

---

**Predictive motifs derived from cytosine methyltransferases**

---

János Pósfai<sup>+</sup>, Ashok S. Bhagwat\*, György Pósfai<sup>1,♠</sup> and Richard J. Roberts

---

Cold Spring Harbor Laboratory, PO Box 100, Cold Spring Harbor, NY 11724 and <sup>1</sup>University of Wisconsin, McArdle Laboratory for Cancer Research, 450 N-Randall Avenue, Madison, WI 53705, USA

---

Received January 19, 1989. Accepted February 23, 1989

**ABSTRACT**

Thirteen bacterial DNA methyltransferases that catalyze the formation of 5-methylcytosine within specific DNA sequences possess related structures. Similar building blocks (motifs), containing invariant positions, can be found in the same order in all thirteen sequences. Five of these blocks are highly conserved while a further five contain weaker similarities. One block, which has the most invariant residues, contains the proline-cysteine dipeptide of the proposed catalytic site. A region in the second half of each sequence is unusually variable both in length and sequence composition. Those methyltransferases that exhibit significant homology in this region share common specificity in DNA recognition. The five highly conserved motifs can be used to discriminate the known 5-methylcytosine forming methyltransferases from all other methyltransferases of known sequence, and from all other identified proteins in the PIR, GenBank and EMBL databases. These five motifs occur in a mammalian methyltransferase responsible for the formation of 5-methylcytosine within CG dinucleotides. By searching the unidentified open reading frames present in the GenBank and EMBL databases, two potential 5-methylcytosine forming methyltransferases have been found.

**INTRODUCTION**

DNA methyltransferases (MTases) recognize specific nucleic acid sequence patterns in their targets and transfer methyl groups from the donor S-adenosylmethionine (SAM) to adenine or cytosine residues (1,2). MTases that are components of bacterial restriction-modification systems protect the DNA against restriction by methylating specific bases within the target sequence for the restriction enzyme. Some bacteriophage-encoded MTases are not associated with restriction enzymes and their role is probably to protect the bacteriophage genome against the restriction enzymes of the hosts. Additionally, several prokaryotic chromosomal MTases are known that are not paired with restriction enzymes and, in one case, the MTase is involved in DNA mismatch repair (3). Following the initial reports of sequence similarity between the first two 5-methylcytosine forming MTases ( $m^5C$  MTase), BspRI and SPR, (4,5) further similarities have been noted as new MTase sequences have been determined (6,7,8,9,10,11). The lack of obvious similarity between the sequences of  $m^5C$  MTases and  $N^6$ -methyladenine forming MTases ( $m^6A$  MTases) and between the sequences of the methylase and restriction components of restriction-modification systems have also been pointed out (6,7,8,9,10,11). Common domains in the  $m^5C$  MTase sequences were suggested to be the sites of common functions, and a global homology for these sequences has been proposed (7,9,11). A proline-cysteine (PC) doublet present in four sequences was postulated to be part of the catalytic site (12). This was based on the analogy to the catalytic mechanism of thymidylate synthase (13), which involves the formation of a covalent bond between the cysteine residue of a PC dipeptide

and the 6-position of the pyrimidine to which the methyl group is to be transferred (13). It has been shown directly that a cysteine is involved in the enzymatic reaction of the *HhaI* MTase (12). Within the bacteriophage MTases strong evidence exists that the variable segments of their sequences are responsible for their interaction with different DNA recognition sequences (14,15,16). A marginal similarity has been reported between some  $m^6A$  MTases and  $m^5C$  MTases, and this led to speculations about the location of the SAM binding site within the sequences (17).

Beginning with a set of twenty-seven DNA MTase sequences we have identified motifs which are present in all thirteen  $m^5C$  MTases. These motifs can serve as anchor points for their global alignment. Software has been developed that allows the detection of these motifs and can assist in the semi-automatic alignment of the sequences. The motifs have been examined extensively for their potential value in predicting  $m^5C$  MTase function.

## MATERIALS AND METHODS

### Sequences

Sequences for the following thirteen MTases which are known to form 5-methylcytosine were used. *BspRI* [recognition sequence: GGCC] (18), *BsuRI* [GGCC] (19), *HpaII* [CCGG] (R.J.R., unpublished), *MspI* [CCGG] (R.J.R., unpublished), *DdeI* [CTNAG] (7), *HhaI* [GCGC] (8), *EcoRII* [CC(A/T)GG] (9), *SinI* [GG(A/T)CC] (10), *HaeIII* [GGCC] (B. Slatko, personal communication) are components of Type II bacterial restriction-modification systems; Dcm [CC(A/T)GG] (A.S.B., unpublished) is encoded by the chromosome of *E. coli*; Phi3T [GGCC, GCNGC] (20), Rho1Is [GGCC, G(A/T/C)GC(T/A/G)C] (6,16) and SPR [GGCC, CCGG, CC(A/T)GG] (4,5) are MTases encoded by *Bacillus* phages.

Sequences were used for the following MTases which catalyze the formation of  $N^6$ -methyladenine. Modification and specificity subunits of Type I restriction-modification systems *EcoR124* (C. Price and T.A. Bickle, personal communication) and *EcoK* (21); *EcoRI* (22,23), *EcoRV* (24), *PstI* (25), *HhaII* (26) and H. Smith, personal communication), *PaeR7I* (27), *TaqI* (28) and *DpnII* (29) are components of Type II bacterial restriction-modification systems; modification subunits of Type III restriction-modification systems *EcoP1* and *EcoP15* (30); T4dam (31) and CviBIII (32) are MTases encoded by the phage T4 and by a *Chlorella* virus respectively, and the Dam (33) MTase is encoded by the *E. coli* chromosome. None of these three MTases have counterpart restriction enzymes. At the final stages of our analysis the sequences of *NgoPII* [GGCC] (34), a component of a Type II restriction-modification system, and of the mammalian  $m^5C$  MTase [GC] (35) became available.

Sequences for the following Type II restriction enzymes were also examined, *BsuRI* (19), *EcoRI* (22,23), *EcoRV* (24), *PstI* (25), *HhaII* (26), *PaeR7I* (27), *TaqI* (28), *DpnII* (29) and *MspI* (R.J.R., unpublished). The sequences of the restriction subunits of the four Type I and Type III enzymes indicated above were also included in the analysis. The following sequence databases were used: PIR (36) release 16; GenBank (37) release 56; EMBL (38) release 15.

### Homology detection

The initial estimates of similarities between pairs of sequences suspected to be related were based on scores by the program FASTP (39). When sequences of comparable length and composition are analyzed higher scores indicate closer similarity. A second program, RDF (39) was used to assess the significance of the individual FASTP scores. Sequences

in each pair were shuffled and the similarity of these randomized sequences was scored using FASTP. The mean and standard deviations of the scores from 150 such shuffles were calculated. This mean was used as a baseline for the comparison of the two real sequences. FASTP scores greater than three standard deviations above the mean were considered significant (39,40,41).

Dot matrix plots were used to display amino acid sequence similarities. Eleven residue long segments were compared by sliding two windows independently over the two sequences. The similarity of the two segments was scored by using the metrics of DIAGON (42).

### *Alignments*

To produce a global alignment of the set of similar sequences we developed a new procedure. A complete description of the algorithms used and their implementation will be presented elsewhere (J.P. and R.J.R., in preparation). In brief, information from both the amino acid and the nucleic acid sequences is used to produce the alignment. The program attempts to reproduce the method of *alignment by eye*, by directly locating globally conserved sequence features. First a search is made for the most significant patterns, which are common to all of the sequences to be aligned. A pattern is defined as a fixed sequence motif of specific amino acids and/or nucleotides which may or may not contain nonspecific positions. For example, YXXXGNxgr, describes a pattern containing a tyrosine at the first position, a glycine at the fifth position, an asparagine at the sixth position and a guanine in the middle of the seventh codon. Upper case letters denote amino acids and lower case letters denote nucleotides. An X or x indicates that any amino acid or nucleotide can occupy that position within the pattern. The length over which the pattern can extend is limited by a preset parameter, which is usually in the range of nine to eleven amino acids. This enables the detection of conserved residues in secondary structure units, such as residues on one face of an amphiphilic helix. Once a set of patterns that occur in all of the sequences has been found, then the one chosen for the initial alignment is based upon its significance, which is estimated from the number of the invariant nucleotide positions in the corresponding DNA sequences. The location of this most significant common pattern in the sequences determines the primary fixed line of the global alignment. Dividing every sequence into two halves by this line, two new sets of sequences are created. The global alignments of these secondary sets are then sought by recursively locating further common patterns, finding further fixed lines of the final alignment and creating new sets of sequence segments to be aligned. Initially the patterns consist of three specific amino acids within a window of up to eleven residues. Once these patterns are exhausted, then a lower stringency pattern consisting of two specific amino acids plus one or two specific nucleotides is used. Further reductions involve just two specific amino acids or one specific amino acid plus one or two specific nucleotides. The procedure stops for a particular set of sequence segments, when no significant common pattern is found. To finish the alignment, gaps are introduced between the fixed lines. No attempt at optimization is performed at this step. For convenience a single insertion is made in the middle of each non-aligned segment.

### *Pattern building and search routines*

The sequence segments around the fixed lines of alignment were used to build consensus patterns by following simple rules. Positions, where more different amino acids occurred in the aligned set of sequences than a preset limit were regarded as nonspecific. This limit varied from one to four. Thus consensus patterns contained positions in which either a single specific amino acid was allowed, positions in which any of two to four amino acids

were allowed and positions in which any amino acid was allowed. For example, the motif (P,T)XXXXXENV has two alternatives at the first position, any amino acid is allowed at the next five positions, and only single amino acids are allowed at the last three positions. A nonspecific position where more than two different amino acids occurs in the alignment is indicated here by X. The sequence databases were searched for these consensus patterns by a program (J.P., R.J.R., in preparation) using the DFS algorithm (43). Note that all sequences containing the specific tripeptide ENV preceded by either a T or a P six residues upstream would score as a match in the example above. Decreased specificity searches allowed mismatches at a limited number of specific positions within the search pattern. The significance of matches during the search was assessed by estimates of the expected number of hits in a random library. These rough estimates were calculated by combinatorial methods. If  $f_T$ ,  $f_P$ ,  $f_E$ ,  $f_N$  and  $f_V$  are the frequencies of amino acids T, P, E, N and V in the library which has Z overlapping 9-mers then the motif of the example will be expected to occur about  $o_{\text{calc}} = (f_T + f_P) * f_E * f_N * f_V * Z$  times in the library. Patterns which are not expected to occur by chance in the library ( $o_{\text{calc}} < 0.5$ ) are candidates for being predictive motifs. Finally, for a sequence motif to be considered predictive, it must occur only in the functionally related set of identified sequences from which it is derived.

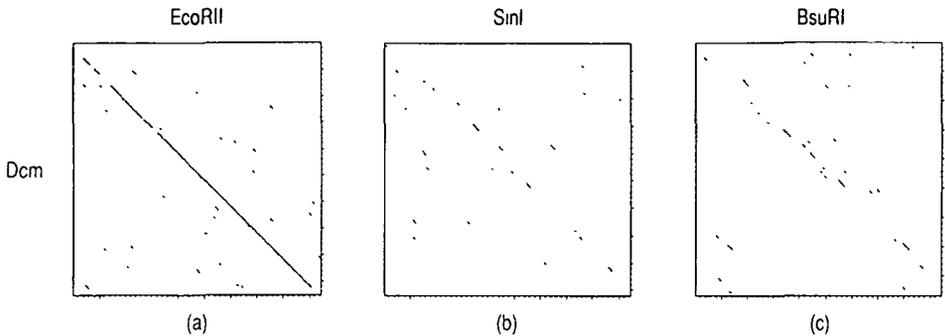
## RESULTS

### Pairwise comparisons

We have collected the sequences of twenty-seven prokaryotic DNA MTases. Twenty are components of restriction-modification systems—two of Type I, sixteen of Type II and two of Type III. The seven other DNA MTases have no known restriction enzyme counterpart. Three of these are encoded by Bacillus phages, one by bacteriophage T4, one by a Chlorella virus, one forms part of the mismatch repair system of *E. coli* and one is of unknown function in *E. coli*. Thirteen of the twenty-seven MTases catalyze the formation of 5-methylcytosine, the others form N<sup>6</sup>-methyladenine. Initially we used FASTP in pairwise comparisons to examine the sequence relatedness of these MTases.

When similarities between each pair of DNA MTase sequences are measured, the FASTP values scatter widely. There are a few scores above 1000 indicating great overall similarity (e.g. *EcoRII* scored 1666 when compared with *Dcm*; the comparison of *EcoRII* to itself gives 2405). However, the lowest scores are typical of the scores of nonrelated sequences. For example, *EcoRII* scores 34 with *HhaII*. Half of the sequences in the PIR database would give higher values when compared with *EcoRII*. When m<sup>5</sup>C MTases are compared to each other, the lowest score (between *EcoRII* and *DdeI*) is 70 while the majority of scores lie between 100 and 300. In contrast, 76 of the 91 m<sup>6</sup>A MTase pairs have scores under 70. Of the 182 mixed pairs (m<sup>5</sup>C vs. m<sup>6</sup>A MTase) only 6 have scores above the limit of 70. Comparisons between the MTases and their cognate restriction enzymes or between sequences of the complete set of all restriction enzymes yielded low scores. These results indicated that the group of m<sup>5</sup>C MTases is the only large subset of MTases with global similarity.

However, the randomization tests prove that even this group of sequences is quite diverse. Some sequences show very strong similarity to each other while others might even be unrelated. Examples of unquestionable relatedness include the sequences of two *E. coli* MTases, *EcoRII* and *Dcm* (Figure 1a), any two of the three *Bacillus* bacteriophage MTases (*SPR*, *Phi3T* or *Rho11s*) or the two MTases *BspRI* and *BsuRI*, encoded by *Bacillus sphaericus* and *Bacillus subtilis*. The opposite extreme in the relationships is exemplified



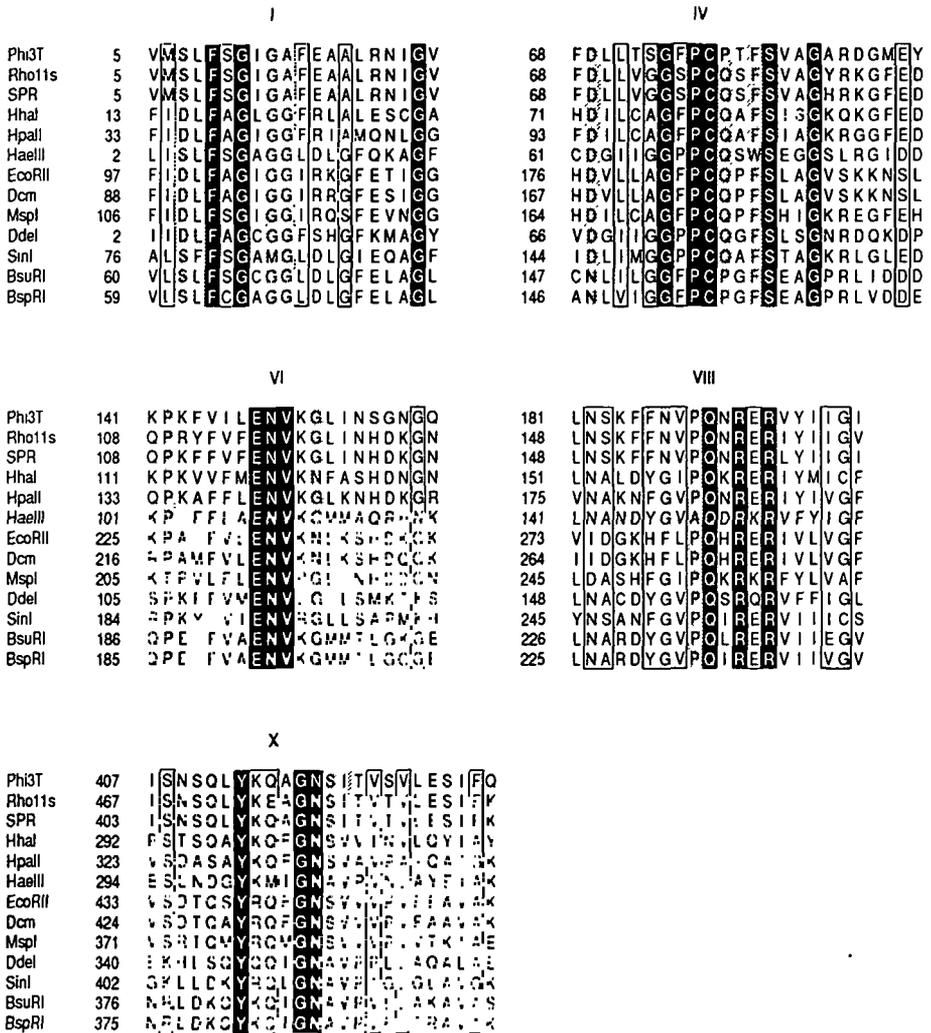
**Figure 1.** Sequence similarities between  $m^5C$  MTases. Dotmatrix comparisons of the sequence of Dcm (vertical axis) to the sequences of three other MTases (horizontal axes); (a) *EcoRII*, (b) *SinI* and (c) *BsuRI*. The DIAGON program with a window length of 11 and a threshold of 130 was used.

in the comparison of *SinI* to Dcm (Figure 1b). Using randomization tests as a measure of their relatedness, the RDF score of 1.7 is well below the value of 3.0, that is generally required as a safe lower limit for a claim of similarity (39,40,41). An analysis which relied solely on the pairwise comparison of the two sequences would suggest the absence of relatedness. In the present case it should be noted that 71 out of the 78 pairwise comparisons yielded RDF scores above the threshold of 3.0, while the lowest value was 0.7.

From dotmatrix plots it is apparent that the similarity between pairs of sequences is not uniform along the length of the sequences (Figure 1c). Patches of similarities and dissimilarities alternate. Because of this patchiness, and of the presence of the long variable N-terminal segments of the sequences, the available automatic multiple sequence alignment programs (44,45) were not able to provide a satisfactory global alignment among all thirteen sequences.

#### Global alignment

To provide a global alignment of the thirteen MTase sequences we developed a new procedure which resembles the method used manually to align sequences that show only patchy homology. Basically our new alignment procedure involves locating sequence patterns common to all sequences, and using these as anchor points for subsequent alignments. The initial application of this method showed that five well-conserved regions were present in all thirteen sequences. Within each of these conserved blocks at least 3 invariant amino acids occur (Figure 2). By searching the regions between these conserved blocks, using a less stringent similarity requirement, five more blocks of conserved sequences were found (Figure 3). These ten blocks provided the anchor points for alignment and arbitrary gaps were introduced as necessary to complete the final global alignment shown schematically in Figure 4. A total of twenty-one amino acids are invariant in the final alignment. Although the spacing between the blocks of similarities shows a definite overall regularity (Figure 4), there are clearly some subsets of sequences that deviate from this regularity. *BsuRI* and *BspRI* contains an insertion between conserved blocks I and II, *EcoRII* and Dcm have an insertion between blocks III and IV, Phi3T contains an insertion between blocks IV and V, and *SinI* has an exceptionally long connection between blocks VI and VII. Half of the sequences have a long and variable N-terminal arm preceding the first common motif. Only the closely related *EcoRII* and Dcm, *BspRI* and *BsuRI* and the three phage MTases show strong overall similarity to each other.



**Figure 2.** Sequences within the highly conserved blocks. Columns with white letters on black background indicate invariant positions. Columns with black letters on shaded background indicate positions where only two different amino acids occur. Columns in frames indicate positions where three different amino acids occur. The number to the left of each sequence indicates the absolute location of the first residue of the segment within the whole sequence. The numbering of the blocks corresponds to the sequential order of all conserved blocks on figure 4.

A notable feature of the sequences is the extremely variable region located between conserved blocks VIII and IX. This variable region is believed to be responsible for sequence recognition in the case of the phage MTases (14,15,16). This conclusion is based on an analysis of mutants that have lesions within this domain (15) and on the results of domain swap experiments (14,16). This region is significantly longer in the multiple recognition specificity MTases than in the single specificity MTases. It is about 210 amino acids in

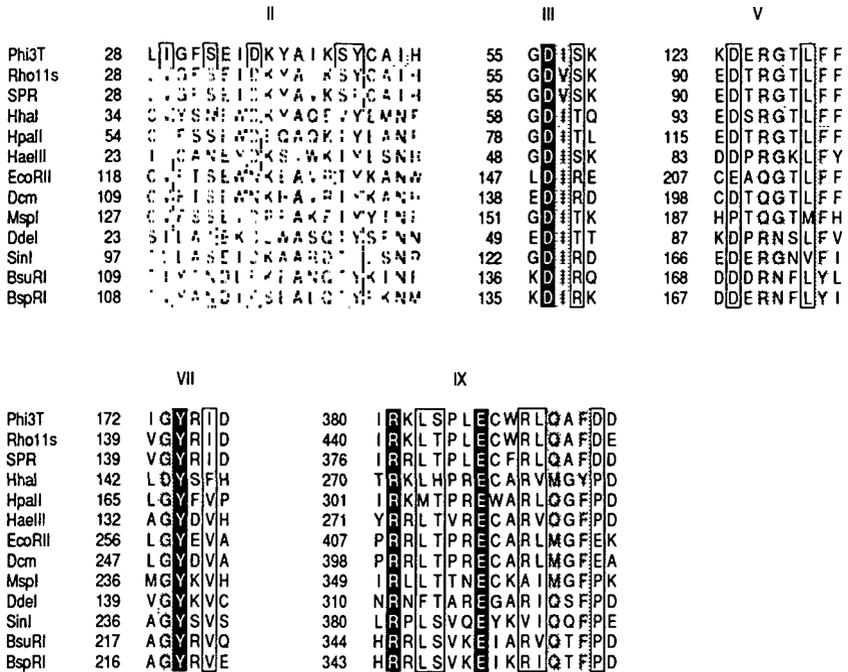


Figure 3. Sequences within the less conserved blocks. The notation is the same as on figure 2

length in the SPR MTase which recognizes three distinct sequences, is 275 and 180 amino acids long in the Rho11s and Phi3T MTases which are reported to recognize two distinct sequences and varies from 90 to 145 amino acids long in the case of the other, single specificity enzymes. When the variable regions are compared against one another, it is

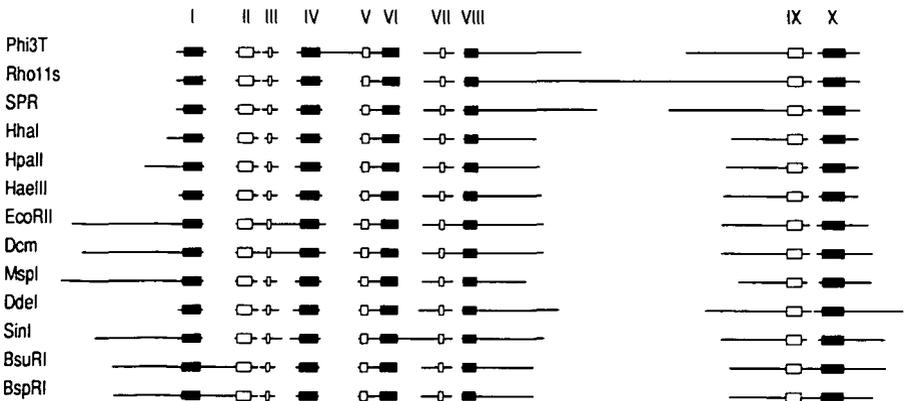
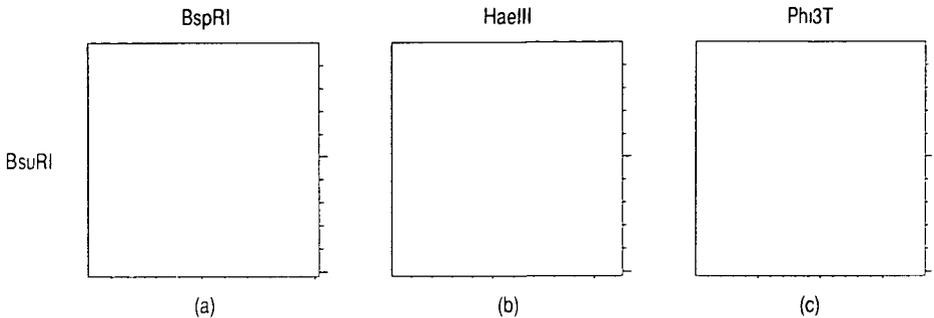


Figure 4. Schematic diagram of the alignment of the thirteen MTase sequences. Each line represents one sequence. Gaps were introduced in the alignment where the lines are interrupted. Boxes indicate where the ten blocks of conserved residues occur. Filled boxes indicate the five highly conserved blocks; the open boxes represent the five less conserved blocks. The variable region lies between blocks VIII and IX.



**Figure 5.** Sequence similarities within the variable regions of MTases recognizing GGCC. The variable region from the end of block VIII to the beginning of block IX of *BsuRI* (vertical axes) is compared to the variable regions of (a) *BspRI*, (b) *HaeIII* and (c) *Phi3T* (horizontal axes). The *DIAGON* program with a window length of 11 and a threshold of 130 was used.

found that some enzymes that recognize the same sequence such as *BsuRI*, *BspRI* and *HaeIII*, which recognize GGCC, show recognizable diagonals in pairwise comparisons (Figure 5a,b). This observation supports the hypothesis (14) that the variable region is responsible for the recognition specificity. However, strong sequence similarity of the variable regions is not a necessary condition for common specificity, since the variable regions of the three multiple specificity enzymes (*Phi3T*, *SPR*, *Rho11s*) which include GGCC as one of their recognition specificities do not exhibit significant similarity to the variable regions of *BspRI*, *BsuRI* or *HaeIII* (Figure 5c). Furthermore *MspI* and *HpaII*, which both recognize the sequence CCGG, differ considerably in their variable regions. This means that either different strategies are used to recognize the same nucleic acid sequence or similarity is not properly defined by concentrating only on amino acid identities.

#### Consensus patterns

Five of the conserved blocks (I, IV, VI, VIII and X in Figure 4) that were used as anchor points to align the MTase sequences contained at least three invariant aligned amino acids each. The sequences present in these blocks were used to define consensus patterns as described in Methods. From each block four different patterns were generated. The first pattern contained only the invariant amino acids within the block. The second pattern contained the invariant residues plus those positions at which two alternative amino acids could occur. The third (or fourth) pattern contained the elements of the second (or third) pattern plus the positions at which three (or four) alternative amino acids could occur. The significances of each of the twenty patterns (five blocks, 4 patterns each) were estimated using the combinatorial expression mentioned earlier. For each block the most conservative pattern (allowing the least variability at each position) with  $\alpha_{\text{calc}} < .5$  is listed in Figure 6. These motifs were used as search patterns against the PIR database. The five  $m^5C$  MTases already present in the database were the only hits. This was true for each of the five motifs and showed that they were able to discriminate between the  $m^5C$  MTases and all other identified proteins in this database. In particular it should be noted that these motifs are not present in any of the sequences of the MTases that form N-6-methyladenosine. This suggested that these motifs might be used to identify new  $m^5C$  MTases.

This was tested by examining the sequence of the *NgoPII*  $m^5C$  MTase which has been determined recently (35), and was not included in the construction of the motifs. This

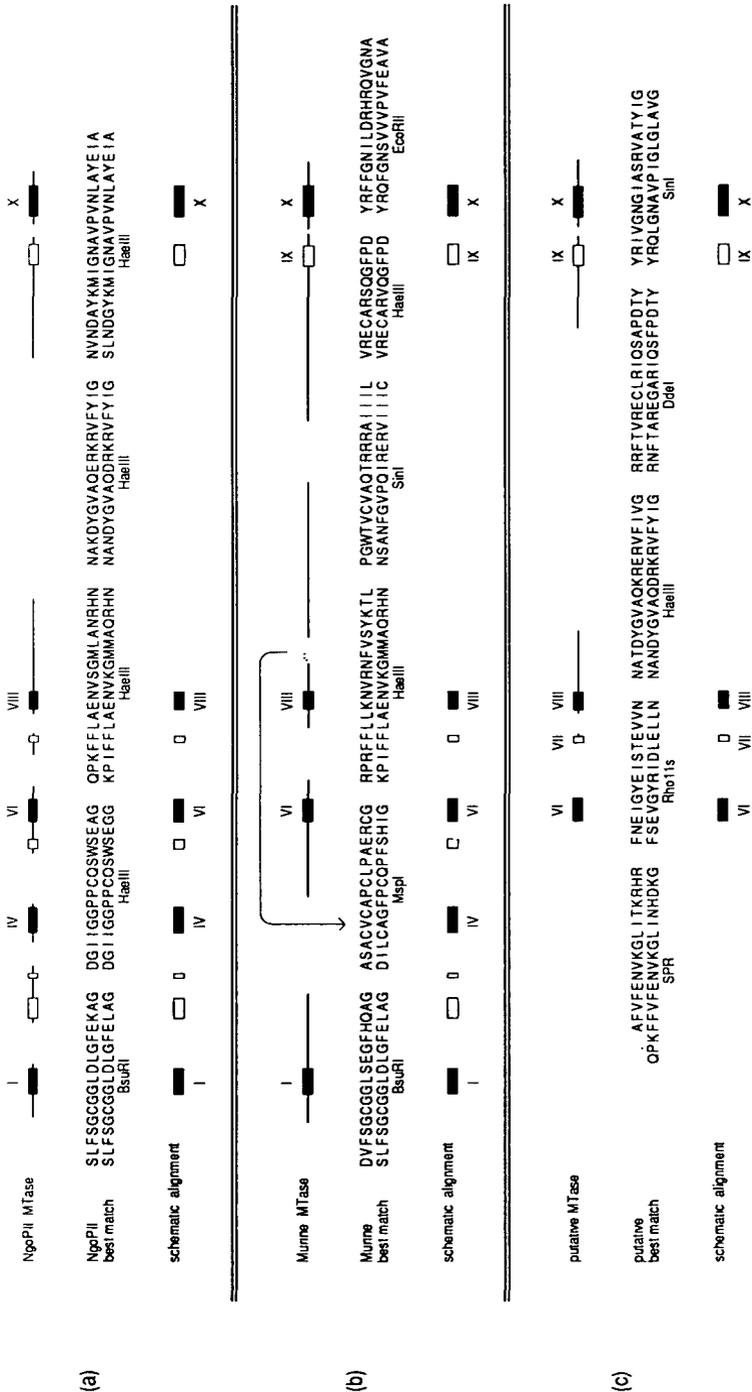
Block number	Motifs
I	DFF - G - GA - - - - - G SL      MG
IV	D - - - - G - PCP - FS - - G N                      Q W
VI	KP - - FF - ENVKGF - A - - - G QT    II            LNI N    K R    LL            P L S    N S    VV            R M T
VI*	KP - - FF - ENV - GF - A - - - G QT    II            NI N    K R    LL            L S    N S    VV            M T
VIII	DA - - FFIAQ - RER - - - EA ID    HGLP      K      IC NS    YNV      Q      VG
X	K - - - - YKE - GNAI - I - A - - - A R      QM      SV P L      F S      RQ            V V      G
X*	YKE - GNAI - I - A - - - A QM      SV P L      F RQ            V V      G

**Figure 6.** Predictive sequence motifs of m<sup>5</sup>C MTases. The block number corresponds to the numbering of figure 4. At positions where more than one amino acid is acceptable the alternatives are listed. Dashes signify that any amino acid can occupy that position. Blocks VI\* and X\* are the modified motifs necessary to accommodate the *Ngo*PII sequence.

sequence also contains three of the five motifs identified above (Figure 7a). The ‘ENV’ and ‘Y ♦ GN’ motifs (the abbreviations within the quotes denote the complete motifs derived from blocks VI and X, ♦ marks a specified distance between two conserved amino acids) are found only when a decreased specificity search is used. At the specific position following the invariant ENV triplet of the motif the serine in *Ngo*PII is not allowed. (A single base change from C to A or G in the corresponding codon would restore the perfect match.) Similarly a mismatch at a single specific position causes the imperfection in the match of *Ngo*PII to the ‘Y ♦ GN’ motif. To include these two motifs as discriminators of all 14 MTases the original conservative definitions need to be relaxed as indicated in Figure 6. The updated motifs were tested against the database as before. They still provide complete discrimination. The other five less conserved blocks are also found in the *Ngo*PII sequence, all of them in the correct order. Indeed the sequence can be aligned very satisfactorily with the thirteen sequences included in our study (Figure 7a).

*The prokaryotic motifs in a mammalian methyltransferase*

It was of interest to search the sequence of the eukaryotic MTase which has recently become available (35), for the presence of the bacterial m<sup>5</sup>C MTase motifs. Four of the five motifs



**Figure 7.** Other sequences containing the methylase motifs. The top lines are the schematic diagrams of the sequences of (a) the *Ngo*Pll MTase, (b) the murine MTase, and (c) a peptide in the corrected sequence of GenBank entry M13488 (c), while the bottom lines show the generalized arrangements of the thirteen bacterial sequences. Filled and open boxes show the locations of the common blocks; the shaded box in (b) indicates the location of the segment containing the PC dipeptide, which shows the closest similarity to the conserved bacterial G-PC---S-G motif. The individual sequence segments of the numbered blocks are shown aligned with the best matching segments of the bacterial sequences. The order of the segments corresponds to the numbering of blocks in the bacterial sequences.

are indeed found in the C-terminal half of the sequence (Figure 7b), when the decreased specificity mode of our search program is used. The invariant triplets (F-G-----G, Q-R-R and Y---GN) of three of the basic blocks (I, VIII and X) are found. Two of the three invariant residues of basic block VI (NV from ENV) are present in the murine sequence, although the match in this region to the other positions of the motif is very good. Some of the five less-specific blocks found in the bacterial sequences can also be identified. For example, within one stretch of twelve amino acids, eleven are identical between *HaeIII* and the murine enzyme (VRECAR[V/S]QGFPD). These lie within conserved block IX. All the above listed patterns lie in the same order as in the bacterial sequences (Figure 7b). Surprisingly the segment of the murine MTase which is proposed (35) to contain the conserved PC sequence and would correspond to the basic block associated with elements of the functional site (block IV) does not fit into this arrangement. In the mammalian sequence the pattern of three of the five invariant amino acids from the block (PC-----G of G-PC---S--G) is located between blocks VIII and IX, which would correspond to the variable region of the prokaryotic MTases. Since it is the variable region that is believed to be responsible for sequence recognition it is entirely reasonable that a catalytic functional unit should lie adjacent to it. Perhaps the bacterial MTases achieve this juxtaposition at the structural level by folding. This may reflect the fact that the bacterial enzymes show exquisite sequence specificity, whereas the mammalian enzyme appears much less fastidious in its sequence recognition.

#### *Potential new methyltransferases*

Given the ability of the MTase sequence motifs to discriminate between m<sup>5</sup>C MTases and all other proteins of known function it was natural to ask if these motifs could be found in any of the previously unidentified open reading frames present in the GenBank or EMBL databases. Searches were made against the six-frame translated versions of both databases using each of the five motifs.

In one entry (M13488) of the unannotated division of GenBank, the 'Q♦R♦R' motif of conserved block VIII appears twice. One of these occurrences is within the known coding region of the bacteriophage Phi3T MTase, which is one of the sequences used to define the motif. However, the second occurrence lies in an open reading frame immediately upstream from the start codon of the known Phi3T MTase. This open reading frame has not previously been assigned a function. Comparing the translated products of this DNA fragment with other MTase sequences in a dotmatrix, it is apparent that one reading frame shows strong similarity beginning within conserved block VI and continuing through blocks VII and VIII. A long non-conserved region follows and with a reading frame change the similarity picks up again with block IX and the beginning of block X. The end of block X can be found by changing back to the original reading frame. Since the sequence information available for this open reading frame only begins within block VI we do not know if the similarity continues upstream. While trying to assess the significance of these similarities which involved changes of reading frame we were gratified to learn that they were the results of errors, present in the original sequence and hence in the GenBank entry, but which had subsequently been corrected (legend to figure 3 in reference 6). After the sequence corrections, the peptide from this reading frame can be aligned very satisfactorily with the known m<sup>5</sup>C MTases (Figure 7c). It should be noted that this new sequence is not simply a direct duplication of the established Phi3T MTase gene. The two sequences are not especially similar to each other and in pairwise comparison with the putative MTase, *BsuRI* MTase and others give higher scores.

A second unidentified open reading frame was also found during the search with the motifs. This was present in the DNA fragment which contains the coding sequence for the *B. subtilis* phage Rho11s MTase (6). Again a putative second MTase appears to be present immediately upstream of the gene already characterized. The sequence is identical, at the nucleotide level, with that of the Phi3T secondary MTase for the length of the published sequence. This identity covers the variable regions between blocks VIII and IX suggesting that both putative MTases have the same recognition specificity. This variable region shows no distinct similarity with the variable region of any other MTase, and so its specificity cannot be predicted. Based on its length of about 100 amino acids, it probably encodes a single specificity.

## DISCUSSION

Analysis of the sequences of m<sup>5</sup>C MTases has revealed a common architecture. This consists of ten conserved blocks of amino acid residues with variable length N-terminal and C-terminal arms. Most of the conserved blocks are separated by short regions of similar length, but with little sequence conservation. Clusters of invariant positions enclose about two hundred amino acids in the N-terminal half of each sequence and about fifty residues in the C-terminal half. A long variable region of 90–275 residues separates conserved blocks VIII and IX and it is likely that this region encodes the protein domain responsible for sequence specificity. The absence of significant sequence similarity between the MTases and the sequences of known DNA binding proteins such as the helix-turn-helix structures (46), zinc fingers (47) or the leucine zippers (48) suggests that DNA recognition takes place by a different mechanism. While the binding of DNA by some of the MTases could be facilitated by the N-terminal arms in a manner similar to that used by lambda repressor (49) or the restriction enzyme *EcoRI* (50), this cannot be a universal feature.

Examination of the sequences of the four available cognate restriction enzymes (*BsuRI*, *SinI*, *MspI* and *DdeI*) shows no similarities among themselves nor with any of their cognate MTases. This is also true for other systems for which sequence information is available and indicates that the complementary components of Type II restriction-modification systems must have evolved separately. This raises the difficult issue as to how the correct pairs of specificities ever found one another. However, if the genes of the components were acquired and incorporated into the genome in separate events this would explain the variation in their genetic organization, relative positions, distances and orientations. It is somewhat surprising that none of the motifs which are characteristic of the m<sup>5</sup>C MTases are found within the sequences of the m<sup>6</sup>A MTases. Since both sets of enzymes share the ability to bind SAM and to transfer the methyl group to DNA, this suggests that several mechanisms must exist to accomplish these common events. The substantial differences probably indicate different evolutionary histories.

The sequence motifs derived from the five best-conserved common blocks are able to discriminate the m<sup>5</sup>C MTases from all other proteins of known sequence. By searching unidentified open reading frames in DNA databases for the presence of these short motifs we were able to detect putative new MTases. During these comparisons it was possible to pinpoint several suspected sequencing errors. The observed homology between the m<sup>5</sup>C MTases and one of the deduced products of a DNA sequence already in the sequence database suddenly disappeared but picked up in another reading frame, hinting at a possible error. This suggests a computational procedure for localizing sequence errors. A program able to scan new sequences for possible homology with sequences already in the databases

at the translated level in all reading frames would be a useful tool. It could be used to highlight any regions of a newly determined sequence that should be checked carefully for possible errors.

The putative MTases that we suggest are encoded by the *B. subtilis* phages Phi3T and Rho1s lie immediately upstream of the known multi-specific MTases of these two phages. Biochemical confirmation of the presence of the putative methyltransferases remains to be done. Having more functional MTases on the same phage would be a simple way to achieve protection against restriction enzymes with different specificity. The tandem configurations of two MTases could be precursors to MTases with multiple specificities.

The motifs that we have found within the m<sup>5</sup>C MTases are probably not directly involved in sequence specific recognition of DNA. Rather it is likely that they define structural elements important for proper folding or elements important for function. In the case of functional motifs it is conceivable that closely related versions of the motifs may also be present in other proteins with related functions such as the RNA MTases that induce resistance to antibiotics, or proteins that interact with the methyl donor S-adenosylmethionine. We have also observed that proteins interacting with glutathione (glutathione reductases and peroxidases) contain patterns that are related to the motifs we have found in these m<sup>5</sup>C MTases. Interestingly glutathione transferases and synthetases contain the DPPY motif that is characteristic of the m<sup>6</sup>A MTases (17).

Since no two of the sequences used in the analysis are identical, each of them contributes to the divergency of the set. For example, if *SinI* were omitted from the analysis, the consensus pattern of the remaining twelve sequences would have four invariant positions in block I, instead of three. Similarly, a decrease in the level of observed conservation, possibly even within the blocks of similarities, is expected with the addition of every new sequence. This is demonstrated by the imperfect matches of some of the motifs to the corresponding segments in the *NgoPII* MTase sequence. This means, that when the motifs are located within new sequences, the similarity blocks upon which the motifs are based will need to be expanded and new consensus patterns must be calculated. These can in turn be tested for their continued predictive power. With the methods described in this paper the step of selecting the functional set of sequences needs human decision, all the other steps are or could be performed automatically. We are currently testing the idea that the algorithms and programs described here can be used to construct and maintain a database of sequence motifs that will be useful in predicting the function of newly determined sequences.

## ACKNOWLEDGEMENTS

We thank Drs. W.R. Pearson and R.F. Doolittle for making their programs available to us. Drs. T. Bestor, B. Slatko, H. Smith, T. Bickle and C. Price were kind enough to release the sequences of the murine MTase, of the *HaeIII* MTase, of the *HhaII* MTase and of the *EcoR124* system, prior to publication. It is a pleasure to acknowledge extremely useful discussions with Drs. G. Otto and R.F. Doolittle. This work was supported by grants from the NSF (DMB-8614032 to RJR) and NIH (LM 04971 to RJR and GM39715 to W. Szybalsky).

\* Present address: Wayne State University, 435 Chemistry, Detroit, MI 48202, USA

Permanent addresses: <sup>+</sup> Institute of Biophysics and <sup>o</sup>Institute of Biochemistry, Biological Research Centre, Hungarian Academy of Sciences, Szeged, PO Box 521, 6701 Hungary

## REFERENCES

1. Adams, R.L.P. and Burdon, R.H. (1985) 'Molecular Biology of DNA Methylation'. Springer-Verlag, New York.
2. Razin, A., Cedar, H. and Riggs, A.D. (1984) 'DNA Methylation. Biochemistry and Biological Significance'. Springer-Verlag, New York.
3. Marinus, M.G. (1976) *J. Bacteriol.* 128, 853–854.
4. Buhk, H.J., Behrens, B., Taylor, R., Wilke, K., Prada, J.J., Gunthert, U., Noyer-Weidner, M., Jentsch, S. and Trautner, T.A. (1984) *Gene* 29, 51–61.
5. Pósfai, G., Baldauf, F., Erdei, S., Pósfai, J., Venetianer, P. and Kiss, A. (1984) *Nucl. Acids Res.* 12, 9039–9049.
6. Behrens, B., Noyer-Weidner, M., Pawlek, B., Lauster, R., Balganesch, T.S. and Trautner, T.A. (1987) *EMBO J.* 6, 1137–1142.
7. Szynter, L.A., Slatko, B., Moran, L., O'Donnell, K.H. and Brooks, J.E. (1987) *Nucl. Acids Res.* 15, 8249–8266.
8. Caserta, M., Zacharias, W., Nwankwo, D., Wilson, G.G. and Wells, R.D. (1987) *J. Biol. Chem.* 262, 4770–4777.
9. Som, S., Bhagwat, A.S. and Friedman, S. (1987) *Nucl. Acids Res.* 15, 313–332.
10. Karreman, C. and de Waard, A. (1988) *J. Bacteriol.* 170, 2527–2532.
11. Chandrasegaran, S. and Smith, H.O. (1988) in R.H. Sarma and M.H. Sarma (eds.), 'From Proteins to Ribosomes', pp. 149–156. Adenine Press.
12. Wu, J.C. and Santi, D.V. (1987) *J. Biol. Chem.* 262, 4778–4786.
13. Santi, D.V. and Brewer, C.F. (1973) *Biochemistry* 12, 2416–2424.
14. Balganesch, T.S., Reiners, L., Lauster, R., Noyer-Weidner, M., Wilke, K. and Trautner, T.A. (1987) *EMBO J.* 6, 3543–3549.
15. Wilke, K., Rauhaut, E., Noyer-Weidner, M., Lauster, R., Pawlek, B., Behrens, B. and Trautner, T.A. (1988) *EMBO J.* 7, 2601–2609.
16. Trautner, T.A., Balganesch, T.S. and Pawlek, B. (1988) *Nucl. Acids Res.* 16, 6649–6658.
17. Lauster, R., Kreibardis, A. and Guschlbauer, W. (1987) *FEBS Lett.* 220, 167–176.
18. Pósfai, G., Kiss, A., Erdei, S., Pósfai, J. and Venetianer, P. (1983) *J. Mol. Biol.* 170, 597–610.
19. Kiss, A., Pósfai, G., Keller, C.C., Venetianer, P. and Roberts, R.J. (1985) *Nucl. Acids Res.* 13, 6403–6421.
20. Tran-Betcke, A., Behrens, B., Noyer-Weidner, M. and Trautner, T.A. (1986) *Gene* 42, 89–96.
21. Loenen, W.A.M., Daniel, A.S., Braymer, H.D. and Murray, N.E. (1987) *J. Mol. Biol.* 198, 159–170.
22. Newman, A.K., Rubin, R.A., Kim, S. and Modrich, P. (1981) *J. Biol. Chem.* 256, 2131–2139.
23. Greene, P.J., Gupta, M., Boyer, H.W., Brown, W.E. and Rosenberg, J.M. (1981) *J. Biol. Chem.* 256, 2143–2153.
24. Bougueleret, L., Schwarzstein, M., Tsugita, A. and Zabeau, M.Z. (1984) *Nucl. Acids Res.* 12, 3659–3676.
25. Walder, R.Y., Walder, J.A. and Donelson, J.E. (1984) *J. Biol. Chem.* 259, 8015–8026.
26. Schoner, B., Kelly, S. and Smith, H.O. (1983) *Gene* 24, 227–236.
27. Theriault, G., Roy, P.H., Howard, K.A., Benner, J.S., Brooks, J.E., Waters, A.F. and Gingeras, T.R. (1985) *Nucl. Acids Res.* 13, 8441–8461.
28. Slatko, B.E., Benner, J.S., Jager-Quinton, T., Moran, L.S., Simcox, T.G., Van Cott, E.M. and Wilson, G.G. (1987) *Nucl. Acids Res.* 15, 9781–9796.
29. Mannarelli, B.M., Balganesch, T.S., Greenberg, B., Springhorn, S.S. and Lacks, S.A. (1985) *Proc. Natl. Acad. Sci. U.S.A.* 82, 4468–4472.
30. Humbelin, M., Suri, B., Rao, D.N., Hornby, D.P., Eberle, H., Pripfl, T., Kenel, S. and Bickle, T.A. (1988) *J. Mol. Biol.* 200, 23–29.
31. MacDonald, P.M. and Mosig, G. (1984) *EMBO J.* 3, 2863–2871.
32. Narva, K.E., Wendell, D.L., Skrdla, M.P. and Van Etten, J.L. (1987) *Nucl. Acids Res.* 15, 9807–9823.
33. Brooks, J.E., Blumenthal, R.M. and Gingeras T.R. (1983) *Nucl. Acids Res.* 11, 837–851.
34. Sullivan, K.M. and Saunders, J.R. (1988) *Nucl. Acids Res.* 16, 4369–4387.
35. Bestor, T., Laudano, A., Mattaliano, R. and Ingram, V. (1988) *J. Mol. Biol.* 203, 971–983.
36. Sidman, K.E., George, D.G., Barker, W.C. and Hunt, L.T. (1988) *Nucl. Acids Res.* 16, 1869–1871.
37. Bilofsky, H.S. and Burks, C. (1988) *Nucl. Acids Res.* 16, 1861–1863.
38. Cameron, G.N. (1988) *Nucl. Acids Res.* 16, 1865–1867.
39. Lipman, D.J. and Pearson, W.R. (1985) *Science* 227, 1435–1441.
40. Feng, D.F., Johnson, M.S. and Doolittle, R.F. (1985) *J. Mol. Evol.* 21, 112–125.
41. Argos, P. (1987). *J. Mol. Biol.* 193, 385–396.

42. Staden, R. (1982) *Nucl. Acids Res.* 10, 2951–2961.
43. Swamy, M.N.S and Thulasiraman, K. (1981) 'Graphs, Networks, and Algorithms'. John Wiley & Sons, New York.
44. Martinez, H.M. (1988) *Nucl. Acids Res.* 16, 1683–1691.
45. Feng, D.F. and Doolittle, R.F. (1987) *J. Mol. Evol.* 25, 351–360.
46. Sauer, R.T., Yocum, R.R., Doolittle, R.F., Lewis, M. and Pabo, C.O. (1982) *Nature* 289, 447–451.
47. Evans, R.M. and Hollenberg, S.M. (1988) *Cell* 52, 1–3.
48. Landschulz, W.H., Johnson, P.F. and McKnight, S.L. (1988) *Science* 240, 1759–1764.
49. Pabo, C.O., Krovatin, W., Jefferey, A. and Sauer, R.T. (1982) *Nature* 298, 441–443.
50. McClarin, J.A., Frederick, C.A., Wang, B.-C., Greene, P., Boyer, H., Grable, J. and Rosenberg, J.M. (1986) *Science* 234, 1526–1541.

**This article, submitted on disc, has been automatically  
converted into this typeset format by the publisher.**