Volume 329, number 3, 233–237 FEBS 12871 August 1993

Minireview

5-Methylcytosine in genes with methylation-dependent regulation

Pietro Volpe[a,b], Paolo Iacovacci[a,b], Richard H. Butler[c] and Tamilla Eremenko[b]

[a]*Department of Biology, University of Rome 'Tor Vergata', Rome, Italy,* [b]*Laboratory of Molecular Biology, Institute of Experimental Medicine, CNR, Rome, Italy* and [c]*Institute of Cell Biology, CNR, Rome, Italy*

Received 6 July 1993

An asymmetric distribution of deoxy-5-methylcytidylic acid-inhibiting restriction sites (dcm-sites) takes place in ten human genes regulated by 5-methylcytosine. These genes are dcm-site enriched upstream and dcm-site poor downstream. Along them, there is a scattering of hypermethylatable introns and hypomethylatable exons with a common code. the 5mCpG dinucleotides characterize promoters; Gp5mCs characterize introns; Tp5mCs and Cp5mCs are in small concentrations in exons. Housekeeping genes contain more dcm-sites when compared with tissue-specific genes. This depends on the higher number of dcm-sites in their promoters and introns. In exons, the relatively lower number of dcm-sites is almost the same in both housekeeping and tissue-specific genes. Going from 5′ to 3′, the average frequency of occurrence of these sites per nucleotide units decreases in introns and increases in exons. This difference is highly discriminated for tissue-specific and less discriminated for housekeeping genes.

5-Methylcytosine; Deoxy-5-methylcytidylic acid; Modified code; Housekeeping gene; Tissue-specific gene

1. INTRODUCTION

It was found that 5-methylcytosine (5mC) is mainly concentrated on promoter and uncoding sequences, while coding sequences contain few 5mCs, if any [1]. In accordance with other information [2,3], this suggested that 5mC might function as a signal in a hypothetical 'modified code' to control transcription [1–6]. In fact, there was a dramatic time-shift between DNA methylation, occurring in S [2], and bulk transcription, mainly occurring in G_1 and G_2 [7]. Translation also showed maximal rates in G_1 and G_2 [8]. Such an ordering during the cell life of the maximal rates of DNA methylation, transcription and translation accounted, thus, for a possible inverse correlation between gene methylation and expression [3]. At the end of the '70s, the literature confirmed that several single genes are expressed if unmethylated and non-expressed if methylated [9–12]. Today, the genes in which methylation is believed to be involved in regulation of transcription are more than fifty. This led us to the search for a language that they should have in common so as to be recognized by the DNA methylase system [13–15].

In this contribution, three HK and seven TS human genes (Table I) have been selected among those which are unequivocally switched-on when unmethylated and unequivocally switched-off when methylated [16–24]. Along these genes, an asymmetric distribution of dcm-sites was found [14,15]. This distribution depends not only on the class of genes (HK or TS) but also on the type of nucleotide sequences inside the same transcriptional unit (promoter, introns and exons). In harmony with the above-mentioned scattered nature of methylation showing in the gene the intermittence of 'hypermethylated domains' and 'hypomethylated islands' [1], it was observed that proceeding in the 5′–3′ direction the analysis of the frequence of dcm-sites, normalised per nucleotide units, allows some deciphering of the 5mC code. Along the double helix, at least four pairs of dinucleotides – two monomethylatable (Tp5mC/ApG; Cp5mC/GpG) and two dimethylatable in *cis–trans* (5mCpG/Gp5mC; Gp5mC/5mCpG) – were identified [15].

2. '5mCpG/Gp5mC' AS A KEY WORD IN THE GENE MODIFIED CODE

The starting point was the consideration that the methylated DNA sequences which are not cut by the dcm-inhibited restriction endonucleases previously functioned as recognition sites for the corresponding bacterial dC-DNA methylases. Since 5mC in eukaryotic DNA is the sole methylated nitrogen base, the presence of these sites in human genes could provide the key to the detection of some variability in the codes available for eukaryotic DNA methylases [13]. Not all dcm-sites were found to be present along the genes taken into consideration (Table II). None of the investigated gene contains dcm-sites for *Sal*I which does, however, cut the globin genes. By contrast, the TK and HPRT genes contain sites for 18 restrictases. This property of the TK

Correspondence address: P. Volpe, Institute of Experimental Medicine, CNR, Marx Street 15, 00137 Rome, Italy.

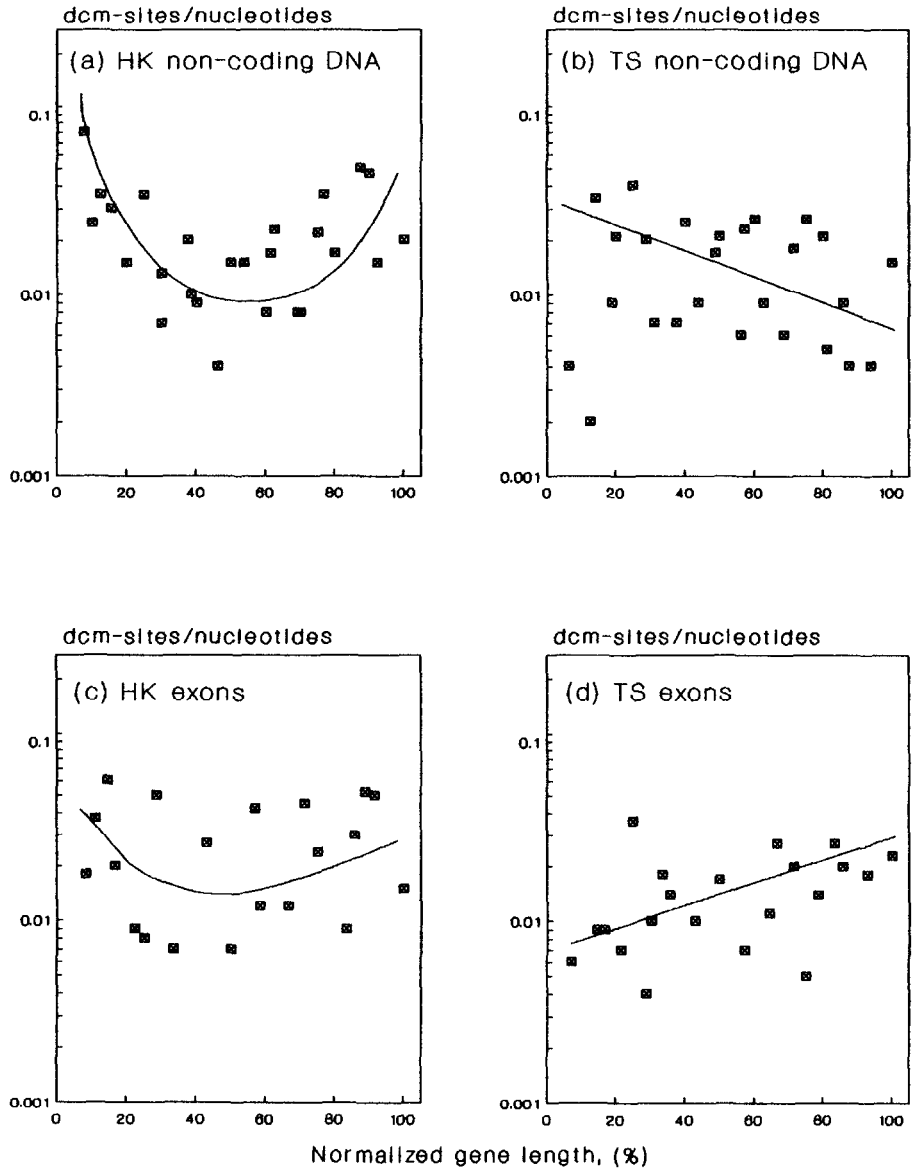

Fig. 1. Average frequency of dcm-sites per nucleotide units along the HK and TS uncoding and coding sequences. HKs and TSs were all those of Table I. The position of sequences along the abscissa was normalized on the basis of their lengths in nucleotide bp [15]. For each gene, the total number of uncoding and coding sequences was considered equal to 100%, so that, if a gene contained 3 exons, for instance, they were localised, along it, in correspondence of 33.3%, 66.6% and 99.9% relative length values (0.1% was not considered). Actually the order number of each single sequence was divided by the total number of sequences present in a given gene according to the expression: (sequence number/total number of sequences) × 100. The average frequence of the dcm-sites is shown on a logarithmic ordinate scale. (a,b) The uncoding sequences, besides the introns, also include the 5′ and the 3′ flanking regions. (c,d) The coding sequences consider the exons exclusively.

and HPRT genes does not depend on their notable length, because the ALB and AFP genes, which are much longer than the TK gene, do not contain several sites. The three HK genes are those which contain the largest number of dcm-sites, when compared with the seven TS genes. 5 enzymes only (*Hpa*II, *Hae*III, *Alu*I, *Mbo*I and *Ava*II) cut sequences in all genes. Among them, however, *Hpa*II is the sole enzyme which in its target sequence recognizes a methylatable CpG dinucleotide. On such a basis, one should conclude that the ten genes considered share at least one word: 5mCpG. As a main signal needed to switch-off promoter from transcription [1,5], this word would correspond to the well-documented presence of the dimethylated, in *cis–trans*, 5mCpG/Gp5mC dinucleotide pairs in eukaryotic DNA [2,25].

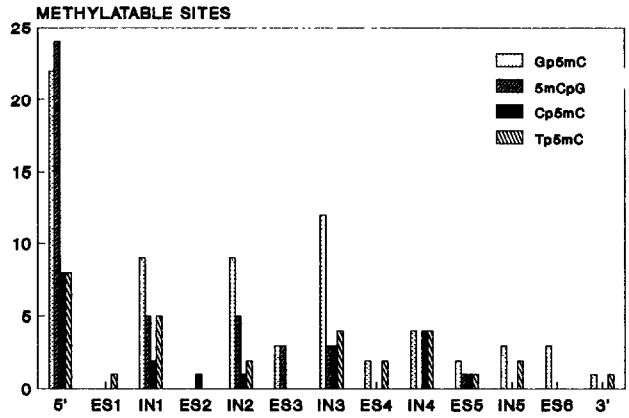

Fig. 2. Differential distribution of four methylatable dinucleotide classes present in dcm-sites along the TS gene for CALC. The dcm-inhibited restriction endonucleases were the following: *Hpa*II, *Hae*III, *Hha*I, *Sma*I, *Apa*I, *Nar*I, *Alu*I, *Bam*HI, *Hin*dIII, *Mbo*I, *Bgl*II, *Ava*I, *Ava*II and *Hae*II (Table II) The ordinate shows the number of dcm-sites containing the 5mC in the following positions: 5mCpG, Gp5mC, Tp5mC and Cp5mC. The abscissa shows the transcriptional unit. 5',3' = Flanking regions containing the promoter (upstream) and the stop codon and the polyadenylation signal (downstream); IN = introns 1, 2, 3, 4 and 5; ES = exons 1, 2, 3, 4, 5 and 6.

3. METHYLATION PROFILES OF HOUSEKEEPING (HK) AND TISSUE-SPECIFIC (TS) GENES

The next goal was to verify whether the average frequency of occurrence of dcm-sites is distributed at random inside the transcriptional units or determined genetically upstream and downstream, with a specific differential occurrence in introns and exons (Fig. 1). It was discovered that this average frequency is higher in HK and lower in TS genes. In both HKs and TSs, it is constantly higher along the uncoding and lower along the coding sequences. Going in the 5'–3' direction, the average frequency of dcm-sites decreases along the TS uncoding and increases along the TS coding sequences (Fig. 1b,d). Lastly, the average frequency of the dcm-sites along the coding and uncoding sequences of HKs (Fig. 1a,c) appears to show some 'crossing' of the two reciprocal tendencies which are well discriminated within the TS genes. In fact, a well-expressed curvilinear pattern appeared showing the minimal value of dcm-sites roughly in the middle of HKs. The curvature is more evident in uncoding (Fig. 1a) and less evident in coding (Fig. 1c) sequences. The crossing might depend on the general tendency of the dcm-site frequency to decrease along the uncoding sequences, from 5' to 3'. Apparently, in the middle of the HK transcriptional units, this tendency might be overlapped by the opposite general tendency of the dcm-site frequency increasing along the exons. Anyway, in HKs, the first tendency might characterize the uncoding sequences, with a more pronounced initial decline and a less pronounced final slope (Fig. 1a), while the second tendency might characterize the exons, with a less pronounced initial decline and a more pronounced final slope (Fig. 1c). Taken as a whole, these results suggested that differentiation of a language in HKs might be less remarkable for coding sequences in the first half of these genes and less remarkable for uncoding sequences in the second half. By contrast, two well-discriminated languages might be correlated with the differentiation of the coding and uncoding parts of the TS genes. This would support the idea that DNA methylation is correlated indeed with differentiation [26,27].

4. VARIABILITY OF THE MODIFIED CODE AS A FUNCTION OF THE GENE ORGANIZATION

An attempt was made to decipher the internal key words of the dcm-sites along all HK and TS transcriptional units. The first observable fact was a general asymmetric distribution of dcm-sites along the genes: there is an extremely high presence of dcm-sites along the 5'-flanking regions and an extremely low presence of dcm-sites along the 3'-flanking regions. This conformed fairly well with the biochemical demonstration that methylation preferentially involves promoter [1].

Table I

Properties of human HK and TS genes.

Their lengths vary from 1,665 to 56,736 bp. In all of them there are different numbers of introns (IN) and exons (EX) also of various lengths (INs oscillate, in number, from 2 to 14 and, in bp, from 82 to 13,057; EXs oscillate, in number, from 3 to 14 and, in bp, from 15 to 720).

Genes	N. nucleotide			IN	EX
	5'-FL	3'-FL	Total		
HK					
ODC	794	648	9,043	11	12
TK	519	742	13,500	6	7
HPRT	1,676	15235	56,736	8	9
TS					
MT-IF	481	492	2,076	2	3
CG	551	33	1,665	2	3
γINF	346	1361	5,961	3	4
CALC	1,833	122	7,637	5	6
APO A-I	213	361	2,385	3	4
ALB	1,775	477	19,002	14	14
AFP	914	1952	22,166	14	14

5'-FL = 5'-Flanking region; 3'-FL = 3'-Flanking region. ODC = ornithine decarboxylase; TK = thymidine kinase, HPRT = hypoxanthine/guanine phosphoribosyl transferase; MT-IF = metallothionein I-F; CG = chorionic gonadotropin; γINF = γ-interferon; CALC = calcitonin; APO A-1 = apolipoprotein A-1; ALB = albumin; AFP = α-fetoprotein These genes are fully sequenced as achievable through computer analysis from the EMBL Nucleotide Sequence Database Release 25.

The second observable fact was a striking intermittence of dcm-enriched and dcm-poor regions along all genes. This also conformed with the demonstration that methylation preferentially occurs on regulatory sequences that do not code for mRNAs, while the coding sequences are 5mC-poor [1,6]. Fig. 2 shows the concrete example of the CALC gene. In Table II it can clearly be seen that the dcm-sites contain 5mC in the following four most probable positions: 5mCpG, Gp5mC, Tp5mC and Cp5mC. These methylated dinucleotides are heterogeneously distributed along the gene (Fig. 2). The 5mCpG and Gp5mC dinucleotides highly dominate in its introns and particularly in its promoter, while there are only few Tp5mCs and Cp5mCs in its exons. In the promoter, the concentration of 5mCpGs overcomes that of Gp5mCs. In introns, this proportion is inverted to favour Gp5mCs, with a maximal value in the center of the gene. This distribution of dinucleotides is rather similar along the transcriptional units of both HKs and TSs, with some quantitative oscillations which apparently seem to store a common language in them.

5. DNA METHYLATION AND GENETIC REGULATION OF TRANSCRIPTION

One should argue, therefore, that the heterogeneous methylated pattern might be correlated with the movement of the RNA polymerase II along the transcriptional unit, for instance, through intermediary 5mC-binding proteins like those described recently [28]. In this case, the high methylation level detected on promoter might signify its involvement in switching-off transcription through such proteins (assuming that these do compete with the transcriptase in promoter recognition). The gradual decrease of the average frequence of dcm-sites along the introns, from 5′ to 3′, might suggest that they probably exert an auxiliary role in modulating the movement of the transcriptase. The

Table II

Presence of dcm-sites in the HK and TS genes.

The cut and uncut sites contain 4–6 letters. The cut sites do not contain 5mC, with the exception of GG/C5mC (cut by *Hae*III), 5mC5mCC/GGG (cut by *Sma*I), GG/CGC5mC (cut by *Nar*I) and G/GATC5mC (cut by *Bam*HI). *Hpa*II, *Hae*III, *Hha*I, *Alu*I and *Mbo*I recognize sites with 4 letters. *Bam*HI and *Ava*II recognize sites with 5 letters. All other enzymes (*Sma*I, *Apa*I, *Nar*I, *Pvu*I, *Hin*dIII, *Bgl*II, *Eco*RI, *Sal*I, *Xba*I, *Xho*I, *Ava*I and *Hae*II) recognize sites with 6 letters. The uncut dcm-sites by these restrictases contain 5 mC in the following positions. 5mCpG (*Hpa*II, *Sma*I, *Nar*I, *Pvu*I, *Hha*I, *Sal*I, *Xho*I and *Ava*I); Gp5mC (*Hae*III, *Apa*I, *Alu*I, *Hin*dIII and *Hae*II); Tp5mC (*Bam*HI, *Pvu*I, *Bgl*II, *Eco*RI, *Sal*I, *Xba*I and *Xho*I), Cp5mC (*Hpa*II, *Sma*I and *Ava*II). There are also some ambiguous codes, such as 5mCpC and 5mCpG (i.e. for *Hpa*II).

Enzymes	Cut sequences	Uncut sequences	HKs			TSs						
			ODC	TK	HPRT	CG	MT-IF	APOA-I	CALC	γIFN	ALB	AFP
*Hpa*II	C/CGG	C5mCGG 5mCCGG	X	X	X	X	X	X	X	X	X	X
*Hae*III	GG/CC GG/C5mC	GG5mCC	X	X	X	X	X	X	X	X	X	X
*Hha*I	GCG/C	G5mCGC GCG5mC	X	X	X	X	X	X	X	0	X	X
*Sma*I	CCC/GGG 5mC5mCC/GGG	CC5mCGGG	X	X	X	X	0	X	X	0	X	0
*Apa*I	GGG/CCC	GGG5mCCC	X	X	X	X	0	X	X	0	0	0
*Nar*I	GG/CGCC GG/CGC5mC	GG/5mCGCC	X	X	X	0	X	X	X	0	X	X
*Alu*I	AG/CT	AG5mCT	X	X	X	X	X	X	X	X	X	X
*Bam*HI	G/GATC G/GATC5mC	GGAT5mCC	X	X	X	0	X	0	X	X	0	X
*Pvu*I	CGAT/CG	CGAT5mCG	X	X	X	0	0	0	0	0	0	0
*Hin*dIII	A/AGCTT	AAG5mCTT	X	X	X	X	0	X	X	X	X	X
*Mbo*I	/GATC	GAT5mC	X	X	X	X	X	X	X	X	X	X
*Bgl*II	A/GATCT	AGAT5mCT	0	X	X	0	X	X	X	0	X	X
*Eco*RI	G/AATTC	GAATT5mC	X	X	X	0	0	0	0	X	X	X
*Sal*I	G/TCGAC	GT5mCGAC	0	0	0	0	0	0	0	0	0	0
*Xba*I	T/CTAGA	T5mCTAGA	X	X	X	0	X	0	0	X	X	X
*Xho*I	C/TCGAG	CT5mCGAG	0	X	X	0	0	X	0	0	X	0
*Ava*I	C/YCGRG	CY5mCGRG	X	X	X	0	X	X	X	0	X	X
*Ava*II	G/GWCC	GGWC5mC	X	X	X	X	X	X	X	X	X	X
*Hae*II	RGCGC/Y	RGCG5mCY	X	X	X	0	X	X	X	0	X	X

'R' represents a purine, 'Y' represents a pyrimidine; 'W' may represent an adenine (A) or a thymine (T). X = Presence of dcm-sites; 0 = Absence of dcm-sites. All sites were collected from Sigma Cel Culture Reagents Catalogue (1990) and from the Boehringer Mannheim Biochemicals for Molecular Biology Catalogues (1990, 1991). Their distribution along the genes was achieved using the Mapdraw DNAstar Program.

countercurrent gradual increase of dcm-sites along the exons might signify an alternative complementary force in modulation of the transcriptase operativity along the gene. The end of transcription might require the absence of dcm-sites with a given code (Gp5mC) on the last introns and vice versa the presence of dcm-sites with another code (Tp5mC or Cp5mC) on the last exons. It seems improbable that the reciprocal patterns of Fig. 1b,d, carrying differential codes, have no significance whatsoever.

Acknowledgements: We are very grateful to Prof. A. Ruffo, of the Lincei National Academy, for stimulating discussion.

REFERENCES

[1] Volpe, P. and Eremenko, T. (1974) FEBS Lett. 44, 121–126.
[2] Geraci, D., Eremenko, T., Cocchiara, R., Granieri, A., Scarano, E. and Volpe, P. (1974) Biochem. Biophys. Res. Commun. 57, 353–361.
[3] Volpe, P. (1976) Horizons Biochem. Biophys. 2, 285–340.
[4] Holliday, R. and Pugh, J.E. (1975) Science 187, 226–232.
[5] Maclean, A. and Hilder, V.A. (1977) Int Rev. Cytol. 48, 54–74.
[6] Tentravahi, V., Guntaka, R.V., Erlanger, B.F. and Miller, O.J. (1981) Proc Natl. Acad. Sci USA 78, 489–493.
[7] Volpe, P., Menna, T. and Eremenko, T. (1976) Bull. Mol. Biol. Med. 1, 18–28.
[8] Eremenko, T. and Volpe, P. (1975) Eur. J. Biochem. 52, 203–210.
[9] Mandel, J.P and Chambon, P. (1979) Nucleic Acids Res. 7, 2081–2103.
[10] McGhee, J.D. and Ginder, G.D. (1979) Nature 280, 419–420.
[11] Sutter, D. and Doerfler, W. (1980) Proc. Natl. Acad. Sci. USA 77, 253–256.
[12] Razin, A. and Riggs, A.D. (1980) Science 210, 604–610.
[13] Cascio O., Petrone G., Fazzio A., Sarpietro M.G., Cambria A. and Volpe P. (1991) Macromol. Funct. Cell 6, 163–178.
[14] Volpe, P., Esposito, C., Iacovacci, P., Butler, R.H. and Eremenko, T. (1993) Macromol. Funct. Cell 7, 59–71.
[15] Volpe, P., Iacovacci, P., Esposito C. and Eremenko, T. (1992) Proc. Natl. Acad. Lincei 3, 383–394.
[16] Holtta, E., Hiervonen, A., Wahlfors, J., Allhonen, L., Janne, J. and Kallio, A. (1989) Gene 83, 125–135.
[17] Kerbel, R.S., Liteplo, R. and Frost, P. (1986) Prog. Clin. Biol. Res. 212, 293–304.
[18] Jones, P.A., Taylor, S.M., Mohandas, T. and Shapiro, L.J. (1982) Proc. Natl. Acad. Sci. USA 79, 1215–1219.
[19] Jahroudi, N., Foster, R., Price-Haughey, J., Beitel, G. and Gedamu, L. (1990) J. Biol. Chem. 265, 6506–6511.
[20] Whitfield, G.K. and Kourides, I.A. (1985) Endocrinology 117, 231–236.
[21] Fukunaga, R., Matsuyama, M., Okamura, H., Nagata, K., Nagata, S. and Sokawa, Y. (1986) Nucleic Acids Res. 14, 4421–4436.
[22] de-Bustros, A., Nelkin, B.D., Silverman, A., Ehrlich, G., Poiesz, B. and Baylin, S.B. (1988) Proc. Natl. Acad. Sci. USA 85, 5693–5697.
[23] Ruiz-Opazo, N. and Zannis, V.I. (1988) J. Biol. Chem. 263, 1739–1744.
[24] Nahon, J.L. (1987) Biochemie 69, 445–459.
[25] Bird, A.P. (1978) J. Mol. Biol. 118, 49–60.
[26] Scarano, E. (1971) Adv. Cytopharmacol. 1, 13–21.
[27] Vanyushin, B.F., Mazin, A.L., Vasiliev, V.K. and Belozersky, A.N. (1973) Biochim. Biophys. Acta 229, 397–404.
[28] Meehan, R., Antequera, F., Lewis, J., McLeod, D., McKay, S., Kleiner E. and Bird, A.P. (1990) Phil. Trans. R. Soc. London B326, 199–205.