated Hpa II sites, as is the case for Echinus (8).

RESULTS AND DISCUSSION

Within the animal kingdom DNA methylation ranges from very low levels in arthropods, through intermediate levels in many non-arthropod invertebrates, to high levels in the vertebrates (4,5,9-12). These degrees of methylation have been classified on the basis of restriction enzyme studies as "insect-type" (indetectable methylation), "echinoderm-type" (partial methylation), and "vertebrate-type" (heavy methylation)[4]. The classification was derived using restriction enzymes Hpa II and Hha I, and so refers only to the sequences CCGG and GCGC respectively. There are, however, reasons for believing that the methylation patterns established with these enzymes apply equally to all CpGs regardless of their sequence setting, since the susceptibility of a CpG to methylation appears to be independent of the surrounding nucleotide sequence (for full discussion see reference 8).

We noticed that organisms with the most extreme CpG deficiency (i.e. the vertebrates (13)) also have the highest levels of DNA methylation. Conversely, poorly methylated genomes (i.e. insect-type) display no significant CpG deficiency (14; Russell and Subak-Sharpe, unpublished results), while partially methylated genomes are deficient in CpG to an intermediate extent (2,15). The relationship for those organisms in which both methylation data and nearest neighbour data are available is presented graphically in Figure 1. The result supports the idea that DNA methylation causes the CpG deficiency (3), because CpG is only infrequent in genomes that are partially or heavily methylated. It also encourages the belief that the methylation patterns established for Hpa II and Hha I sites are typical of CpGs flanked by other nucleotide sequences, since it is unlikely that the overall CpG frequency would correlate with methylation of the small proportion of CpGs detected with these enzymes alone.

A plausible molecular explanation for the relationship is suggested by the results of a study of mutations in E. coli. Coulondre et al (16) have elegantly shown that in the I gene of the lac operon mutational "hotspots" are caused by an abnormally high rate of mutation from 5mC to T. We have made use of the available nearest neighbour sequence data in order to test whether 5mC mutations could account for the CpG deficiency in methylated animal DNA. If, over evolutionary time, mCpG were to mutate relatively

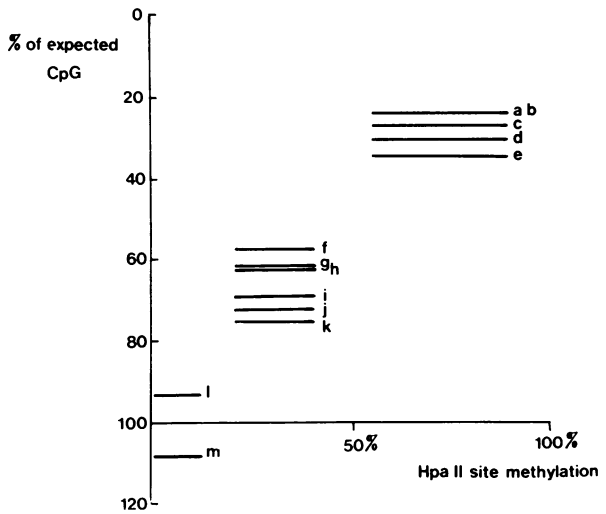

Figure 1. A correlation between CpG deficiency and the level of Hpa II site methylation in various animal DNAs. Levels of methylation are expressed as lines rather than points because of difficulties in accurately quantitating differences between Hpa II and Msp I ethidium bromide staining patterns. CpG deficiency has been expressed as a percentage of the expected frequency calculated from the base composition of the DNA concerned. The figures are taken from a collection of nearest neighbour data made by Setlow (26), and also from reference 15 and unpublished data of G. Russell, D. McGeoch and J. Subak-Sharpe (bee, fruit-fly and sea anemone). (a) man, (b) chick, (c) mouse, (d) rabbit, (e) BHK cells (hamster), (f) starfish, (g) sea urchin (Echinus), (h) sea urchin (Paracentrotus), (i) sea anemone, (j) sea squirt, (k) sea cucumber, (l) fruit fly, (m) bee.

frequently compared to other dinucleotides, the observed deficiency of CpG should be matched by a corresponding accumulation of the complementary dinucleotides TpG and CpA (Figure 2a). Since one 5mC change would cause loss of two CpGs and the gain of one TpG and one CpA, this should be discernable as a correlation between CpG deficiency and TpG plus CpA excess. Figure 2b shows that in animals with varying levels of DNA methylation such a correlation does exist. Genomes low in CpG (vertebrates) are high in TpG plus CpA to a roughly equivalent extent, whereas genomes whose CpG frequency is close to the value calculated from base composition (insects) have a normal TpG plus CpA frequency. This result provides strong evidence for the conversion of mCpG to TpG plus CpA during evolution. In addition, it shows that the excess of TpG:CpA in vertebrates is as dramatic as the deficiency of CpG:CpG, but has escaped notice because CpG is a self-complementary

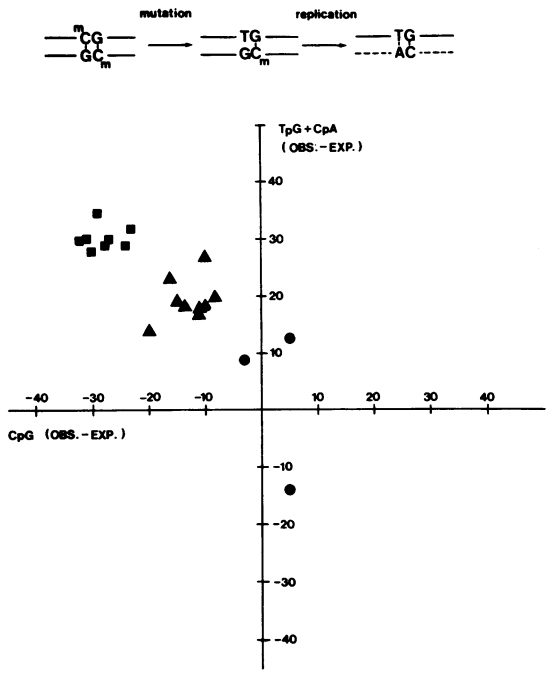

Figure 2. Evidence that mCpG tends to mutate to TpG. (A) Scheme of mutation from mCpG to TpG and CpA. Mutation of 5mC initially causes a mismatched T-G pair, which, if not repaired, gives rise to a T-A pair after DNA replication. Two CpG doublets are lost and one TpG plus one CpA are gained. (B) The correlation between CpG deficiency and TpG + CpA excess. Doublet frequencies are expressed as the difference between the observed frequency per one thousand dinucleotides and the expected frequency calculated from base composition. For CpG this is (CpG) obs - (CpG) exp. For TpG + CpA the figure is [(TpG) obs - (TpG) exp] + [(CpA) obs - (CpA) exp]. The figures were calculated from sources cited in the legend to Figure 1. Squares, vertebrates; triangles, non-arthropod invertebrates (including coelenterates, molluscs, brachiopods, echinoderms and invertebrate chordates); circles, arthropods (all insects).

dinucleotide whose deviations from the expected frequency are consequently doubled. The frequency of CpG does not correlate well with other dinucleotide frequencies. Also TpG and CpA are nearly always the most excessive over expectation of any dinucleotides in the DNA of non-arthropod animals.

The possibility that mCpG is evolutionarily unstable by virtue of its methylation has been proposed previously to explain the CpG deficiency in vertebrate DNA (3). Also, Comings (17) has speculated that the low average G + C content of vertebrate DNA is related to the high rate of 5mC mutation.

Others have taken a different viewpoint, suggesting that CpG, whether methylated or not, is discriminated against in coding DNA by natural selection, and therefore tends to be eliminated during evolution (18,19). Our results strongly support the suggestion of Salser (3) that methylation of CpG renders this dinucleotide unusually mutable, and further, they argue that the direction of mutation is mainly from 5mC to T. Selective discrimination against CpG in coding DNA, though possible, now seems a less important cause of the observed CpG deficiency in vertebrates.

It follows from the results discussed above that mCpG will be a mutational hotspot wherever it occurs. This expectation is given preliminary support by an examination of closely related 5S RNA sequences. Between the oocyte-type and somatic-type 5S RNAs of Xenopus laevis there are six nucleotide changes, two of which involve interconversion of TpG and CpG (positions 30 and 47) (20-22). Although we cannot determine in which direction the mutations have occurred, the observations that all six CpGs in the oocyte-type coding sequence are normally methylated (23), and that CpG appears to be involved in the divergence of these RNAs abnormally frequently, are both consistent with the possibility that mCpG has mutated to TpG. Further evidence for the instability of mCpG has emerged from studies of the chicken ovalbumin genes (24). Certain Hha I sites in this region were found to be either present, absent or apparently heterozygous in different individuals.

More than 10 years ago Scarano and co-workers proposed a mechanism for cell differentiation involving deamination of 5mC to T during development (25). While this suggestion foreshadowed the findings presented here, it should be stressed that there have so far been no clear demonstrations that such programmed 5mC to T mutations do occur during development.

Might the function of DNA methylation be to increase the mutation rate? It is difficult to argue conclusively for or against this possibility. One point against it, however, is that CpG would become progressively rarer in the DNA, thereby cancelling out the importance of its mutability. An alternative possibility is that the function of DNA methylation lies in some other direction, whose advantages outweigh the attendant disadvantage of high mutability. What that function might be is not yet clear.

ACKNOWLEDGEMENTS

I am grateful to Graham Russell for making available unpublished nearest neighbour frequencies, and to Ed Southern and Peter Walker for reading the manuscript.

REFERENCES

1. Josse, J., Kaiser, A.A. and Kornberg, A. (1961). J. Biol. Chem. 236, 864-875.
2. Swartz, M.N., Trautner, T.A. and Kornberg, A. (1962). J. Biol. Chem. 237, 1961-1967.
3. Salser, W. (1977). Cold Spring Harbour Symp. Quant. Biol. XLII, 98-1103.
4. Bird, A.P. and Taggart, M.H. (1980). Nucl. Acids Res. 8, 1485-1497
5. Rae, P.M.M. and Steele, R.E. (1979). Nucl Acids Res. 6, 2987-2995.
6. Waalwijk, C. and Flavell, R.A. (1978). Nucl Acids Res. 5, 3231-3236.
7. Singer, J., Roberts-Ems, J. and Riggs, A.D. (1979). Science 203, 1019-1021.
8. Bird, A.P., Taggart, M.H. and Smith, B.A. (1979). Cell 17, 889-901.
9. Wyatt, G.R. (1951). Biochem. J. 48, 584-590.
10. Antonov, A.S., Favorova, O.O. and Belozersky, A.N. (1962). Dokl. Akad. Nauk. SSSR 147, 1480-1481.
11. Vanyushin, B.F., Racheva, S.G. and Belozersky, A.N. (1970). Nature 225, 948-949.
12. Chargaff, E., Lipshitz, R. and Green, C. (1952). J. Biol. Chem. 195, 155-160.
13. Russell, G.J., Walker, P.M.B., Elton, R.A. and Subak-Sharpe, J.H. (1976). J. Mol. Biol. 108, 1-23.
14. McGeoch, D.J. (1970). Ph.D. Thesis, Institute of Biochemistry, University of Glasgow.
15. Russell, G.J. and Subak-Sharpe, J.H. (1977). Nature 266, 533-536.
16. Coulondre, C., Miller, J.H., Farabough, P.J. and Gilbert, W. (1978). Nature 274, 775-780.
17. Comings, D.E. (1972). Exp. Cell Res. 74, 383-390.
18. Subak-Sharpe, J.H., Burk, R.R., Crawford, L.V., Morrison, J.M., Hay, J. and Keir, H.M. (1966). Cold Spring Harbour Symp. Quant. Biol. 31, 737-747.
19. Elton, R.A. (1975). J. Molec. Evol. 4, 323-346.
20. Brownlee, G.G., Cartwright, E., McShane, T. and Williamson, R. (1972). FEBS Lett. 25, 8-12.
21. Wegnez, M., Monier, T. and Denis, H. (1972). FEBS Lett. 25, 13-20.
22. Ford, P.J. and Southern, E.M. (1973). Nature New Biol. 241, 7-10.
23. Miller, J.R., Cartwright, E.M., Brownlee, G.G., Fedoroff, N.V. and Brown, D.D. (1978). Cell 13, 717-725.
24. Mandel, J.L. and Chambon, P. (1979). Nucl. Acids Res. 7, 2081-2103.
25. Scarano, E., Iaccarino, M., Grippo, P. and Parisi, E. (1967). Proc. Nat. Acad. Sci. 57, 1394-1400.
26. Setlow, P. In CRC Handbook of Biochemistry and Molecular Biology: Nucleic Acids (Ed. G.D. Fasman, Volume 2, 32). CRC Press, Cleveland, Ohio.