Predictive motifs derived from cytosine methyltransferases

Cold Spring Harbor Laboratory, PO Box 100, Cold Spring Harbor, NY 11724 and [1]University of

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 *Hha*I 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

for their global alignment. Software has been developed that allows the detection of these

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).

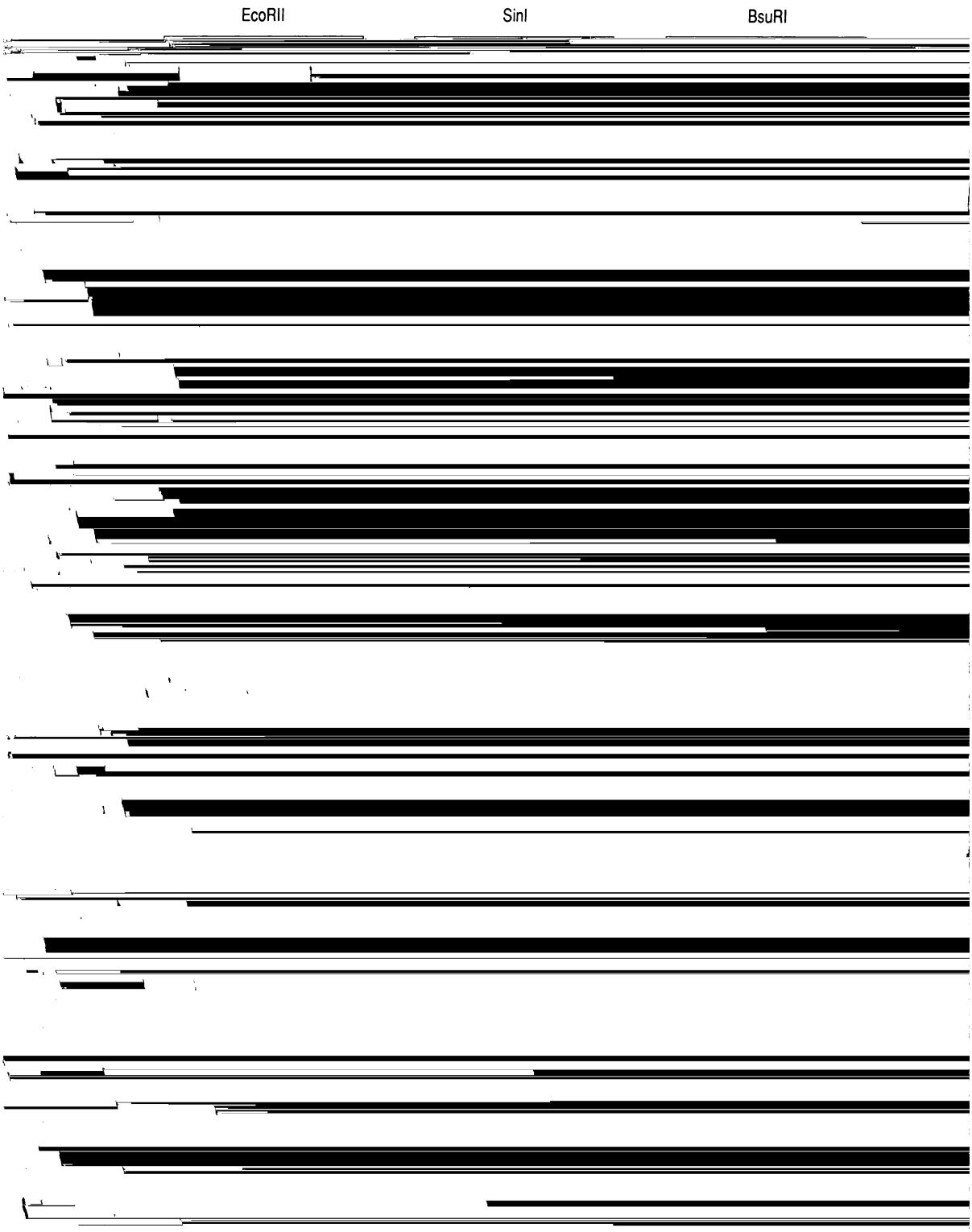
Dot matrix plots were used to display amino acid sequence similarities. Eleven residue long segments were compared by sliding two windows independently over the two sequences. The similarity of the two segments was scored by using the metrics of DIAGON

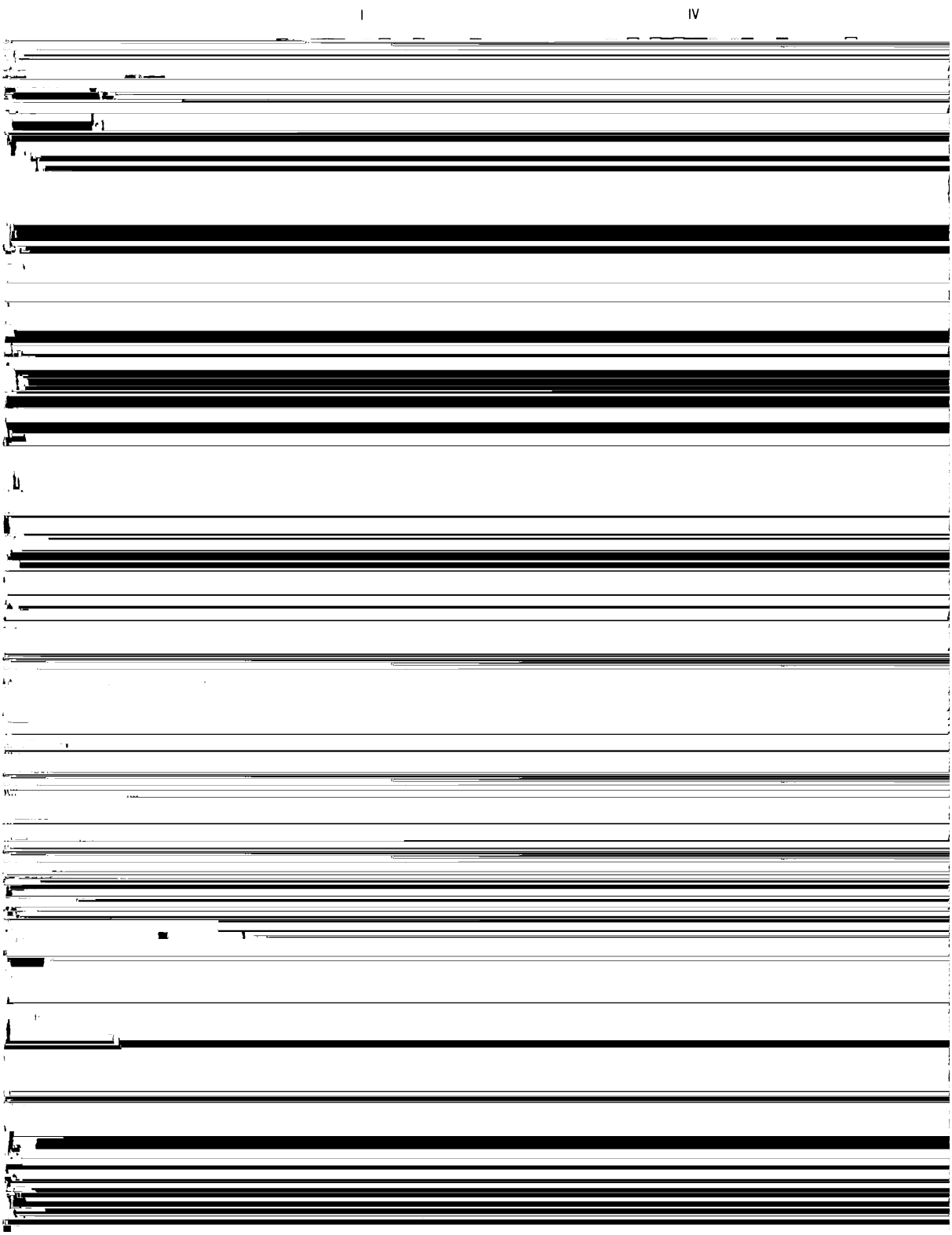
Alignments

were allowed and positions in which any amino acid was allowed. For example, the motif (**P,T**)*XXXXX***ENV** 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.

[redacted]

is indicated here by *X*. The sequence databases were searched for these consensus patterns


EcoRII
SinI
BsuRI


I
IV


Phi3T 28 L I G F S E I D K Y A I K S Y C A I H 55 G D I S K 123 K D E R G T L F F
Rho11s 28 L V G F S E I D K Y A I K S Y C A I H 55 G D V S K 90 E D T R G T L F F
SPR 28 L V G F S E I D K Y A V K S F C A I H 55 G D V S K 90 E D T R G T L F F
Hhal 34 C V Y S N E W D K Y A Q E V Y E M N F 58 G D I T Q 93 F D S R G T L F F

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 *Bsu*RI (vertical axes) is compared to the variable regions of (a) *Bsp*RI, (b) *Hae*III 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 *Bsu*RI, *Bsp*RI and

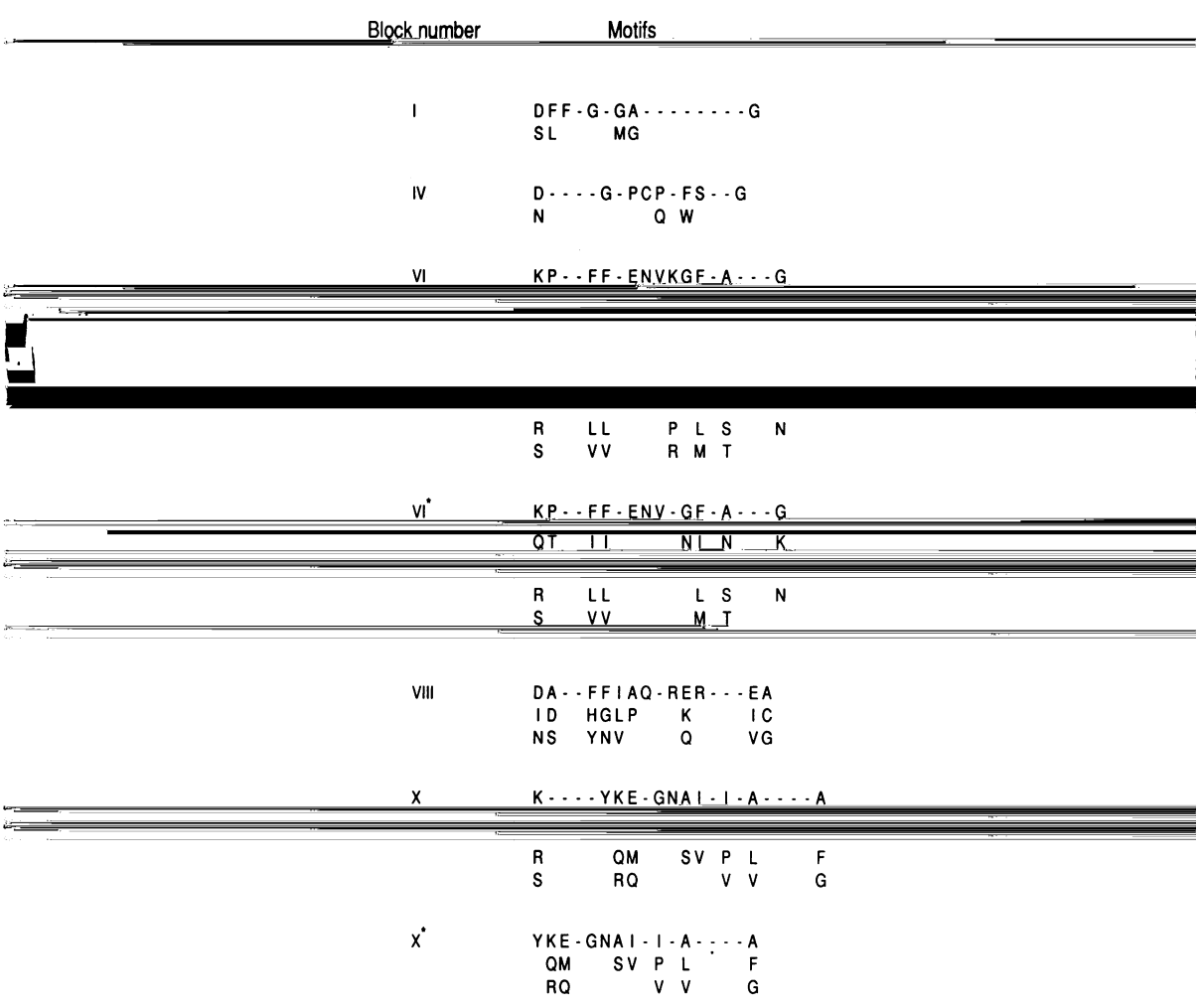
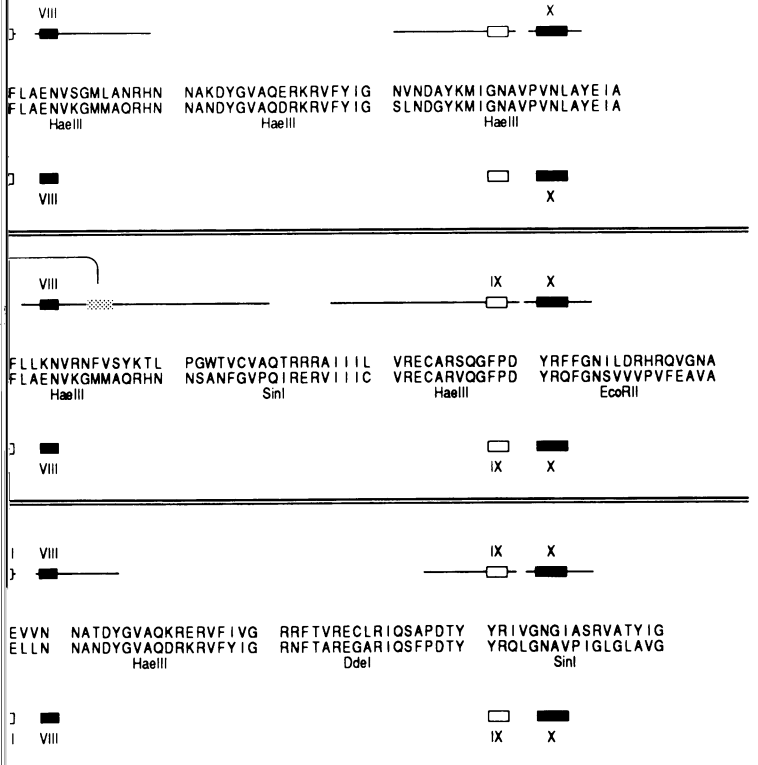


Figure 6. Predictive sequence motifs of m^5C MTases. The block number corresponds to the numbering of figure 4. At positions where more than one amino acid is acceptable the alternatives are listed. Dashes signify that any amino acid can occupy that position. Blocks VI* and X* are the modified motifs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



re the schematic diagrams of the sequences of (a) the *Ngo*PII MTase, (b) the murine 3488 (c), while the bottom lines show the generalized arrangements of the thirteen non blocks; the shaded box in (b) indicates the location of the segment containing rial **G-PC---S--G** motif. The individual sequence segments of the numbered blocks s. The order of the segments corresponds to the numbering of blocks in the bacterial

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**,

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.

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

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

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 Rho11s lie immediately upstream of the known multi-specific MTases of these two phages.

be done. Having more functional MTases on the same phage would be a simple way to

REFERENCES

1. Adams, R.L.P. and Burdon, R.H. (1985) 'Molecular Biology of DNA Methylation'. Springer-Verlag, New

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.

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,

6. Behrens, B., Noyer-Weidner, M., Pawlek, B., Lauster, R., Balganesh, T.S. and Trautner, T.A. (1987)

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.

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.