Volume 15 Number 23 1987 Nucleic Acids Research

Thymine methyls and DNA–protein interactions

Robert Ivarie

Department of Genetics, University of Georgia, Athens, GA 30602, USA

Received September 14, 1987; Revised and Accepted November 4, 1987

SUMMARY

Evidence is summarized showing that thymine methyls are as important in the recognition of specific sequences by proteins as are the more widely recognized hydrogen bonding sites of bases in the major groove (1). Strongest evidence has come from experiments using functional group mutagenesis (2) in which thymines in a specific recognition sequence (e.g., promoters, operators and restriction sites) are replaced by oligonucleotide synthesis with methyl-free uracil or cytosine and 5-methylcytosine. Such experiments have shown that thymine methyls can provide contact points via van der Waals interactions with amino acid side chains of specific DNA binding proteins. Actual contact between a thymine methyl and carbons of a glutamine side chain has been observed in a cocrystal of the phage 434 repressor and its operator by X-ray analysis. The issue of why thymine occurs in DNA is discussed in light of these findings.

INTRODUCTION

It comes as no surprise to biologists that virtually all organisms contain thymine in their DNA. What is often overlooked, however, is that DNA is heavily methylated as a consequence via the C5-methyls of the pyrmidine. Despite this enormous background of natural DNA methylation, low level methylation of cytosine C5 (4-7% in higher eukaryotes) and adenine N6 (up to 3% in E. coli) catalyzed by specific DNA methylases after replication can exert profound effects on the structure and function of the genome, e.g., transcription, recombination and repair (reviewed in 3). Methyl groups at adenine N6 and pyrimidine C5 occur in the major groove of B DNA where the vast majority of contacts are made between DNA and site-specific binding proteins (reviewed in 4). It is not surprising, therefore, that thymine methyls in some sequences via hydrophobic interactions with amino acid side chains of DNA binding proteins should exert effects on gene function comparable to those caused by the other two methylated bases.

Thymines and Restriction Endonuclease Activity

That pyrimidine methyls interfere in the recognition of specific sequences by proteins has long been known from analaysis of microbial restriction/modification systems. Methylation of cytosine C5 in a variety of restriction enzyme sites prevents cleavage by the endonucleases (see 5). In terms of structure, cytosines are methylated symmetrically in palindromic recognition sequences and introduce a pair of methyls across the major groove in the helix. The vertebrate methylation dinucleotide 5'-CpG-3' is also symmetrically methylated and two T-containing dinucleotides, TpA and ApT, also place symmetrical methyls across the major groove.

The effect of thymine methyls on restriction enzyme activity was initially analyzed on DNA from Bacillus subtilis phage PBS2 (6,7). PBS2 DNA contains uracil in place of thymine and is, therefore, unmethylated at all pyrmidine C5's. Enzymes having only G-C in their recognition sites (HpaII and HhaI) cut PBS2 DNA as readily as they cut thymine-containing DNA. However, three enzymes with A-T in their recognition sites (HpaI, GTTAAC; HindII, GTPyPuAC; and HindIII, AAGCTT) cut the phage DNA inefficiently while EcoRI (GAATTC) and BamHI (GGATCC) activities were unaffected. PBS2 DNA was also a poor substrate for EcoRI methylase.

Although these observations implied a role for thymine methyls in the site-specific interactions, restriction of the unmethylated DNA could have been influenced by the absence of thymines external to the recognition sites. This problem was addressed by functional group mutagenesis in which thymines in two restriction enzyme sites, GAATTC (EcoRI) and GATATC (EcoRV), were individually replaced with uracil (8) and the sites assayed for cleavage by the enzymes. EcoRI activity was slightly altered in that the enzyme cut uracil-containing sites with a 30% decrease in rate. Recent X-ray analysis of cocrystals of EcoRI and its recognition sequence at 3 Å resolution has revealed the absence of hydrophobic contacts between the four central thymine methyls and amino acids in the active site (9). By contrast, EcoRV could not cut its site when either T was replaced by U. A greater than 50% reduction in the rate of cleavage has also been found for EcoRII on a chemically synthesized substrate (CC[U/A]GG); the enzyme bound weakly to the unmethylated substrate and had a substantially increased rate of dissociation (10).

Taken together, these observations support the idea that thymine methyls may be important contact points in some restriction enzyme sites. They also show that the C5 methyls of thymine exert an effect opposite to the C5

methyls of cytosines in at least one site. Cleavage by EcoRII is blocked by methylation of the external C in the recognition sequence whereas methyls of the internal thymines enhance enzyme activity. Dissimilar effects on activity most likely depends on the position of the methyls within the site and therefore, with respect to the protein. As noted below, for the lac repressor, 5-methylcytosines can have the same effect as thymine methyls and vice versa at the same position.

Thymine Methyls in Operators

Although it has been recognized that thymine methyls are important contact points for repressors and RNA polymerase in operator and promoter sites, respectively, the experimental method for detecting the interactions between C5 methyls and amino acid side chains is indirect. A DNA site substituted with 5-bromouracil (5-BU) is subjected to UV light before and after binding the protein (11). UV cleaves bromine from the C5 position of deoxyuridine leading to a single-strand break at the site in the absence of bound protein or, presumably, to a crosslink with a neighboring amino acid side chain in the presence of bound protein. Hence, 5-BU-substituted sites are protected from strand scission by bound proteins from which it is inferred that thymine methyls themselves lie in close proximity to the protein and represent points of contact. However, contact with thymine methyls is not actually demonstrated. Hydrogen bonding sites in an A-T basepair at the site could supply the contact point(s) and indirectly quench the cleavage reaction by other means (12).

There are, nonetheless, unequivocal examples showing that the pyrimidine C5 methyl is the chemically important group at the A-T site. The lac operator is 22 bp in length of which 15 are A-T. UV photochemical cleavage experiments identified 12 thymines as crosslinking sites (11). By functional group mutagenesis, Caruthers and his colleagues (13,14) replaced the thymine at position 13 with uracil or cytosine and found that in vitro binding of repressor to the substituted operator was strongly inhibited. Binding by repressor was restored, however, when the site contained 5-methylcytosine. Hence, the methyl was the essential group recognized by the repressor. The lac operator also contains a symmetrical thymine at position 19 because of the dyad symmetry of the site, but thymine-19 is apparently unimportant in contacting repressor side groups (13,14). Furthermore, a complex series of base analog substitution experiments in lac operator (14) have shown that 4 other thymine methyls (positions 6,7 + 25,26) make hydrophobic contacts with repressor side chains but the contacts are likely to be weak.

Model building and chemical protection experiments have also suggested that the methyl of alanine-49 in the lambda CI repressor makes van der Waals contacts with thymine methyls at +3 and -5 in the O_R1 operator (4). Although not yet shown for the lambda repressor, X-ray crystallographic analysis of cocrystals of the phage 434 repressor bound to its A-T rich operator (ACAATATATATTGT) at 3.2-4.5 Å resolution has shown that β and γ carbons of glutamine-29 form van der Waals contacts with the methyl of thymine-12 (15). Moreover, a glutamine-to-alanine mutation at residue 28 in the 434 repressor abolished binding to the operator unless a compensating mutation was introduced in the operator (16). Glutamine-28 normally contacts adenine-1 by a bidentate hydrogen bond to N7 and N6. Replacing adenine-1 with thymine restored binding of the mutant repressor most likely via a van der Waals bond between the methyls of alanine-28 and thymine-1. Replacing the thymine with uracil or cytosine abolished mutant repressor binding while 5-methylcytosine substitution allowed detectable but reduced binding of the mutant repressor.

Thymine Methyls and E. coli Promoters

E. coli promoters are A-T rich and the -35 (consensus: TTGACA) and -10 (TATAAT) sequences each contain two "invariant" thymines, noted in bold. The first 2 thymines in the -35 sequence occur 92 and 94 times, respectively, while the second and last thymines in the -10 sequence occur 106 and 108 times, respectively, in 112 promoters (17). Less well conserved sites external to -35 and -10 sequences are more often an A-T than a G-C bp. DNAse I footprints of RNA polymerase bound to promoters extend for ~70 bp beginning upstream from the -35 region to well past the transcription start site, and the -35 and -10 thymines are protected by bound RNA polymerase (18). Photochemical cleavage of the lac UV5 promoter substituted with 5-BU has shown that at least 10 thymine methyls are covered by RNA polymerase (12,19). Three occur in the -35 site including the two invariant thymines, four occur in the -10 site including one of the invariant thymines, and three occur in the transcription start site region which is also A-T rich. Furthermore, two thymine methyls in the initiation site have been shown to form crosslinks with α and β subunits of RNA polymerase (20).

Using functional group mutagenesis, Caruthers and his colleagues (2) systematically replaced 12 thymine methyls in the lambda PR promoter with uracil and analyzed the ability of RNA polymerase to bind functionally to the modified sites. Only two thymine methyls at -34 and -35 were crucial for promoter function when assayed by either run-off transcription or abortive initiation. Hence, uracil at 10 other thymine sites had little to no effect

on promoter activity. Single replacement of thymines with uracil at -34 and -35 increased the lag time of open complex formation 4- to 5-fold while double substitution completely abolished the ability of the promoter to function in both assays. Double substitution of the -34/-35 thymines with 5-methylcytosine did not restore promoter function, perhaps because of crosschain clashing of guanines in the minor groove. As they point out, the critical substitution would be 5-methylcytosine/inosine at the sites.

These experiments with restriction enzymes, repressors and RNA polymerase interacting with their recognition sequences have several implications. First, photochemical cleavage experiments do not necessarily show that thymine methyls are contact points. Nonetheless, those thymine methyls that are contact points are protected from photochemical cleavage. Second, only a fraction of the thymine methyls in A-T rich binding sequences are likely to be strong contact sites. Third, invariant A-T basepairs in the -10 region must contain contact points other than the C5 methyl that are not found in a C-G bp (e.g., thymine 4-carbonyl and adenine 6-amino). Fourth, apart from their C5 methyls, A-T rich sequences have two other properties absent from G-C rich sequences: they are more readily denatured and can bend more easily providing the right sequence combination occurs (21). Denaturability is especially relevent to proteins that interact with and melt DNA. For example, RNA polymerase denatures 12 bp in the promoter from the middle of the -10 region through the transcription start site during open complex formation (19).

Other DNA Sequences

Although a long list of A-T rich sequences can be compiled in which thymine methyls may serve as contact points, at least a few are worth mentioning for comparison:

1. Site-specific recombination of bacteriophage lambda into the 4.2×10^6 bp chromosome of *E. coli* takes place within an A-T rich, 15 bp sequence (GCTTTTTTATACTAA) found once in both bacterial and phage DNA (22). Recombination takes place within the string of thymines.

2. Many promoters in nonhousekeeping genes of higher eukaryotes harbor A-T rich sequences at -80/-100 (CAAAT) and at -25/-35 (TATAAA) both of which are required for normal transcription by RNA polymerase II (23).

3. Two A-T rich sequences have been found in clusters in the 5' and 3' flanking regions of several genes of *Drophosphila melanogaster*

(24). Fifteen A-rich boxes with the sequence AATAAA(T/A)AAA and nine T-rich boxes with the sequence TT(A/T)T(T/A)TT(T/A)TT were specifically bound to the nuclear scaffold proteins.

4. Replication of the E. coli chromosome is initiated at the 255 bp oriC sequence that is 56% A+T (25). Phylogenetic comparison among 6 bacterial species has shown conservation at 129 bp of which 67% are A-T bp. E. coli DnaA protein is required for initiation of replication and binds to oriC at A-T rich "R" sequences, TTAT(C/A)CA(C/A)A. Furthermore, 9 Dam methylation sites (GATC) occur in oriC and full methylation is required for oriC activity.

The Dam Methylation Site

When the adenines are methylated, the Dam site is structurally unusual in that four methyls span the major groove across the internal ApT dinucleotide. Furthermore, methylation of N6 adenine eliminates a hydrogen bond donor site (1). Genetic dissection of Salmonella phage p22 Mnt repressor/operator interactions has uncovered an unusual repressor mutation whose binding to the operator can be restored by symmetric G:C->A:T transitions in the 17 bp operator (26). A his -> pro mutation in the Mnt repressor bound wild-type operator 1000-fold less efficiently; full binding capacity was restored by two symmetric transition mutations in the operator creating symmetric Dam methlyation sites and subsequent N6-adenine methylation. The authors proposed that the planar ring of proline can make van der Waals contacts in the major groove on the surface of the four methyls occurring in the methylated Dam site.

Partial Substitution of Thymine with Uracil Is Not Lethal in E. coli

If, as the foregoing implies, thymine methyls are important for gene regulation and promotor recognition, replacing thymines with uracils in the E. coli chromosome should be lethal to the bacterium. Although 100% replacement has not been achieved, dut ung double mutants deficient in dUTPase and N-uracil-DNA glycosylase can replace up to 20% of the thymines with uracil (27). This level of substitution reduced the growth rate by 50%. T4 DNA substituted to a comparable extent with uracil did not lead to discernible effects on T4 infection (28). At face value, these observations imply a relatively minor role for thymine in overall gene structure and function. However, T4 DNA is heavily modifed with all cytosines containing a bulky glucosyl-hydroxymethyl group at C5 and the virus encodes most enzymes necessary to replicate and express such DNA (29). Furthermore, only one or a few thymines per gene are likely to be the important ones to replace with

uracil so that the level of expression of most uracil-substituted genes would still be quite high, sufficiently high perhaps to enable the cell to complete another round of replication and resubstitute thymine in the sites. A more stringent test of the idea will be to measure the effects of uracil substitution on basal and induced enzyme levels, such as ß-galactosidase, where crucial thymine methyls have already been located with regard to promoter and operator function.

Why Does Thymine Occur in DNA?

The de novo pathway of thymine biosynthesis is peculiar with respect to the other four bases and suggests that in the course of cell evolution uracil may have antedated thymine in DNA. C, A and G enter deoxyribonucleotide pools from their ribonucleoside diphosphates by ribonucleotide reductase: NDP -> dNDP which is then converted to dNTP by a nonspecific dNDP kinase. Uracil is also converted by this pathway to dUTP but its incorporation into DNA is blocked by dUTPase which removes pyrophosphate leaving dUMP. dUMP is then methylated at C5 via thymidylate synthetase and converted to the diphosphate and triphosphate via deoxynucleotide kinases. Hence, thymine enters DNA de novo by an additional four enzymatic steps. If this reflects the fact that uracil was used in early genetic material in place of thymine, an important question is why thymine eventually replaced uracil. There are at least two conventional answers and I propose a third based on thymine methyls and sequence recognition.

The first explanation is based on the fact that methylation of uracil at C5 greatly enhances the resistance of the DNA to UV damage (30). In uracil-containing DNA, the major UV photoproduct is uracil-hydrate which has been reported to base-pair with guanine (31-35; but also see 36). Hence, U-containing DNA would be highly mutagenic, especially on oxygen-depleted primordial earth lacking an ozone layer to screen UV light. Thymine hydrate forms at a thousand-fold lower rate than uracil hydrate. Hence, including thymine in DNA would greatly have enhanced its UV-resistance. This argument gains plausibility given that bacterial species inhabiting high UV-flux environments have a low A + T content compared to those in low UV-incident habitats (37).

The second explanation centers on the need for organisms to repair G-U mismatches arising by cytosine deamination to uracil (38). Besides G-U mismatches, N-uracil-glycosylase excises uracil from an A-U but not from an A-T basepair. Apparently, methylation of C5 is sufficient to inhibit the enzyme. Replacement of uracil with thymine, therefore, would have enabled

the cell to evolve an efficient repair pathway for cytosine deamination.

A third reason may have to do with gene expression. That is, thymine methyls may have increased or decreased the ability of enzymes and/or regulatory proteins to bind to specific sequences. An example is the lac repressor which binds nonspecifically to poly(dA-dU) with 20-fold greater affinity than it does to poly(dA-dT) (39). If this were a general effect, then by reducing nonspecific binding alone, the inclusion of thymine in a primordial cell's genetic material would potentially increase the difference between specific and nonspecific binding constants. This difference, in turn, could have provided several selective advantages to the progenitor cell. For example, for genes controlled by repressors, the basal level of gene expression would be substantially reduced under noninducing conditions enabling more stringent control of expression. Similarly, fewer repressor molecules would be needed to locate operators in such a cell (40). Fewer repressors and basal enzyme molecules would represent an energy savings to the cell.

Introduction of thymine methyls in primordial genetic material would also have increased the number of ways in which and A-T(U) bp could be distinguished from a G-C bp. For example, at least two contacts per bp are required to distinguish one bp from any other (1). In uracil-containing DNA, pyrimidine C5's are unmethylated and incapable of hydrogen bonding. However, introducing a methyl at C5 of uracil would have allowed a new contact point for the A-T(U) bp via a van der Waals bond. Its methylation and proximity to purine N7 in thymine-purine dinucleotides would also potentially affect hydrogen bonding at the N7 site as well (1).

ACKNOWLEDGEMENTS

The author thanks Larry Shimkets, Rick Gourse and Roger Wartell for their comments on the manuscript, Larry Gold for a long discussion and Pieter deHaseth for a preprint of the functional group mutagenesis paper before publication. The author's research is supported by a grant from National Cancer Institute (CA34066).

REFERENCES

1. Seeman, N.C., Rosenberg, J.M., and Rich, A. (1976) Proc. Nat. Acad. Sci., U.S.A. 73, 804-808.
2. Dubendorff, J.W., deHaseth, P.L., Rosendahl, M.S. and Caruthers, M.H. (1987) J. Biol. Chem. 261, in press.
3. Doerfler, W. (1983) Ann. Rev. Biochem. 52, 93-124.
4. Pabo, C.O. and Sauer, R.T. (1984) Ann. Rev. Biochem. 53, 293-321.
5. McClelland, M. and Nelson, M. (1985) Nucleic Acids Res. 13, r201-207.

6. Berkner, K.L. and Folk, W.R. (1977) J. Biol. Chem. 252,3185-3193.
7. Berkner, K.L. and Folk, W.R. (1979) J. Biol. Chem. 254, 2552-2560.
8. Fliess, A., Wolfes, H., Rosenthal, A., Schwellnus, K., Blocker, H., Frank, R., and Pingoud, A. (1986) Nucleic Acids Res. 14, 3463-3474.
9. McClarin, J.A., Frederick, C.A., Wang, B-C., Greene, P., Boyer, H.W., Grable, J. and Rosenberg, J.M. (1986) Science 234, 1526-1541.
10. Yolov, A.A., Vinogradova, M.N., Gromova, E.S., Rosenthal, A., Cech, D., Veiko, V.P., Metelev, V.G., Kosyhk, V.G., Buryanov, Y.I., Bayev, A.A., and Shabarova, Z.A. (1985) Nucleic Acids Res. 13, 8983-8998.
11. Ogata, R. and Gilbert, W. (1977) Proc. Nat. Acad. Sci., U.S.A. 74, 4973-4976.
12. Simpson, R.B. (1979) Proc. Nat. Acad. Sci., U.S.A. 76, 3233-3237.
13. Fisher, E.F. and Caruthers, M.H. (1979) Nucleic Acids Res. 7, 401-415.
14. Goeddel, D.V., Yansura, D.G., and Caruthers, M.H. (1978) Proc. Nat. Acad. Sci., U.S.A. 75, 3578-3578.
15. Anderson, J.E., Ptashne, M. and Harrison, S.C. (1987) Nature 326, 846-852.
16. Wharton, R.P. and Ptashne, M. (1987) Nature 326, 888-891.
17. Hawley, D.K. and McClure, W.R. (1983) Nucleic Acids Res. 11, 2237-2255.
18. Spassky, A.. Kirkegaard, K., and Buc, H. (1985) Biochem. 24, 2723-2731.
19. Siebenlist, U., Simpson, R.B., and Gilbert, W. (1980) Cell 20, 269-281.
20. Simpson, R.B. (1979) Cell 18, 277-285.
21. Koo, H-S., Wu, H-M. and Crothers, D.M. (1986) Nature 320, 501-506.
22. Landy, A. and Ross, W. (1977) Science 197, 1147-1169.
23. Bucher, P. and Trifonov, E.N. (1986) Nucleic Acids Res. 14, 10009-10026.
24. Gasser, S.M. and Laemmli, U.K. (1986) Cell 46, 521-530.
25. Zyskind, J.W. and Smith, D.W. (1986) Cell 46, 489-490.
26. Vershon, A.K., Youderian, P., Weiss, M.A., Susskind, M.M., and Sauer, R.T. (1985) in Sequence Specificity in Transcription and Translation, Alan R. Liss, Inc., pp. 209-218.
27. Warner, H.R., Duncan, B.K., Garrett, C. and Neuhard, J. (1981) J. Bacteriol. 145, 687-695.
28. Warner, H.R. and Duncan, B.K. (1978) Nature 272, 32-34.
29. Mathews, C.K., Kutter, E.M., Mosig, G. and Berget, P.B., eds. (1983) Bacteriophage T4, Am. Soc. Microbiol., Inc.
30. Lesk, A.M. (1969) J. Theoret. Biol. 22, 537-540.
31. Grossman, L. (1962) Proc. Nat. Acad. Sci., U.S.A. 48, 1609-1614.
32. Grossman, L. (1963) Proc. Nat. Acad. Sci., U.s.A. 50, 657-663.
33. Ottensmeyer, F.P. and Whitmore, G.F. (1968a) J. Mol. Biol. 38, 1-16.
34. Ottensmeyer, F.P. and Whitmore, G.F. (1968b) J. Mol. Biol. 38, 17-24.
35. Remson, J.F. and Cerutti, P.A. (1972) Biochem. Biophys. Res. Commun. 48, 430-436.
36. Singer, B. and Fraenkel-Conrat, H. (1970) Biochem. 9, 3694-3701.
37. Singer, C.E. and Ames, B.N. (1970) Science 170, 822-826.
38. Coulondre, C., Miller, J.H., Farabaugh, P.J., and Gilbert, W. (1978) Nature 274, 775-780.
39. Lin, S. and Riggs, A.D. (1971) Biochem. Biophys. Res. Commun. 45, 1542-1547.
40. von Hippel, P.H. and Berg, O.G. (1986) Proc. Nat. Acad. Sci., U.S.A. 83, 1608-1612.