Nucleic Acids Research, 1994, Vol. 22, No. 1 1–10

The DNA (cytosine-5) methyltransferases

Sanjay Kumar, Xiaodong Cheng[1], Saulius Klimasauskas[1], Sha Mi[1], Janos Posfai, Richard J.Roberts* and Geoffrey G.Wilson
New England Biolabs, 32 Tozer Road, Beverly, MA 01915 and [1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

Received November 30, 1993; Accepted December 6, 1993

INTRODUCTION

The modification of DNA by methylation commonly occurrs in organisms as diverse as bacteria, plants, and mammals. DNA methylation is sequence-specific and MTases with single and multiple sequence-specificities exist. In prokaryotes, methylation of cytosine and adenine residues is primarily involved in restriction-modification systems that serve as 'immune responses' to phage infection (1). Adenine methylation in prokaryotes is also involved in regulating the initiation of DNA replication (2) and in targeting the correction of errors in DNA replication (3). In higher eukaryotes, methylation of cytosine residues appears to participate in the control of gene expression, developmental regulation, genomic imprinting and X-chromosome inactivation (4). Abberant DNA methylation may be mutagenic in mammals (5,6,7) and plays a role in the development of certain human diseases (8). A eukaryotic enzyme that carries out this methylation localizes to DNA replication foci in a cell cyle-dependent manner (9) and is essential for normal embryonic development in mice (10).

DNA methyltransferases (MTases) fall into three classes based on the type of methylation catalyzed. Two classes modify exocyclic nitrogens, converting adenine to N6-methyladenine or cytosine to N4-methylcytosine. The third class, the subject of this review, methylates the 5-carbon of the pyrimidine ring of cytosine, creating 5-methylcytosine. From an evolutionary perspective, the DNA-(cytosine-5) methyltransferases (m5C-MTases) appear to be unique. They can be found in both eukaryotes and prokaryotes, whereas the exocyclic MTases have been isolated only from prokaryotes, and they share a large set of well-conserved blocks of amino acid sequence that simplify their identification from primary sequence data and serve as a natural targets for functional studies.

The chemistry of cytosine-5 methylation has been extensively studied and is well understood. A methyl group is transferred from S-adenosyl-L-methionine (AdoMet) to the 5-carbon of cytosine in a manner postulated by Wu and Santi (11,12) to be analogous to other enzymes that catalyze one-carbon transfers to the 5-position of pyrimidines, e.g., thymidylate synthase, tRNA-(uracil-5) MTase, and dCMP hydroxymethylase (13,14). A key feature of this process is the formation of a transient covalent complex between the protein and the pyrimidine being modified. As shown in Fig. 1, a cysteine thiol on the enzyme serves as a nucleophile that attacks the 6-carbon of cytosine and forms a covalent DNA-protein intermediate. The addition of the nucleophile activates the 5-carbon allowing transfer of the methyl from AdoMet and release of S-adenosyl-L-homocysteine (AdoHcy). Following the methyl transfer, the proton at the 5-position is abstracted by a basic residue on the enzyme which is eliminated from the 6-position by β-elimination. A Cys embedded within a Pro-Cys dipeptide that is highly conserved in all m5C-MTases and also occurs in thymidylate synthase has been shown to be the thiol nucleophile (see section on conserved motifs).

M.*Hha*I is an m5C-MTase (EC 2.1.1.37); it recognizes the sequence 5′-GCGC-3′ in double-stranded DNA and methylates the inner cytosine to produce 5′-GmeCGC-3′. The enzyme was originally isolated from the bacterium *Haemophilus haemolyticus* (15) in which it serves as part of a restriction-modification system. The gene for this MTase has been cloned and sequenced, and Fig. 2 shows its 327 amino acid derived protein sequence. Overexpression of the protein in *E. coli* (16,17) and large-scale purification has allowed its crystallization (18) and led to the determination of its complete atomic structure (19,20). The availability of this structure significantly enhances our understanding of the architecture and mechanism of m5C-MTases. This review summarizes the relationship between the recently determined structure of M.*Hha*I and the known functions of this class of enzymes. Broader perspectives on DNA methylation and restriction-modification systems are also available (1,4,21,22).

Conserved motifs in m5C-MTases

Cloning and characterization of restriction and modification genes has progressed at a remarkable rate (23). Currently, fifty m5C-MTase sequences are available (Table I). Comparative analysis has shown that these proteins share an ordered set of sequence motifs which alternate with non-conserved regions (24–27). Depending on the criteria used to define the limits of the conserved blocks, up to ten motifs can be identified (24–26) (Figs. 2,3). For this review, we will use the nomenclature of Posfai *et al.* (26). All ten motifs can be identified in the majority of the known sequences, including the C-terminal 500 amino acids of the eukaryotic CpG MTases (Fig. 3). In the original analysis of 13 m5C-MTases (26), five motifs were considered highly conserved (I, IV, VI, VIII, and X), and the remaining five moderately conserved. Reanalysis of 36 sequences (19) resulted in the inclusion of a sixth motif, motif IX, within the highly conserved set (Fig. 4). That reanalysis, as well as the current one based on a subset of forty-five of the fifty available

* To whom correspondence should be addressed

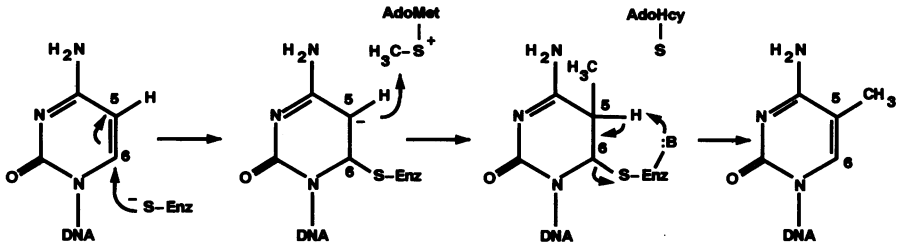


Figure 1. Schematic representation of the reaction pathway, based on the mechanism proposed by Wu and Santi (11,12) for thymidylate synthase and tRNA-(uracil-5) methyltransferase. The attack on carbons 5 and 6 occurs from above or below the plane of the pyrimidine ring.

Figure 2. Amino acid sequence of M.*Hha*I based upon the translation of Genbank entry J02677 (17). The six highly conserved motifs are highlighted in color (red – motif I, FGG; yellow – motif IV, PC; green – motif VI, ENV; cyan – motif VI, QRR; magenta – motif IX, RE; blue – motif X, GN). The intervening minor motifs (II, III, V, VII, in order from the N-terminus) are highlighted in black and white. The shading is described in Fig. 3. Brackets above and below the motif indicate the extent of the corresponding block boxed in Fig. 4.

sequences, shows only minor changes in the degree of conservation at each residue position within the highly conserved motifs. However, a few sequences have been detected in which unambiguous assignments for some of the motifs are not possible (indicated by the missing boxes in Fig. 4). These motifs (II, III, IX, X) are either more tolerant to sequence variation or non-essential for function in some of the proteins. The largest non-conserved or 'variable' region lies between motifs VIII and IX, and often varies greatly in size. The sequential order of the motifs appears to be important as well, since no deviation is seen in the 45 sequences examined here (Fig. 3).

The highly conserved motifs of the m5C-MTases provide tantalizing clues to the presumably common underlying architecture of these enzymes. DNA MTases must carry out two functions: recognition of a specific DNA sequence and catalysis of methyl transfer. The sequence organization of these enzymes suggests that these two functions are segregated into different domains. The conserved motifs are likely sites for the chemistry common to all MTases, while the variable region is a natural candidate for sequence specificity.

To date, two motifs have been assigned functional roles related to the common chemistry of these enzymes. Motif I (FxGxG) (Fig. 3, red) of the m5C-MTases is similar to a weakly conserved motif shared by other AdoMet-dependent MTases (27,28) and is presumed to be at the cofactor binding site. Motif IV (Fig. 3, yellow) contains an invariant Pro-Cys dipeptide that is known to be part of the catalytic site. It contains the nucleophilic thiol proposed by Wu and Santi (11), based on several criteria: analogy to thymidylate synthase which also contains a conserved Pro-Cys dipeptide (13,14); mutagenesis leading to loss of function in M.EcoRII (29), M.*Hha*I (30), M.*Hae*III (31), and Dcm (32); direct identification of the residue covalently attached to carbon-6 in the trapped intermediates formed with the suicide substrate 5-fluorodeoxycytosine (5FdC) (29,34). This covalent bond between Cys81 and carbon-6 of cytosine is clearly visible in the M.*Hha*I–DNA structure, since an oligonucleotide containing 5FdC was used in the co-crystal.

The variable region (between motifs VIII and IX) is known to define the sequence specificity of both mono-specific and multi-specific m5C-MTases (16,35–37). In the multi-specific enzymes, a point mutation in the variable region is capable of abolishing one target specificity while leaving the others intact. By mapping these mutations as well as determining the specificity of chimeric multi-specific MTases, Trautner's group was able to define and eventually swap target recognition domains (TRDs) within the variable region (35–37). For a thorough review, see Noyer-Weidner and Trautner (2). Hybrid swap experiments in the mono-specific MTases established that not only did the variable region determine the sequence-specificity, but also the choice of the specific base to be methylated within the target sequence (16,38). The monospecific MTase experiments also indicated that motif IX was capable of increasing the activity of the hybrid when it was swapped along with the variable region.

Structrural orientation of the motifs

Fig. 5 shows the three-dimensional structure of M.*Hha*I with the motifs color-coded as in Fig. 2. The enzyme folds into two domains separated by a large cleft that holds the DNA. The

Table 1.

Name[a]	Sequence	[b]Organism	Size (aa)[c]	Accession number[d]	Reference
*Alu*I	AGCT	*Arthrobacter luteus*	525	Z11841	51
M.*Apa*I	GGGCCC	*Acetobacter pasteurianus*	350		52
M.*Aqu*I	CYCGRG	*Agmenellum quadruplicatum*	248+139	M28051	48
M.*Asc*I	GGCGCGCC	*Arthrobacter sp.*	>430		48
M.*Ava*II	GGWCC	*Anabaena variabilis*	477		53
M.*Ban*I	GGYRCC	*Bacillus aneurinolyticus*	428	D00704	54
M.*Bbv*I	GCWGC	*Bacillus brevis*	374		55
M.*Bep*I	CGCG	*Brevibacterium epidermis*	403	X13555	56
M.*Bsp*RI	GGCC	*Bacillus sphaericus R*	424	X15758	57
M.*Bsu*FI	CCGG	*Bacillus subtilis*	409	X51515	58
M.*Bsu*RI	GGCC	*Bacillus subtilis R*	436	X02988	59
CpG MTase	CG	*Arabidopsis thaliana*	1534	L10692	60
CpG MTase	CG	Human	1495	X63692	61
CpG MTase	CG	Mouse	1502	X14805 M84387	62,63
M.*Cvi*JI	RGCB	*Chlorella* virus IL-3A	367	M27265	64
M.*Cvi*JJ	(RGCB)	*Chlorella* virus PBCV1	367	M83739	65
M.*Dde*I	CTNAG	*Desulfovibrio desulfuricans*	15	Y00449	39
M.*Eco*Dcm	CCWGG	*E. coli* K-12	472	X13330 M32307	66,67
M.*Eco*RII	CCWGG	*E. coli* plasmid N3	477	X05050	68
M.*Fnu*DI	GGCC	*Fusobacterium nucleatum* D	344		69
M.ϕ3TI	GCNGC GGCC	*Bacillus subtilis* phage F3T	443	M13488	70
M.*Hae*III	GGCC	*Haemophilus aegyptius*	330	M24625	71
M.*Hga*I-1	GCGTC	*Haemophilus gallinarum*	357	D90363	72
M.*Hga*I-2	GACGC	*Haemophilus gallinarum*	358	D90363	72
M.*Hgi*BI	GGWCC	*Herpetosiphon giganteus* Hpg5	437	X55137	73
M.*Hgi*CI	GGYRCC	*Herpetosiphon giganteus* Hpg9	420	X55138	74
M.*Hgi*CII	GGWCC	*Herpetosiphon giganteus* Hpg9	437	X55139	75
M.*Hgi*DI	GRCGYC	*Herpetosiphon giganteus* Hpa2	309	X55140	76
M.*Hgi*DII	GTCGAC	*Herpetosiphon giganteus* Hpa2	354	X55141	77
M.*Hgi*EI	GGWCC	*Herpetosiphon giganteus* Hpg24	437	X55142	78
M.*Hgi*GI	GRCGYC	*Herpetosiphon giganteus* Hpa1		X55143	79
M.*Hha*I	GCGC	*Haemophilus Haemolyticus*	327	J02677	17
M.*Hin*P1I	GCGC	*Haemophilus influenzae* P1	>300		55
M.*Hpa*II	CCGG	*Haemophilus parainfluenzae*	358	X51322	80
M.H2I	GDGCHC GCNGC (CCWGG) GGCC	*Bacillus amyloliquefaciens* phage H2	503	M72412	81
M.*Msp*I	CCGG	*Moraxella* sp.	418	X1419	82
M.*Mth*TI	GGCC	*Methanobacterium thermoformicicum*	330	M97222	83
M.*Ngo*MI	GCCGGC	*Neisseria gonorrhoeae* MS11	312	M86915	84
M.*Ngo*PII	GGCC	*Neisseria gonorrhoeae*	330	X06965	85
M.*Nla*X	?	*Neisseria lactamica*	313	X54485	86
M.$\varrho 11_s$I	GDGCHC (GCNGC) (CCWGG) GGCC	*Bacillus subtilis phage* $r11_s$	503	X05242	87
M.*Sau*3AI	GATC	*Staphylococcus aureus* 3A	412	M32470	88
M.*Sau*96I	GGNCC	*Staphylococcus aureus* PS96	430	X53096	89
M.*Scr*FIa	CCNGG	*Lactococcus lactis* subsp. *cremoris*	389	M87289	90
M.*Scr*FIb	CCNGG	*Lactococcus lactis* subsp. *cremoris*	360	L12227	91
M.*Sin*I	GGWCC	*Salmonella infantis*	461	J03391	92
M.SPβ	GCNGC GGCC	*Bacillus subtilis* phage SPb	443	M19513	89
M.SPR	CCWGG CCGG GGCC	*Bacillus subtilis* phage SPR	439 X01670	K02124	93,94
M.*Sso*II	CCNGG	*Shigella sonnei* 47 plasmid P4	379	M86545	95
M.*Sss*I	CG	*Spiroplasma species* MQ1	386	X17195	40

[a]The nomenclature is that of Smith and Nathans (50).

[b]Sequences in parentheses signify targets not normally recognized, but which can be recognized in the presence of 'activating' mutations (96). When known, the target cytosine is underlined. ? indicates the recognition sequence is undetermined. R=A or G; Y=C or T; W=A or T; B=not A; D=not C; H=not G; N=any base.

[c]Where size is listed as >, it is because there is ambiguity in the location of the initiator codon. In the case of M.SPβ and *Hgi*GI, no size is listed because the database entries contain only fragmentary information. M.*Aqu*I consists of two peptides.

[d]Numbers refer to GenBank or EMBL database entries. Sequences without accession numbers can be obtained from G.Wilson

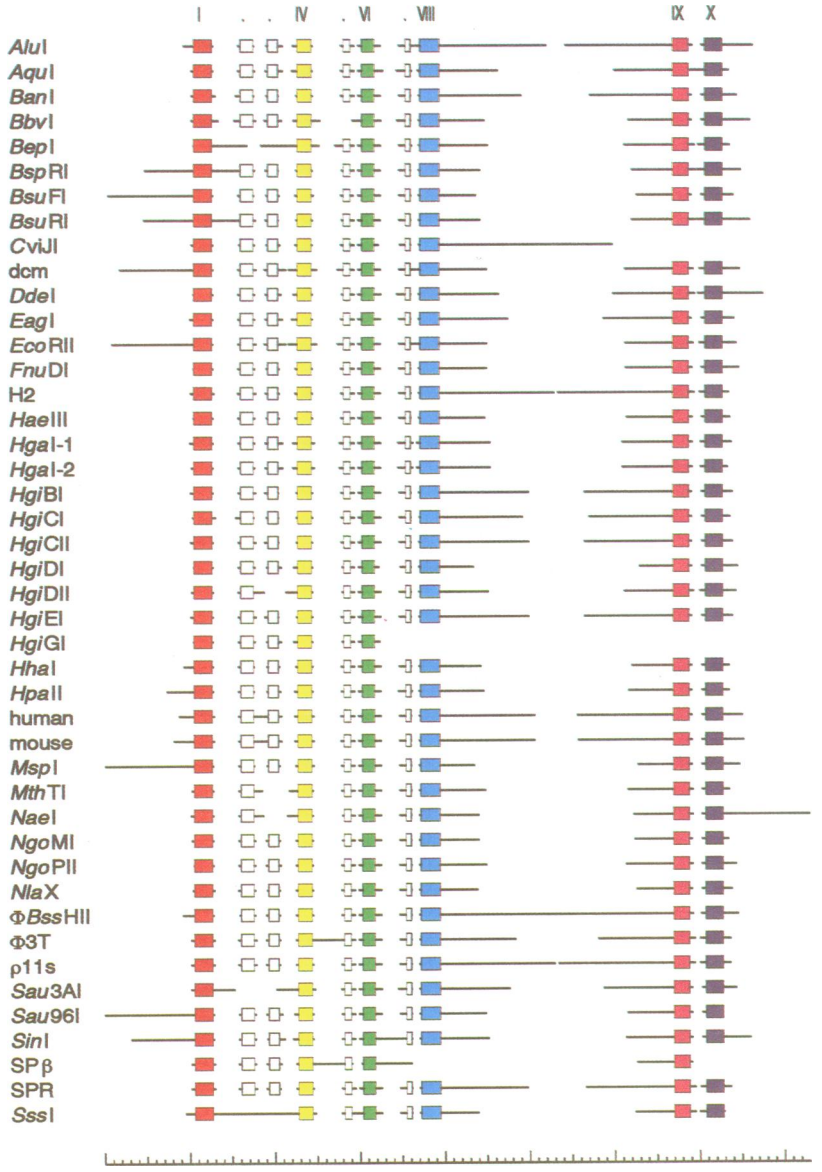


Figure 3. Schematic showing the alignment of 45 m5C-MTases. Sequences were aligned as previously described (26). After the analysis was completed, five more sequences were added to Table I, but were not included in the alignment. The color coding is as in Fig. 2. Where boxes are missing, the motif could not be unambiguously assigned. Breakpoints within unaligned segments were arbitrarily placed in the center of the segments. The sequence *Aqu*I consists of two separate peptides (48). Sequences SPβ and *Hgi*GI are fragmentary.

distribution of conserved sequences between the domains is highly asymmetric. The large domain (left in Fig. 5) encompasses motifs I through VIII and most of motif X which includes the residues essential for catalysis and cofactor binding. The highly conserved motifs I (red), IV (yellow), VI (green), and VIII (cyan) form the 'core' structure of this segment of the molecule. Most of the invariant residues in these motifs are situated immediately adjacent to the secondary structures in loops which face the cleft and are clustered around the active site of the molecule. The invariant Pro-Cys dipeptide, containing the catalytic Cys81 (in motif IV, yellow), is situated on a large flexible loop ('catalytic loop') that hovers over the cleft. In the presence of DNA, this loop moves substantially to come into contact with the DNA (Fig. 5C). The cofactor binding pocket is embedded within the large domain in the cleft.

The small domain of M.*Hha*I (right in Fig. 5) is dominated by the variable region which begins as a stalk emanating from the large domain, traverses the length of the protein at its surface, and then folds to form the bulk of the small domain (Fig. 5B). The only conserved motif in the small domain is motif IX, which has many interactions with the variable region. Hybrid swaps of variable regions which include motif IX are more likely to fold the small domain correctly, explaining the enhanced catalytic activity observed (16,38).

Given the invariance in motif order and the mostly conserved spacing between motifs I−VIII and IX−X, it is likely that all of the m5C-MTases will have a domain with a structure closely resembling the large domain of M.*Hha*I. In the few cases where the motif spacing is significantly altered, as between motifs IV and V in M.ϕ3TI, and motifs VI and VII in M.*Sin*I, the difference

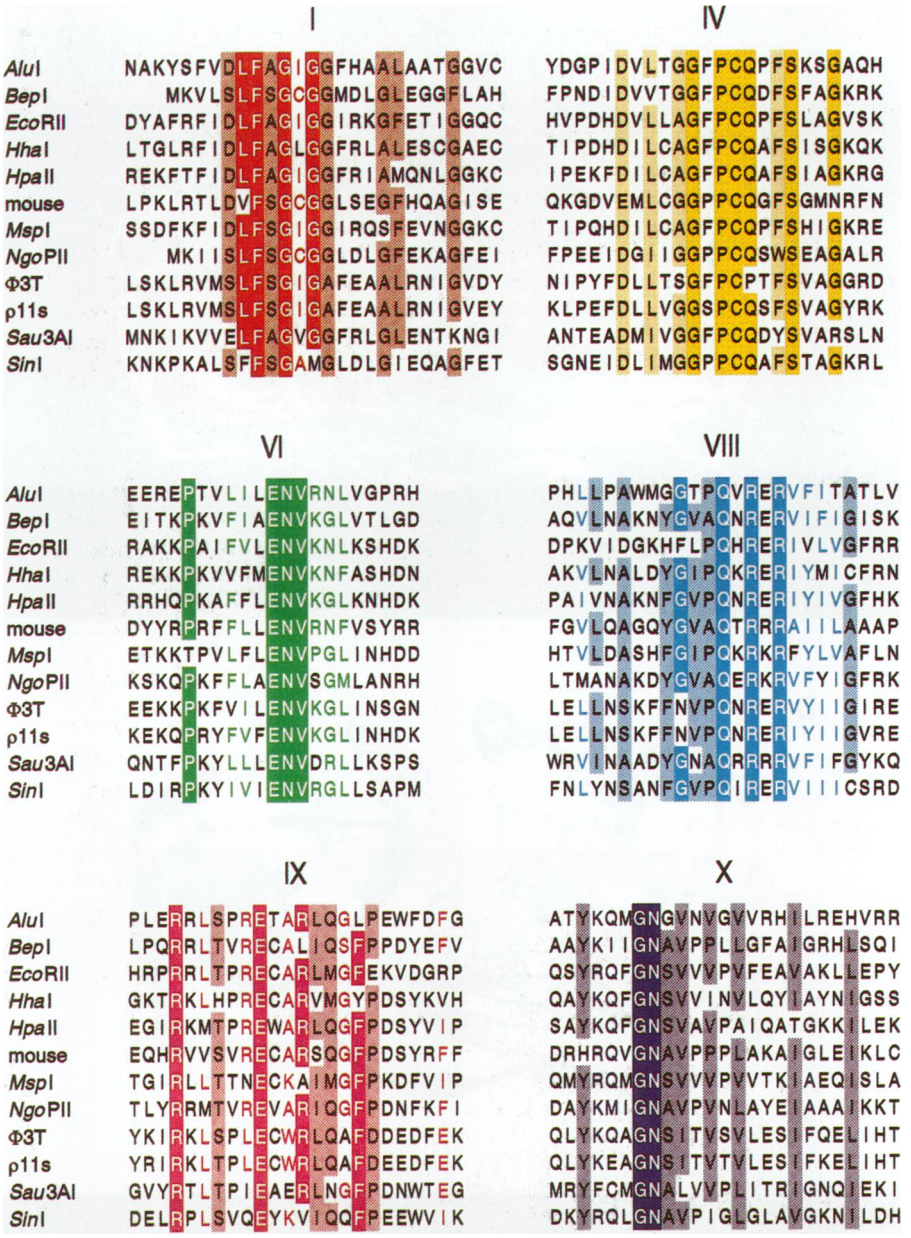


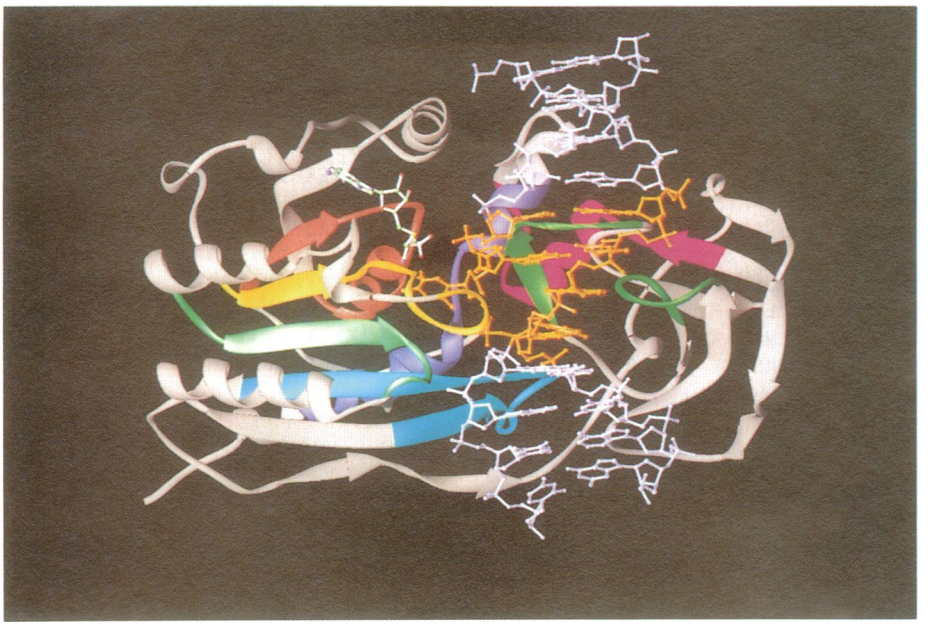
Figure 4. Representative sequences of the six highly conserved motifs. Four degrees of conservation ranging from complete to none are denoted in order of decreasing conservation by: solid (dark) colored background with white text, stippled colored background with black text, white background with colored text, and white background with black text. In the case of motif IV, only black text was used to aid visibility, so only three levels of conservation are indicated. The color scheme is as in Fig. 2.

is localized to a loop region near the protein's surface and should be easily accomodated. In contrast, the length of the variable region can show a 3-fold difference between different enzymes. The difference is most pronounced between the mono-specific MTases and the multi-specific enzymes, and correlates with the difference in the number of target sequences recognized by the two types of MTases. While the overall structure of the variable region in other enzymes cannot be extrapolated from the small domain of the M.*Hha*I structure, certain structural motifs may be present within it (see section on target recognition domains).

Interactions with the target cytosine

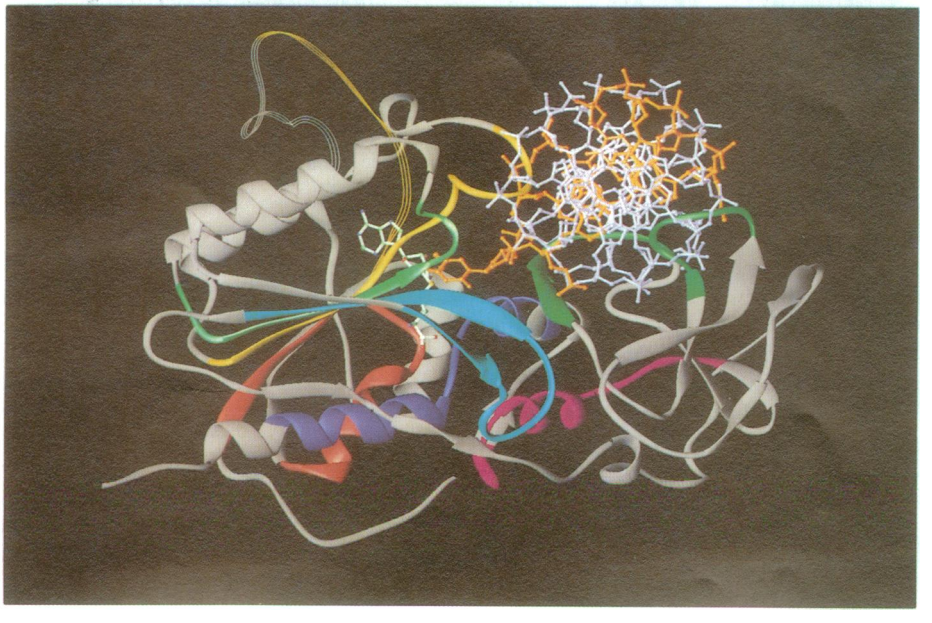
The stereochemistry of the methylation reaction dictates the spatial relationship between the reacting groups: the attack on the 5-position has to occur at right angles to the plane of the pyrimidine ring. In normal B-DNA, carbon-5 of cytosine is too deeply buried in the helix to allow this reaction to proceed, so a distortion of the helix was expected. However, the distortion that occurs is as surprising as it is elegant: the m5C-MTases cleanly extend the target cytosine out of the helix and into the catalytic site, without seriously disturbing the rest of the DNA helix (Fig. 5 and cover). In the process, the protein undergoes a major conformational change upon binding DNA; the tip of the catalytic loop containing motif IV (yellow) moves nearly 25 Å toward the cleft and into the minor groove of the DNA (Fig. 5C), at the same time pulling Cys81 into the region that will become the active site. Concurrently, the two domains move slightly toward the cleft and two Gly-rich 'recognition loops' from

A



B

C



the variable region (Fig. 5, dark green) contact the helix from the major groove side.

The details of how the target cytosine is trapped outside the helix are not clear. In M.*Hha*I, the gap left by the evicted base is filled by Gln237 which infiltrates the DNA helix from one of the recognition loops in the small domain. Gln237 provides hydrogen bonds to the orphan guanine residue and maintains the stacking of the helix (Fig. 6). Ser87 from the catalytic group interacts with Gln237, apparently to stabilize the conformation of the Gln. Conservation of the Gln cannot fully be assessed because of the difficulty in preparing alignments of the variable regions. However, neither the Ser nor the Gln appear to be well-conserved. Of the 45 sequences examined, only 2 others (M.*Dde*I and M.*Alu*I) carried Ser while 30 contained Ala at this position; several different amino acids filled this position in the remaining sequences. In M.*Dde*I, Gln also occurs at the corresponding position in the variable region (39). Interestingly, Ser and Gln also occur at the corresponding positions in motif IV and the variable region in M.*Sss*I (40), but their order is reversed.

Once outside the helix, the target cytosine is buried deep within the cleft and held in place by four residues, Phe79, Cys81, Arg165, and Glu119, from motifs IV (yellow), VI (green) and VIII (cyan), respectively, which converge at the active site (Figs. 6,7). Cys81, Glu119, and Arg165 are completely conserved across all known m5C-MTases. Phe79 interacts via backbone carbonyls and is not absolutely conserved. Interactions between these three motifs and a portion of the variable region may explain why they are so highly conserved. Glu119, which hydrogen bonds to the N3 and N4 of cytosine, may determine the substrate specificity for cytosine. In thymidylate synthase, which carries out a mechanistically similar methylation on dUMP, a conserved Asn interacts with N3 and O4 of the substrate. Mutation of this Asn to Asp or Glu converts the substrate specificity to dCMP (41). Data on thymidylate synthase and cytidylate hydroxymethylase (41,42) show that protonation of N3 of cytidine is essential for catalysis due to its ability to stabilize covalent intermediates (31).

Adomet binding

AdoMet binds to the large domain of M.*Hha*I in a pocket facing the cleft. Residues from all five of the N-terminal motifs (I–V) and the C-terminal motif X contribute to the binding pocket (Fig. 7). Motif I (FxGxG, red), long suspected to be part of the binding site, forms a tight loop in the first turn of a β_1-α_A-β_2 tructural unit (19). The conserved Gly residues appear important in allowing a tight turn to occur in the structure (19,37), and only four sequences with a total of three different substitutions have been seen naturally; all were small residues (Ser, Cys, Ala). The conserved Phe in Motif I interacts with the adenosyl moiety of the cofactor, with its aromatic ring perpendicular to the plane of the purine ring (Fig. 7). With the exception of two sequences in which it is replaced by Cys, this residue is completely conserved and would appear to be an essential component of the cofactor binding pocket. However, it should be noted that the corresponding motif in other MTases often substitutes Gly and sometimes other amino acids for Phe (27,28). Two invariably acidic residues from motifs II and III, Glu40 and Asp60, also interact with the cofactor.

The combined region of motifs I, II, and III around this portion of the binding pocket bears a strong resemblance to the Rossmann fold of the dinucleotide-binding motif, which is built upon a similar structure. It also contains a Gly-rich consensus (VxGxGxxG) and has analogous acidic residues (43,44). Motifs II and III cannot be identified in several of the m5C-MTases; however, each motif contains only one highly conserved position, namely the acidic residue, and a few moderately conserved sites and may be quite tolerant to flanking sequence variability. The remainder of the pocket is formed by Pro80 (motif IV), Leu100 (motif V), and Trp41 (motif II) (Fig. 7). While Pro and Leu are absolutely conserved, the Trp is not. Trp41 lies parallel to the cofactor purine ring and provides substantial interaction with the cofactor in M.*Hha*I. The proximity of the AdoMet binding site to the catalytic site explains why in photolabelling experiments with M.*Eco*RII, a tritium-labelled methyl group of AdoMet became attached to the Cys of the Pro-Cys dipeptide (motif IV) (45). The previously unsuspected interaction of motif X with AdoMet or AdoHcy is mediated through Asn304 of the conserved Gly-Asn dipeptide (Fig. 7).

Target recognition domains

The target recognition functions of m5C-MTases reside in the variable region between motifs VIII and IX. As discussed above, much of the data supporting this has come from Trautner's work on the multi-specific m5C-MTases, which have large variable regions that can recognize up to four target DNA sequences. By determining the effects of random point mutations (37), site-directed mutations (36,37), and chimeric constructs (35,36) on the target specificities of multi-specific phage MTases, Trautner's group defined several individual target recognition domain (TRD) sequences. The domains spanned approximately 40 amino acids each; individual TRDs did not overlap and were separated by a single amino acid (2,35). Sequence comparisons between TRDs suggested a weak consensus sequence: T(V/I/L)XXXXXXG(V/L) (25).

DNA binds to M.*Hha*I in the cleft between the domains with its major groove facing the variable region in the small domain and its minor groove toward the protein's catalytic site in the large domain. None of the usual DNA-recognizing structural motifs are present in the structure (46). Instead, three loops contact the target sequence: the catalytic loop (yellow) from the large domain, and two Gly-rich recognition loops (dark green, I: aa 233–240, GKGGQGER; II: aa 250–257, TLSAYGGG)

Figure 5. The structure of M.*Hha*I covalently linked to a DNA oligonucleotide containing 5-fluorocytosine at the target. The product of the reaction, AdoHcy, is also present (white). The color coding for the motifs is as in Fig. 2, with the exception of DNA recognition loops which are colored dark green. The target sequence is colored gold. **A.** Side view of the DNA, looking down into the active site cleft of the enzyme. The large domain is pictured on the left. **B.** Same view as in A, but without DNA. The variable region is shaded black. C. View looking down the helix axis. The loop near the top of the structure, drawn with three thin lines and partially colored in yellow, is the catalytic loop that undergoes a major conformational change when DNA is bound (see text). The thin lines show the conformation without DNA present. In the presence of DNA this loops moves towards the DNA into the position shown by the solid-filled loop with the same color scheme.

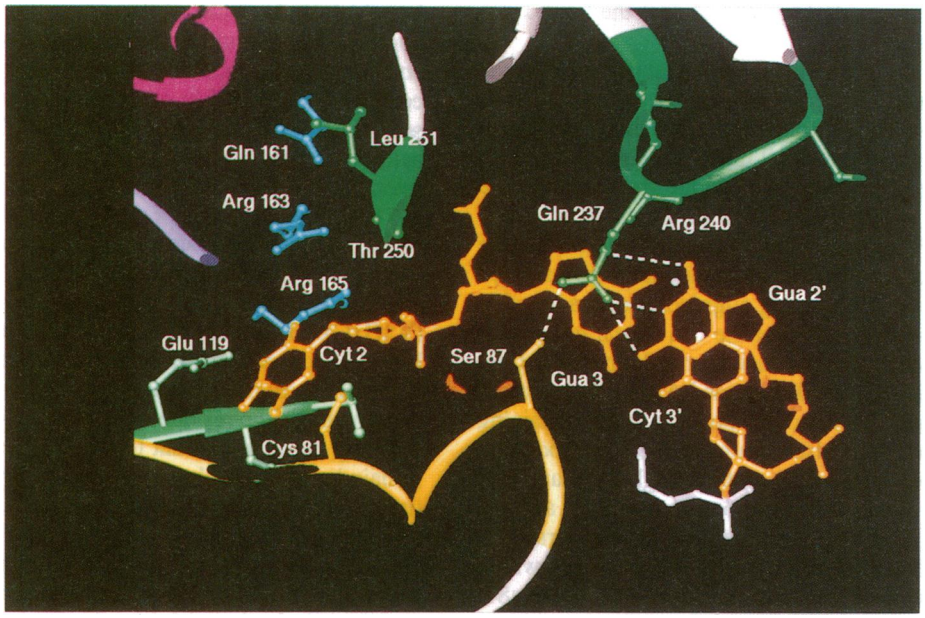


Figure 6. Closeup of the interaction between Gln 237, Ser 87, and the target cytosine. The target sequence is numbered 1 to 4 (5′ to 3′). The sequence on the complementary strand is numbered 4′ to 1′ (5′ to 3′). Only the 2−2′ and 3−3′ base pairs are shown. The color scheme is as in Fig. 5. Portions of the structure were removed for clarity.

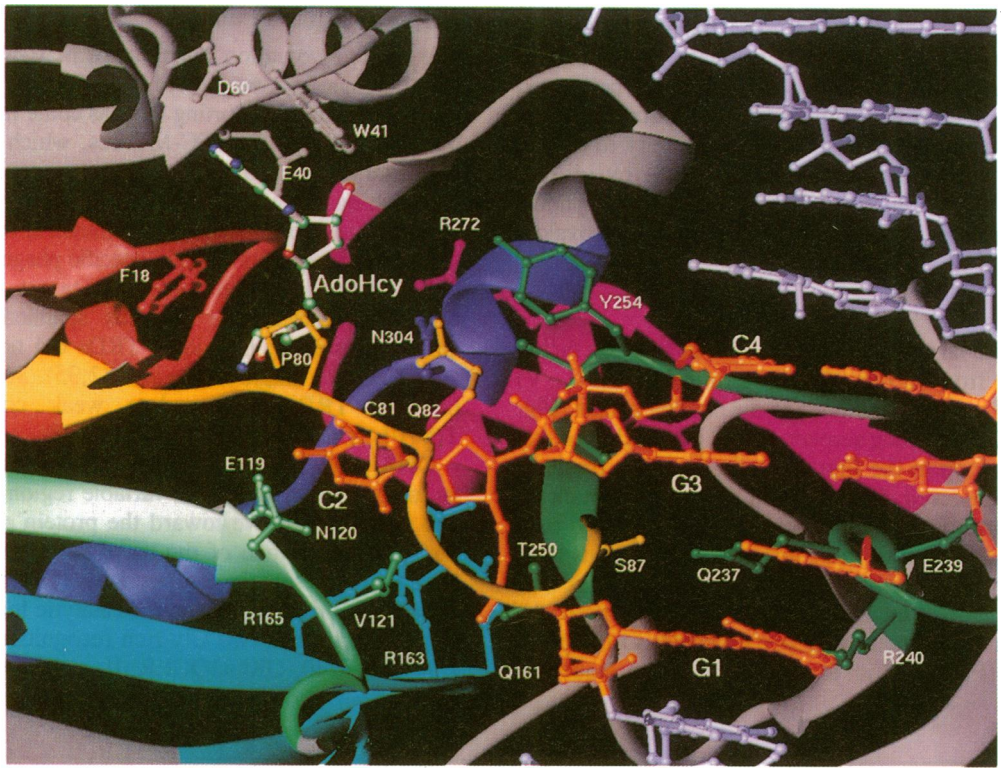


Figure 7. The active site of M.*Hha*I. The viewing angle is very close to that shown in Fig. 5A. The color coding is as in Fig. 5. Portions of the structure were removed for clarity. Residues are labelled with one-letter codes.

from the small domain (Figs. 5−7). The two recognition loops contact non-overlapping portions of both strands of the target in a diagonal fashion. Loop I primarily contacts the 5′ half of the recognition sequence in the DNA strand complementary to the one carrying the target cytosine and provides the residue (Glu 237) that pairs with the orphan guanine. Loop II interacts extensively with the backbone of the DNA strand containing the target cytosine, and contacts three bases in the 3′ half of the

double-stranded recognition sequence. A detailed discussion of the DNA-protein contacts seen in this complex has appeared (20).

The two Gly-rich recognition loops fall within the putative TRD of M.*Hha*I predicted by Lauster *et al.* (25); loop II, in fact, falls entirely within the TRD consensus sequence. Given the weakness of the TRD consensus used to anchor the alignments, the accuracy of their prediction is remarkable. When DNA is bound in the cleft, the Thr-Leu dipeptide from the TRD consensus (Thr250-Leu251) appears to be important in positioning recognition loop II relative to the target cytosine. In the absence of DNA, the Thr-Leu dipeptide interacts with motif VIII (20).

TRD mapping in the multi-specific MTases shows they are excluded from the N-terminal portion of the variable region (37,25). This portion forms a long stalk that connects the two domains of the molecule (bottom-most chain in Figs. 5A,C; the other connector is formed by portions of motifs IX and X) and is not positioned within the cleft. Loss of general function mutations that map to the N-terminal half of the variable region in the phage multi-specific MTases (37) suggest that a major perturbation of the stalk might reposition the entire small domain relative the catalytic site in the large domain. The segregation of TRDs into a separate domain and their localization within loops may explain their ability to be replaced by other TRDs or even random DNA fragments without disturbing neighboring TRDs or the catalytic site in multi-specif enzymes (47). However, swapping the entire variable region from a mono-specific MTase into a phage TRD element failed to confer the mono-specific MTase's specificity to the phage MTase (47); in all likelihood, the mono-specific TRD, which is a subdomain of the variable region, would be incorrectly positioned relative to the catalytic site.

Unusual m5C-MTases

Although the structure of M.*Hha*I helps to resolve many questions about m5C-MTases, numerous problems remain. For example, it comes as no surprise that mutation of the catalytic Cys abolishes methyltransferase activity (29–32,34). What is surprising is that in M.*Hha*I and M.*Eco*RII replacement of the Cys with Gly causes cytotoxicity, whereas replacement with others does not (29,30,34). The source of the the cytoxicity appears to be unusually tight binding of the DNA substrate, though the mechanism for this effect is not understood.

An interesting pair of sequences are the two open reading frames (ORF-a and ORF-b) that encode the M.AquI m5C-MTase (48). ORF-a encodes the N-terminal 248 residues of the enzyme, while ORF-β codes for the C-terminal 139 residues. Both peptides are needed for methylase activity. All 10 motifs are present in the combined peptides (Fig. 3), with the breakpoint between the two sequences falling within the center of the variable region. An artificial analogy to M.*Aqu*I was generated using M.*Bsu*RI and the closely related M.*Bsp*RI (49). Individual N- and C-terminal peptides expressed from separate vectors were shown to be capable of complementation to generate functional m5C-MTases. Pairs of peptides that resulted in gapped regions (missing motifs) were unable to complement each other. Surprisingly, peptides with very large overlaps (in one case duplicating motifs II–VIII and the N-terminal half of the variable region) were still able to complement. It is not clear how these latter fragments interact with each other or the DNA substrate.

Although the mechanism of methyl transfer appears quite clearly defined, an interesting observation has been made about the two m5C-MTases M.Dcm and M.*Bsp*RI. Incubation of these enzymes with AdoMet in the absence of DNA results in transfer of the methyl group to the protein (32, A. Kiss, personal communication). In the case of M.*Bsp*RI, two Cys residues (Cys156, Cys181)were identified as the recipients of the methyl group; one of these, Cys156, was part of the conserved Pro-Cys doublet.Subsequent reaction of the M.*Bsp*RI methylated at Cys156 with DNA allows transfer of the methyl group to the DNA, but the enzyme is kinetically incapacitated and carries out the reaction at a significantly slower rate. It is not known whether these methylated proteins serve as intermediates in the normal reaction pathway for these particular enzymes.

SUMMARY

The m5C-MTases form a closely-knit family of enzymes in which common amino acid sequence motifs almost certainly translate into common structural and functional elements. These common elements are located predominantly in a single structural domain that performs the chemistry of the reaction. Sequence-specific DNA recognition is accomplished by a separate domain that contains recognition elements not seen in other structures. This, combined with the novel and unexpected mechanistic feature of trapping a base out of the DNA helix, makes the m5C-MTases an intriguing class of enzymes for further study. The reaction pathway has suddenly become more complicated because of the base-flipping and much remains to be learned about the DNA recognition elements in the family members for which structural information is not yet available.

ACKNOWLEDGEMENTS

We thank S.Linn and R.Gumport for critical comments on the manuscript. This work was supported in part by funding from a NIH fellowship (GM 15262) to S.K., a grant from the NIH (GM 49245) to X.C., an American Cancer Research Postdoctoral Fellowship to M.S., and grants from the NIH (GM46127) and NSF (DMB-8917650) to R.J.R.

REFERENCES

1. Roberts, R.J. and Halford, S.S. (1993) In Linn, S.M., Lloyd, R.S. and Roberts, R.J. (eds.) Nucleases. Cold Spring Harbor Press. In press.
2. Noyer-Weidener, M. and Trautner, T.A. (1993) In Jost, J.P. and Saluz, H.P. (eds.) DNA Methylation: Molecular Biology and Significance. Birkhauser Verlag, Basel, pp 39–108.
3. Modrich, P. (1991) Annu. Rev. Genet. **25**: 229–253.
4. Jost, J.P. and Saluz, A.P. (1993) DNA Methylation: Molecular Biology and Significance, Birkhauser Verlag, Basel.
5. Cooper, D.N. and Youssoufian, M. (1988) Hum. Genet. **78:** 151–155.
6. Rideout, W.M., Coetzee, G.A., Olumi, A.F. and Jones, P.A. (1990) Science **249**: 1288–1290.
7. Shen, J.C., Rideout, W.M. and Jones, P.A. (1992) Cell **71**: 1073–1080.
8. Oberle, I., Rousseau, F., Heitz, D., Kretz, C., Devys, D., Hanauer, A., Boue, J., Bertheas, M.F. and Mandel, J.L. (1991) Science **252:** 1097–1102.
9. Leonhardt, H., Page, A.W., Weier, H.U. and Bestor, T.H. (1992) Cell **71:** 865–873.
10. Li, E., Bestor, T.H. and Jaenisch, R. (1992) Cell **69:** 915–926.
11. Wu, J.C. and Santi, D.V. (1985) In Cantoni, G.I. and Razin, A. (eds.) Biochemistry and Biology of DNA Methylation. Alan R. Liss Inc., New York. pp 119–129.
12. Wu, J.C. and Santi, D.V. (1987) J. Biol. Chem. **262**: 4778–4786.
13. Santi, D.V. and Danenberg, P.V. (1984) In Blakely, R.L. and Benkovic, S.J. (eds.) Folates and Pterins. Wiley-Interscience, New York. Vol **1**, pp. 343–396.
14. Santi, D.V. and Hardy, L.W. (1987) Biochemistry **26:** 8599–8606.
15. Roberts, R.J., Myers, P.A., Morrison, A. and Murray, K. (1976) J. Mol. Biol. **103:** 199–208.

16. Klimasauskas, S., Nelson, J.L. and Roberts, R.J. (1991) Nucl. Acids Res. **19:** 6183–6190.
17. Caserta, M., Zacharias, W., Nwankwo, D., Wilson, G.G. and Wells, R.D. (1987) J. Biol. Chem. **262:** 4770–4777.
18. Kumar, S., Cheng, X., Pflugrath, J.W. and Roberts, R.J. (1992) Biochemistry **31:** 8648–8653.
19. Cheng, X., Kumar, S., Posfai, J., Pflugrath, J.W. and Roberts, R.J. (1993) Cell **74**: 299–307.
20. Klimasauskas, S., Kumar, S., Roberts, R.J. and Cheng,X. (1994) Cell in press.
21. Wilson, G.G. and Murray, N.E. (1991) Ann. Rev. Genet. **25:** 585–627.
22. Bestor, T.H. (1990) Phil. Trans. R. Soc. Lond. **326**: 179–187.
23. Wilson, G.G. (1991) Nucl. Acids Res. **19:** 2539–2566.
24. Som, S., Bhagwat, A.S. and Friedman, S. (1987) Nucl. Acids Res. **15:** 313–23.
25. Lauster, R., Trautner, T.A. and Noyer-Weidner, M. (1989) J. Mol. Biol. **206:** 305–312.
26. Posfai, J., Bhagwat, A.S., Posfai, G. and Roberts, R.J. (1989) Nucl. Acids Res. **17:** 2421–2435.
27. Klimasauskas, S., Timinskas, A., Menkevicius, S., Butkiene, D., Butkus, V. and Janulaitis, A.A. (1989) Nucl. Acids Res. **17:** 9823–9832.
28. Ingrosso, D., Fowler, A.V., Bleibaum, J. and Clarke, S. (1989) J. Biol. Chem. **264:** 20130–20139.
29. Wyszynski, M.W., Gabbara, S., Kubareva, E.A., Romanova, E.A., Oretskaya, T.S., Gromova, E.S., Shabarova, Z.A. and Bhagwat, A.S. (1993) Nucl. Acids Res. **21**: 295–301.
30. Mi, S. and Roberts, R.J. (1993) Nucl. Acids Res. **21:** 2459–2464.
31. Chen, L., MacMillan, A.M. and Verdine, G.L. (1993) J. Amer. Chem Soc. **115:** 5318–5319.
32. Hanck, T., Schmidt, S. and Fritz, H-J. (1993) Nucl. Acids. Res. **21**:303–309.
33. Chen, L., MacMillan, A.M., Chang, W., Ezaz-Nikpay, K., Lane, W.S. and Verdine, G.L. (1991) Biochemistry **30:** 11018–11025.
34. Wyszynski, M.W., Gabbara, S. and Bhagwat, A.S.. (1992) Nucl. Acids Res. **20:** 319–326.
35. Balganesh, T.S., Reiners, L., Lauster, R., Noyer-Weidner, M., Wilke, K. and Trautner, T.A. (1987) EMBO J. **6:** 3543–3549.
36. Trautner, T.A., Balganesh, T.S. and Pawleck, B. (1988) Nucl. Acids Res. **16:** 6649–6657.
37. Wilke, K., Rauhut, E., Noyer-Weidener, M., Lauster, R., Pawlek, B., Behrens, B. and Trautner, T.A. (1988) EMBO J. **7:** 2601–2609.
38. Mi, S. and Roberts, R.J.. (1992) Nucl. Acids Res. **20:** 4811–4816
39. Sznyter, L.A., Slatko, B., Moran, L., O'Donnell, K.H. and Brooks, J.E. (1987) Nucl. Acids Res. **15:** 8249–8266.
40. Renbaum, P., Abrahamove, D., Fainsod, A., Wilson, G.G., Rottem, S. and Razin, A. (1990) Nucl. Acids Res. **18:** 1145–1152.
41. Liu, L. and Santi, D.V. (1993) Biochem. **31**: 5100–5104.
42. Graves, K.L., Butler, M.M. and Hardy, L.W. (1992) Biochemistry **31:** 10315–10321.
43. Rossmann, M.G., Moras, D. and Olsen, K.W. (1974) Nature **250:** 194–199.
44. Wierenga, R.K., Terpstra, P. and Hol, W.J.G. (1986) J. Mol. Biol. **187:** 101–107.
45. Som, S. and Friedman, S. (1991) J. Biol. Chem. **266:** 2937–2945.
46. Harrison, S.C. (1991) Nature **353:** 715–719.
47. Walter, J., Trautner, T.A. and Noyer-Weidener, M. (1992) EMBO J. **11:** 4445–4450.
48. Karreman, C. and de Waard, A. (1990) J. Bacteriol. **172:** 266–272.
49. Posfai, G., Kim, S.C., Szilak, L., Kovacs, A. and Venetianer, P. (1991) Nucl. Acids Res. **19**: 4843–4847.
50. Smith, H.O. and Nathans, D. (1973) J. Mol. Biol. **81:** 419–423.
51. Zhang, B., Tao, T., Wilson, G.G. and Blumenthal, R.M. (1993) Nucl. Acids Res. **21:** 905–911.
52. Hall, D., Moran, L.S., Slatko, B.E. and Kong, H. unpublished results.
53. Lunnen, K.D., Moran, L.S., Slatko, B.E. and Wilson, G.G. unpublished results.
54. Maekawa, Y., Yasukawa, H. and Kawakami, B. (1990) J. Biochem. **107:** 645–649.
55. Barsomian, J.M. and Wilson, G.G. unpublished results.
56. Kupper, D., Zhou, J.-G., Venetianer, P. and Kiss, A. (1989) Nucl. Acids Res. **17:** 1077–1088.
57. Posfai, G., Kiss, A., Erdei, S., Posfai, J. and Venetianer, P. (1983) J. Mol. Biol. **170:** 597–610.
58. Walter, J., Noyer-Weidner, M. and Trautner, T.A. (1990) EMBO J. **9:** 1007–1013.
59. Kiss, A., Posfai, G., Keller, C.C., Venetianer, P. and Roberts, R.J. (1985) Nucl. Acids Res. **13:** 6403–6421.
60. Finnegan, E.J. and Dennis, E.S. (1993) Nucl. Acids Res. **21:** 2383–2388.
61. Yen, R.W., Vertino, P.M., Nelkin, B.D., Yu, J.J., el-Deiry, W., Cumaraswamy, A., Lennon, G.G., Trask, B.J., Celano, P. and Baylin, S.B. (1992) Nucl. Acids Res, **20:** 2287–2291.
62. Bestor, T., Laudano, A., Mattaliano, R. and Ingram, V. (1988) J. Mol. Biol. **203:** 971–983.
63. Rouleau, J., Tanigawa, G. and Szyf, M. (1992) J. Biol. Chem. **267:** 7368–7377.
64. Shields, S.L., Burbank, D.E., Grabherr, R. and Van Etten, J.L. (1990) Virology, **176:** 16–24.
65. Zhang, Y., Nelson, M. and Van Etten, J.L. (1992) Nucl. Acids Res. **20:** 1637–1642.
66. Hanck, T., Gerwin, N. and Fritz, H.-J. (1989) Nucl. Acids Res. **17:** 5844.
67. Sohail, A., Lieb, M., Dar, M. and Bhagwat, A.S. (1990) J. Bacteriol. **172:** 4214–4221.
68. Som, S., Bhagwat, A.S. and Friedman, S. (1987) Nucl. Acids Res. **15:** 313–332.
69. Zhang, B.-H., Van Cott, E.M., Benner, J.S. and Wilson, G.G. unpublished results.
70. Tran-Betcke, A., Behrens, B., Noyer-Weidner, M. and Trautner, T.A. (1986) Gene, **42:** 89–96.
71. Slatko, B.E., Croft, R., Moran, L.S. and Wilson, G.G. (1988) Gene, **74:** 45–50.
72. Sugisaki, H., Yamamoto, K. and Takanami, M. (1991) J. Biol. Chem. **266:** 13952–13957.
73. Dusterhoft, A., Erdmann, D. and Kroger, M. (1991) Nucl. Acids Res. **19:** 3207–3211.
74. Erdmann, D., Dusterhoft, A. and Kroger, M. (1991) Eur. J. Biochem. **202:** 1247–1256.
75. Erdmann, D., Horst, G., Dusterhoft, A. and Kroger, M. (1992) Gene, **117:** 15–22.
76. Dusterhoft, A., Erdmann, D. and Kroger, M. (1991) Nucl. Acids Res. **19:** 1049–1056.
77. Dusterhoft, A. and Kroger, M. (1991) Gene, **106:** 87–92.
78. Erdmann, D. (1991) Ph. D. Thesis. Justas-Liebig Univ., Geissen, W.Germany.
79. Dusterhoft, A. (1990) Ph. D. Thesis. Justas-Liebig Univ., Geissen, W.Germany.
80. Card, C.O., Wilson, G.G., Weule, K., Hasapes, J., Kiss, A. and Roberts, R.J. (1990) Nucl. Acids Res. **18:** 1377–1383.
81. Lange, C., Noyer-Weidner, M., Trautner, T.A., Weiner, M. and Zahler, S.A. (1991) Gene, **100:** 213–218.
82. Linn, P.M., Lee, C.H. and Roberts, R.J. (1989) Nucl. Acids Res. **17:** 3001–3011.
83. Nolling, J. and De Vos, W.M. (1992) J. Bacteriol. **174:** 5719–5726.
84. Stein, D.C., Chien, R. and Seifert, S. (1992) J. Bacteriol. **174:** 4899–4906.
85. Sullivan, K.M. and Saunders, J.R. (1988) Nucl. Acids Res. **16:** 4369–4387.
86. Labbe, D., Holtke, H.J. and Lau, P.C.K. (1990) Mol. Gen. Genet. **224:** 101–110.
87. Behrens, B., Noyer-Weidner, M., Pawlek, B., Lauster, R., Balganesh, T.S. and Trautner, T.A. (1987) EMBO J. **6:** 1137–1142.
88. Seeber, S., Kessler, C. and Gotz, F. (1990) Gene, **94:** 37–43.
89. Szilak, L., Venetianer, P. and Kiss, A. (1990) Nucl. Acids Res. **18:** 4659–4664.
90. Davis, R., Van der Lelie, D., Mercenier, A., Daly, C. and Fitzgerald, G.F. (1993) Appl. Environ. Microbiol. **59:** 777–785.
91. Twomey, D.P., Davis, R., Daly, C. and Fitzgerald, G.F., unpublished results.
92. Karreman, C. and de Waard, A. (1988) J. Bacteriol. **170:** 2527–2532.
93. Buhk, H.-J., Behrens, B., Tailor, R., Wilke, K., Prada, J.J., Gunthert, U., Noyer-Weidner, M., Jentsch, S. and Trautner, T.A. (1984) Gene, **29:** 51–61.
94. Posfai, G., Baldauf, F., Erdei, S., Posfai, J., Venetianer, P. and Kiss, A. (1984) Nucl. Acids Res. **12:**9039–9049.
95. Karyagina, A.S., Lunin, V.G., Degtyarenko, K.N., Uvarov, V.Y. and Nikolskaya, I.I. (1993) Gene, **124:** 13–19.
96. Lange, C., Jugel, A., Walter, J., Noyer-Weidener, M. and Trautner, T.A. (1991) Nature **352**: 645–648.