Distribution Patterns for 5-Methylcytosine among Apurinic DNAs from Several Sources

David P. Ringer,* Boyd A. Howell, David C. Beck, Joseph A. Clouse, and Donald E. Kizer
Biomedical Division, The Samuel Roberts Noble Foundation, Inc., Ardmore, Oklahoma 73402
Received February 5, 1985

ABSTRACT: Purified DNA from the liver of rats, mice, rabbits, and guinea pigs, from guinea pig lymph nodes, from hyperplastic nodules induced in rat liver by feeding with 2-(acetylamino)fluorene, and from *Escherichia coli* cells was made apurinic by reaction with diphenylamine. After chromatographic separation of pyrimidine tracts (isostichs or isoplyths) according to the number of contiguous pyrimidines, semilog plots of tract frequency vs. the number of contiguous pyrimidines were linear, plots for DNA from several sources differed from one another, and all deviated significantly from randomness. Similar semilog plots for coding sequences among 60 mammalian genomes or 28 rat tissue genomes were intermediate among slopes for isolated DNA. Individual isostichs were hydrolyzed, and their constituent pyrimidine bases were analyzed by high-pressure liquid chromatography. Among isostichs from isolated DNAs, the distribution of Thy and Cyt contents differed markedly from the distribution of 5-methylcytosine (5-Me-Cyt); e.g., although isostich 1 contained 45–49% of 5-Me-Cyt, amounts of Thy or Cyt did not exceed 25%. Semilog plots of normalized values for tract frequency or the content of 5-Me-Cyt vs. isostich number were essentially superimposable; thus, among the first five pyrimidine tracts of a particular tissue or *E. coli* DNA, the number of tracts per 5-Me-Cyt moiety was essentially constant. The data showed that 5-Me-Cyt and/or dCyd-dGuo dinucleotides have a distribution throughout DNA structure that superimposes the distribution of pyrimidine tract frequency and suggests that regulatory 5-Me-Cyt moieties are principally located at 3′ termini of pyrimidine tracts.

The only modified base occurring naturally in vertebrate DNA is 5-methylcytosine (5-Me-Cyt)[1] (Ehrlich & Wang, 1981). Most 5-Me-Cyt moieties occur in dCyd-dGuo sequences (Razin & Riggs, 1980; Ehrlich & Wang, 1981; Doerfler, 1983), but only 50–70% of the dCyd-dGuo dinucleotides are methylated (Razin & Friedman, 1981; Doerfler, 1983). Methylation of Cyt moieties in eukaryotic DNA is a key element among mechanisms that control accessibility and expression of genes (Razin & Riggs, 1980; Ehrlich & Wang, 1981; Razin & Friedman, 1981; Boehm & Drahovsky, 1983; Doerfler, 1983; Jaenisch & Jahner, 1984). Although hypermethylation is correlated with inaccessibility of genes and hypomethylation with gene accessibility, some 5-Me-Cyt residues appear to exert no appreciable influence on gene regulation (Razin & Riggs, 1980; Ehrlich & Wang, 1981; Razin & Friedman, 1981; Boehm & Drahovsky, 1983; Doerfler, 1983; Jaenisch & Jahner, 1984). Doerfler (1983) proposed that DNA methylation be viewed as a pluripotent signal that assumes different meanings depending on sites and sequences involved and suggested that future research should focus on patterns of methylation at all dCyd-dGuo sites in genes and their regulatory sequences.

Information concerning 5-Me-Cyt content and/or distribution has come from direct analyses on whole or fractionated DNAs or has been inferred from data on the sensitivity of these DNAs to certain restriction endonucleases; however, restriction enzyme analyses may only probe 10% of methylated sites (Razin & Riggs, 1980). Sequence determinations, as currently used, do not fix the position of 5-Me-Cyt in DNA, but its position can be inferred by dCyd-dGuo nearest-neighbor determinations. Owing to the limitations of these approaches for accessing possible methylation patterns imposed on DNA, we have searched for methylation patterns in apurinic DNA. The development of sensitive methods for direct determination of 5-Me-Cyt (Kuo et al., 1980), together with the knowledge that most 5-Me-Cyt occur in dCyd-dGuo sequences (Razin & Riggs, 1980; Ehrlich & Wang, 1981; Doerfler, 1983), suggested that apurinic DNA might faithfully reflect a methylation pattern, should a pattern exist. Furthermore, availability of a number of sequences for mammalian DNAs also permitted computer-assisted determination of dCyd-dGuo dinucleotide distributions among pyrimidine tracts in coding sequences for comparison with actual 5-Me-Cyt distribution patterns.

Materials and Methods

DNA Sources. Livers were dissected from male Sprague-Dawley rats (SASCO Inc., Omaha, NE), male Hartley albino guinea pigs (BioLab, St. Paul, MN), male C3HeB/FeJ mice (Jackson Laboratories, Bar Harbor, ME), and male New Zealand white rabbits (purchased locally). Hyperplastic nodules were induced in rat liver with 2-(acetylamino)fluorene (AAF) by the procedure of Teebor & Becker (1971), except that AAF was fed at 0.05% and rats were exposed to five cycles of feeding. Peritoneal lymph nodes were dissected from guinea pigs 14 days following injection with complete Freund's adjuvant. Lyophilized strain B *Escherichia coli* cells (E.C. 11303) were purchased from Sigma Chemical Co., St. Louis, MO. Rat liver DNA was labeled by partially hepatectomizing rats (Higgins & Anderson, 1931) and, 22-h postoperation, injecting each with 25 μCi of [*methyl*-^{3}H]dThd and allowing 2 h of incorporation.

Isolation of DNA and Pyrimidine Tracts. Tissues were homogenized, nuclear pellets were isolated by centrifugation, and DNA was isolated as previously described (Kizer et al., 1983). DNA was depurinated with diphenylamine in formic acid, and the liberated pyrimidine tracts were separated according to the number of contiguous pyrimidine moieties per

[1] Abbreviations: 5-Me-Cyt, 5-methylcytosine; AAF, 2-(acetylamino)fluorene.

0006-2960/85/0424-7498$01.50/0

tract by elution from DEAE-cellulose columns (Cerny et al., 1969; Spencer et al., 1969; Lieberman et al., 1971).

Isolation of Pyrimidine Bases. Isostichs were desalted by passage through Sephadex G-10 columns (Pharmacia Fine Chemicals, Piscataway, NJ) and then hydrolyzed in 88% formic acid (Carrier & Setlow, 1971; Williams, 1978). Pyrimidines were separated by high-pressure liquid chromatography (HPLC) on a C_{18} column (Kuo et al., 1980), and amounts of each were quantified by comparing their absorbance at 280 nm to that of knowns. Known amounts of radioactively labeled rat liver DNA were added to unlabeled DNA sources; radioactivity was monitored (Bruno & Christian, 1961; Bush, 1963) throughout analytical manipulations to give a basis for estimating product recoveries.

Computations. The nucleotide content (*Cn*) in tracts from isolated DNA is the sum of Thy, Cyt, and 5-Me-Cyt contents in tract hydrolysates as determined by HPLC analysis (Kuo et al., 1980). The frequency (*Fn*) of a pyrimidine tract is the ratio Cn/n, where n is the number of contiguous pyrimidines in the tract.

Pyrimidine tracts distributed among DNAs isolated from natural sources were compared to average tract distributions among 60 coding sequences from various mammalian tissues or among 28 coding sequences from several rat tissues (for references, see footnotes for Table I). Coding sequences from each source were entered into a Pertec PCC 2000 computer, and programs were developed to analyze pyrimidine tract and nucleotide distributions.

A program for generating random numbers in the TRS-80C computer was used to generate random-number DNA molecules and compute average values for tract frequency and Thy or Cyt contents; it was clear that these averages were valid estimates of F_n and C_n as theoretically derived earlier by Fitch (1977). Direct comparison of pyrimidine tract data arising from HPLC analyses or computer analyses was facilitated by normalizing, e.g., expressing a value for tract 1 as a percent of the sum of similar values for tracts 1–11.

Results

Thy over Cyt Ratios. Thy over Cyt ratios for pyrimidine tracts from isolated DNA were determined and compared when possible with published data to ensure that procedures of depurination, tract isolation, tract hydrolysis, and base separation had not biased our estimates of Thy, Cyt, and 5-Me-Cyt contents. In our hands, values ranged from 0.92 for *E. coli* to 1.43 for mouse liver. The values for mouse and rat liver DNAs, i.e., 1.43 ± 0.01 and 1.27 ± 0.04, respectively, differed slightly from those reported by Shapiro (1976), i.e., 1.51 ± 0.04 and 1.39 ± 0.03, respectively; however, the differences may have resulted from the combining of 5-Me-Cyt and Cyt contents in our computations, whereas Shapiro (1976) listed no 5-Me-Cyt values for these DNAs. We noted with interest that Thy/Cyt values for DNA from hyperplastic DNA were appreciably higher than those for regenerating rat liver, i.e., 1.4 ± 0.06 vs. 1.27 ± 0.04, respectively.

Frequency of Pyrimidine Tracts. The distribution of pyrimidine tracts among various isolated DNAs, random-number DNA, and coding sequence regions from 60 mammalian genes (coding sequences) and 28 rat genes [coding sequences (rat)] are summarized in Table I. The data are reported as slopes of semilog plots of pyrimidine tract frequency vs. pyrimidine tract length; all plots were linear as observed earlier with tracts from bacterial DNA (Shapiro et al., 1965). Slope estimates for all isolated DNAs and coding-sequence DNAs were significantly lower ($P < 0.05$) than slope estimates for random-number DNA, indicating a departure from random distribution occasioned by an underrepresentation among shorter pyrimidine tracts and an overrepresentation among longer tracts. Similar differences in distribution were reported earlier (Spencer & Chargaff, 1963; Sneider & Stone, 1971). Furthermore, a comparison of slope values for the distribution of

Table I: Comparison of Slope Values for the Distribution of Pyrimidine Tract Frequency[a]

DNA source[b]	correlation coefficient	slope ± SE[c]
random number	−0.9999 (12)[d]	−0.301 ± 0.001
E. coli	−0.998 (3)	−0.17 ± 0.003[e]
rat liver	−0.9992 (8)	−0.25 ± 0.001[e]
rabbit liver	−0.996 (3)	−0.25 ± 0.003[e]
coding sequences[f]	−0.998 (60)	−0.24 ± 0.002[e]
coding sequences (rat)[g]	−0.997 (28)	−0.24 ± 0.003[e]
hyperplastic nodule	−0.998 (3)	−0.24 ± 0.002[e]
guinea pig liver	−0.996 (3)	−0.23 ± 0.003[e]
mouse liver	−0.997 (3)	−0.23 ± 0.003[e]
lymph nodes	−0.997 (3)	−0.22 ± 0.002[e]

[a] Slopes were computed from the regression of log of percent values vs. number of contiguous pyrimidines for isostichs 1–8. [b] DNA isolation and pyrimidine tract analysis were performed as described under Materials and Methods. [c] Variance about slope estimates, S_{bl}^2, was calculated as described by Natrella (1963). Standard error was S_{bl}^2/n, where n was 8 in all cases. [d] Numbers in parentheses are number of experiments or sequences. [e] Values less than ($P < 0.05$) the value for random-number DNA. [f] References for 60 coding sequences published in recent literature or in *Nucleotide Sequences* (1984, IRL Press, Oxford, England). Below is an alphabetical listing of authors, followed by the journal name or the *Nucleotide Sequences* (1984) entry name. For this listing, journal names were abbreviated in the following manner: NAS, *Proc. Natl. Acad. Sci. U.S.A.*; NAR, *Nucleic Acids Res.*; Nat, *Nature (London)*; EJB, *Eur. J. Biochem.*; FEBS, *FEBS Lett.*; BBRC, *Biochem. Biophys. Res. Commun.*; JBC, *J. Biol. Chem.*; JMB, *J. Mol. Biol.*; Sci, *Science (Washington, D.C.)*; BJ, *Biochem. J.* References for the 60 coding sequences are as follows: Blackburn et al., RATCASB; Claesson et al., NAS, *80*, 7395, 1983; Cooper & Crain, NAR, *10*, 4081, 1982; Deschenes et al., NAS, *81*, 726, 1984; Diamong et al., Sci, *225*, 516, 1984; Dickson et al., JBC, *256*, 8407, 1981; Dodgson et al., JBC, *258*, 12685, 1983; Ellison et al., HUMIGG1C; Fugii-Kuriyama et al., RATCYP450; Hall et al., EJB, *138*, 585, 1984; Harris et al., BOVCHYMOB; Heilig et al., NAR, *10*, 4363, 1982; Hennighausen et al., MUSECA; Holbrook et al., NAS, *81*, 1634, 1984; Horikawa et al., Nat, *306*, 611, 1983; Horwich et al., Sci, *224*, 1068, 1984; Imai et al., NAS, *80*, 7405, 1983; Jakowlew et al., NAR, *12*, 2861, 1984; Jansen et al., Nat, *306*, 609, 1983; Kan et al., NAS, *81*, 3000, 1984; Kimura et al., NAR, *12*, 2917, 1984; Knott et al., BBRC, *120*, 734, 1984; Kost et al., NAR, *11*, 8287, 1983; Lai et al., JBC, *259*, 5536, 1984; Leibold et al., JBC, *259*, 4327, 1984; Lemischka et al., RATTUBAL1; Le Moullec et al., FEBS, *167*, 93, 1984; Lin et al., Sci, *224*, 843, 1984; Maki et al., Nat, *309*, 722, 1984; Malissen et al., HUMHLA; Montminy et al., NAS, *81*, 3337, 1984; Nickerson & Piatigorsky, NAS, *81*, 2611, 1984; Noda et al., Nat, *305*, 818, 1983; Orkin et al., JBC, *258*, 12753, 1983; Phillips et al., JBC, *259*, 7947, 1984; Pinsky et al., NAS, *80*, 7486, 1983; Poncin et al., EJB, *140*, 493, 1984; Quax-Jeuken et al., NAS, *80*, 3548, 1983; Raugei et al., NAR, *11*, 5811, 1983; Rosen et al., NAR, *12*, 4893, 1984; Sandell et., JBC, *259*, 7826, 1984; Sharpe et al., NAR, *12*, 3917, 1984; Sherman et al., HUMSOD1; Stewart et al., NAR, *12*, 3895, 1984; Takahashi et al., NAR, *11*, 6847, 1983; Taniguchi et al., HUMIL2; Taylor et al., BJ, *219*, 223, 1984; Wiginton et al., NAR, *12*, 2439, 1984; Wodnar-Filipowicz et al., NAS, *81*, 2295, 1984; Yabusaki et al., NAR, *12*, 2929, 1984; Zakut-Houri et al., Nat, *306*, 594, 1983. [g] References for rat nucleotide sequences as cited in *Nucleotide Sequences* (1984, Part I): Amara et al., RATCAL; Amara et al., RATCALP; Amara et al., RATCGRP; Barta et al., RATGH1; Blackburn et al., RATCASB; Burnstein et al., RATIGKJ; Dandekar & Qasba, RATALPHALA; Gordon et al., RATFABP; Innis & Miller, RATAFP; Jacobs et al., RATCALCITN; Jagodzinski et al., RATAFPM; Liao et al., RATAFPA; MacDonald et al., RATAMYLASE; MacDonald et al., RATELAI; MacDonald et al., RATELAII; Moorman et al., RATCRYALPH; Moorman et al., RATCRYGAMA; Ohkubo et al., RATANG; Page et al., RATGH2; Quinto et al., RATCBXPA; Ricca et al., RATAGPA1; Sargent et al., RATALBM; Schmale et al., RATAVP1; Schmale et al., RATAVP2; Shani et al., RATACTA; Tsutsumi et al., RATALDB; Ullrich et al., RATINSI; Unterman et al., RATGBA2UM; Yu-Lee & Rosen, RATCASG11; Zakut et al., RATACTSK.

Table II: Pyrimidine Contents among Isostichs 1–5 for DNAs from Various Sources

DNA source	percent pyrimidine content[a]														
	Thy isostich					Cyt isostich					5-Me-Cyt isostich				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
random number	25	25	19	12	8	25	25	19	12	8	50[b]	25[b]	13[b]	6[b]	3[b]
E. coli	16	25	19	13	10	19	30	19	12	8	45	21	16	7	4
rat liver	22	20	17	13	10	19	20	19	14	10	44	21	14	9	5
rabbit liver	25	19	16	12	8	21	20	16	13	9	50	20	12	8	4
coding sequences	23	21	17	14	10	20	24	18	14	9	48[b]	26[b]	16[b]	11[b]	6[b]
coding sequences (rat)	23	23	17	13	10	20	26	19	12	11	48[b]	27[b]	18[b]	12[b]	8[b]
hyperplastic nodules	21	18	17	14	10	20	19	17	14	10	48	23	14	8	4
guinea pig liver	20	20	18	12	8	18	20	18	13	9	42	23	14	10	4
mouse liver	21	17	18	13	9	18	19	19	14	9	45	20	13	9	1
lymph nodes	14	22	18	15	10	19	30	19	12	8	45	21	16	7	4

[a] Values are averages. The number of experiments was listed in Table I. [b] Values listed are values for the dinucleotide dCyd-dGuo not for 5-Me-Cyt.

tracts among whole DNAs isolated from mammalian and bacterial sources vs. those from coding sequences indicated values were in the same range. Thus, factors in addition to exon enrichment exerted influence upon the extent to which tract frequencies deviated from randomness.

Distribution of Pyrimidine Bases or dCyd-dGuo Dinucleotides among Tracts. Table II compares the distribution of the three pyrimidine constituents of isolated DNA. Since upward of 90% of 5-Me-Cyt moieties have a 3′-adjacent Gua moiety (Razin & Riggs, 1980; Ehrlich & Wang, 1981; Doerfler, 1983; Cooney et al., 1984; Pollack et al., 1984), we also compared the 5-Me-Cyt distribution among tracts from isolated DNAs with the distribution of dCyd-dGuo dinucleotides in tracts from random-number DNA and coding sequences. For clarity and ease of presentation, data are shown only for isostichs 1–5; among the DNAs, these five isostichs contained 81–89% of total Thy or Cyt moieties and 93% or more of 5-Me-Cyt moieties or dCyd-dGuo dinucleotides. Among all DNAs, the distribution of dCyd-dGuo doublets and 5-Me-Cyt differed markedly from distributions for Thy or Cyt; e.g., isostich 1 contained 42–50% dCyd-dGuo doublet or 5-Me-Cyt, whereas amounts of Thy or Cyt did not exceed 25%. Furthermore, as the number of contiguous pyrimidines increased, amounts of dCyd-dGuo doublets or 5-Me-Cyt moieties diminished twice as fast as amounts of Thy or Cyt moieties; e.g., slopes of plots of log of percent pyrimidine content vs. pyrimidine tract length for Thy, Cyt, and 5-Me-Cyt distributions among rat liver DNA isostich were −0.0889, −0.0688, and −0.2286, respectively. Plots of dCyd-dGuo or 5-Me-Cyt distribution among pyrimidine tracts, i.e., plots of log of percent of dCyd-dGuo or 5-Me-Cyt vs. pyrimidine tract length, were linear with correlation coefficients of 0.991 or better and had slopes of −0.22 for dCyd-dGuo distribution among the 60 coding sequences and −0.24 ± 0.01 for 5-Me-Cyt distribution among the isolated DNAs. The close similarity between these slopes supports confinement of five methylations to Cyt of the dCyd-dGuo dinucleotide as suggested in earlier reports using different methods, e.g., Doerfler (1983). In addition, since dCyd-dGuo dinucleotides in coding sequences resided, ipso facto, at 3′ termini of pyrimidine tracts, it follows that 5-Me-Cyt moieties also occur principally at 3′ termini of tracts.

Nearest-Neighbor Analyses of dCyd-dGuo Dinucleotides among Coding Sequences. 5-Me-Cyt moieties and dCyd-dGuo dinucleotides are highly conserved among eukaryotic DNA (Vanyushin et al., 1970, 1973; Russell et al., 1976; Shapiro, 1976; Brown & Burdon, 1977; Kuo et al., 1980; Ehrlich & Wang, 1981; Pollack et al., 1984; Tykocinski & Max, 1984), raising the possibility that similarities in distribution were owing to low levels of both in DNA. To investigate whether dCyd-dGuo dinucleotides in the coding sequences were uniquely conserved among pyrimidine tracts, its distribution was compared with other dinucleotide distributions. A comparison of dCyd-dGuo occurrence with respect to (a) the calculated observed/expected ratio (Tykocinski & Max, 1984) for the 16 nearest neighbors of coding sequences among 60 mammalian tissues, Table III, or (b) the distribution of other pyrimidine–purine dinucleotides at 3′ termini of pyrimidine tracts from coding sequences, Figure 1, gave similar patterns of distribution. In both comparisons, dCyd-dGuo as well as dThd-dAdo dinucleotides were distributed in patterns suggesting their occurrence was conserved among coding sequences. The observed/expected ratios in Table III for dThd-dAdo and dCyd-dGuo, viz., 0.50 ± 0.029 and 0.45 ± 0.033, respectively, indicated that they occurred only half as often as expected, a pattern previously observed (Smith et al., 1983). The lower slopes for the semilog plots of dCyd-dGuo and dThd-dAdo distribution through the first five pyrimidine tracts shown in Figure 1 might also reflect the known conservation of dCyd-dGuo in vertebrate DNA. Conservation

Table III: Observed over Expected Frequency Ratios for Nearest Neighbor among Coding Sequences[a]

	A	T	C	G
A	1.01 ± 0.030	0.86 ± 0.025[b]	0.89 ± 0.020[b]	1.21 ± 0.022[b]
T	0.50 ± 0.029[b]	0.97 ± 0.026	1.00 ± 0.023	1.46 ± 0.027[b]
C	1.17 ± 0.027[b]	1.34 ± 0.036[b]	1.09 ± 0.030[b]	0.45 ± 0.033[b]
G	1.21 ± 0.021[b]	0.82 ± 0.025[b]	0.97 ± 0.025	0.99 ± 0.021

[a] Following computer enumeration of dinucleotide content in coding sequences, values were computed essentially as described by Tykocinski & Max (1984). [b] Values that differ significantly (at $P = 0.05$ or greater) from 1.

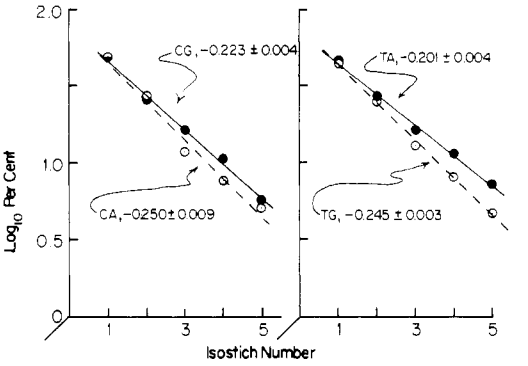

FIGURE 1: Semilog plots of the distribution of dinucleotides at 3′ termini among pyrimidine tracts from coding sequences. Values are normalized averages for 60 published coding sequences; for references, see the footnotes for Table I. Isostichs 1–5 accounted for not less than 96% of the dinucleotides enumerated. CG and CA, dCyd-dGuo and dCyd-dAdo dinucleotides; TG and TA, dThd-dGuo and dThd-dAdo dinucleotides.

Table IV: Number of Pyrimidine Tracts for Each 5-Methylcytosine Moiety among Isolated DNAs or for Each Dinucleotide at 3′ Termini of Tracts among Coding-Sequence DNAs[a]

DNA source	pyrimidine tracts per 5-methylcytosine					
	1–5	1	2	3	4	5
E. coli	100 ± 9	91 ± 30	133 ± 25	86 ± 22	98 ± 8	98 ± 18
rat liver	26 ± 1	29 ± 1	28 ± 1	25	23 ± 1	24 ± 1
mouse liver	25 ± 1	26 ± 1	26 ± 2	26 ± 1	21 ± 1	26 ± 1
hyperplastic nodule	25 ± 1	26 ± 3	24 ± 1	25 ± 1	25 ± 1	27 ± 1
rabbit liver	23 ± 2	24 ± 2	28 ± 6	24 ± 3	19 ± 4	21 ± 6
guinea pig liver	22 ± 2	25 ± 4	23 ± 5	23 ± 4	19 ± 2	21 ± 3
lymph nodes	19 ± 1	20 ± 1	20 ± 2	21 ± 1	15 ± 1	21 ± 2
	pyrimidine tracts per dinucleotide					
coding sequences	1–5	1	2	3	4	5
CG-	9 ± 0.6	12 ± 2	10 ± 1	7 ± 1	7 ± 0.6	7 ± 0.8
TA-	9 ± 0.4	11 ± 0.8	9 ± 0.7	10 ± 1	7 ± 0.8	6 ± 0.9
TG-	3 ± 0.1	3 ± 0.1	3 ± 0.1	3 ± 0.2	3 ± 0.2	3 ± 0.2
CA-	3 ± 0.1	3 ± 0.1	3 ± 0.1	3 ± 0.1	4 ± 0.4	3 ± 0.2

[a] For DNAs isolated from tissues or cells, number of tracts was nanomoles of Thy + Cyt + 5-Me-Cyt per number of contiguous pyrimidines; among coding sequences, Thy, Cyt, and dCyd-dGuo content was determined by computer enumeration.

of dCyd-dGuo has been postulated to be associated with deamination of methylcytosine moieties in those doublets to form dThd-dGuo or its complimentary dinucleotide, dCyd-dAdo (Bird, 1980; Adams & Eason, 1984; Tykocinski & Max, 1984). Our ratios were >1 for dCyd-dAdo and dThd-dGuo in Table III and supported the postulate. On the other hand, this rationale does not account for the marked conservation of dThd-dAdo dinucleotides and several other doublets with ratios that differed significantly from 1, see Table III. These data indicate that, in coding sequences, the conservative distribution of dCyd-dGuo either at the 3′ termini of pyrimidine tracts or as the observed/expected nearest-neighbor ratio is not unique and suggests that factors in addition to codon requirements and deamination of 5-Me-Cyt may exert influence upon the occurrence of certain dinucleotides.

Comparison of Tract Frequencies with 5-Me-Cyt and Dinucleotide Contents. As shown in Figure 2, semilog plots of normalized values for pyrimidine tract frequency and for 5-Me-Cyt or dCyd-dGuo contents were essentially superimposable. This suggested that the frequency of pyrimidine tract and the frequency of 5-Me-Cyt moieties and/or dCyd-dGuo dinucleotides were interrelated and would possibly yield ratios approaching constant values throughout DNA structures that give rise to the first five pyrimidine isostichs.

Table IV lists ratios for pyrimidine tract frequency per 5-Me-Cyt content among isolated DNAs. In *E. coli* DNA, for tracts containing one through five contiguous pyrimidines, a 5-Me-Cyt moiety occurred once among 100 ± 9 pyrimidine tracts; however, among mammalian DNAs, 5-Me-Cyt moieties occurred with a 4–5-fold higher frequency. In the lower portion of Table IV, ratios are listed for tract frequency per nearest-neighbor dinucleotide. Under an assumption that about half of the dCyd-dGuo doublets in DNA are methylated (Razin & Friedman, 1981; Doerfler, 1983), the dCyd-dGuo values shown for tracts 1–5 suggest that coding sequences may have had about one 5-Me-Cyt moiety for 18 pyrimidine tracts. The ratio for dThd-dAdo doublet was identical with that for dCyd-dGuo, but both dThd-dGuo and dCyd-dAdo doublets occurred 3-fold more frequently among pyrimidine tracts. Furthermore, pyrimidine tracts/dCyd-dGuo ratios decreased as the number of contiguous pyrimidines increased; thus, ratios for isostichs 1, 3, 4, and 5 differed significantly from the average ratio for tracts 1–5. A somewhat similar pattern also was seen for the distribution of dThd-dAdo. In contrast, however, ratios for dThd-dGuo and dCyd-dAdo were essentially constant.

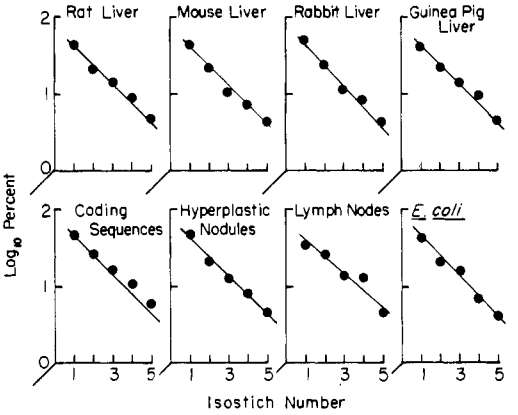

FIGURE 2: Semilog plots of pyrimidine tract frequency (—) and 5-Me-Cyt or dCyd-dGuo contents (●) vs. isostich number for tracts of DNA from various sources. The line represents the regression of average values for isostichs 1–5, while the closed circles depict average values for each isostich; *n* was as listed in Table I. Closed circles in the panel labeled coding sequences represent values for dCyd-dGuo content among 60 published coding sequences.

DISCUSSION

Purines can be chemically stripped from the DNA to yield an apurinic acid in which the pyrimidine deoxynucleotides remain aligned as in the parent nucleic acid (Shapiro & Chargaff, 1960). This allows preservation of 5-Me-Cyt in pyrimidine tracts and provides an opportunity to compare the distribution of 5-Me-Cyt throughout DNA with the manner in which purines are inserted throughout DNA. In polymers resulting from random insertion of the four deoxynucleotides, Fitch (1977) showed that pyrimidine tracts (and, vis-à-vis, purine tracts) occurred with a frequency (*Fn*) described mathematically by the expression $Fn = 0.5^n$ (where n is the number of contiguous pyrimidines) and that the Thy and Cyt content (*Cn*) of these tracts was estimated by the expression $Cn = (0.25n)(0.5^n)$. Our analysis of DNA from bacteria, tissues, and 60 known coding sequences showed tract frequency and Thy or Cyt content distributions that were reasonably estimated by the theoretical expressions of Fitch (1977). However, a similar analysis of 5-Me-Cyt content showed a pattern of distribution that deviated from that for Thy and Cyt and was best represented by the mathematical expression that estimates pyrimidine tract frequency, viz., $Fn = 0.5^n$. It appears that 5-Me-Cyt moieties are inserted into the alignment of pyrimidine deoxynucleotides with a regularity that equals that for insertion of purines. The unique distribution for

5-Me-Cyt was predictable upon considering that upward of 90% of 5-Me-Cyt are 5′ to a Gua moiety (Razin & Riggs, 1980; Ehrlich & Wang, 1981; Doerfler, 1983; Cooney et al., 1984; Pollack et al., 1984). The close correlation between 5-Me-Cyt and dCyd-dGuo dinucleotides was further supported by our analysis of dCyd-dGuo distribution among 60 known coding sequences. It showed that the dCyd-dGuo content was also distributed according to the expression $Fn = 0.5^n$. An observation not predictable nor previously recorded was the remarkable constancy of 5-Me-Cyt or dinucleotide insertions among the pyrimidine tracts of a particular DNA, Table IV.

Several observations from the analysis of 5-Me-Cyt and dCyd-dGuo distributions among pyrimidine tracts suggest that they may be associated with regulation of gene expression. For instance, the observation that 5-Me-Cyt was inserted 4–5-fold more often among pyrimidine tracts from mammalian DNA than among tracts from *E. coli* DNA (Table IV) may be rationalized by the further observation that the bacterial chromosome only accomodates about 2900 genes (von Hippel et al., 1984). Insertion frequencies for dinucleotides at 3′ termini among pyrimidine tracts in coding sequences fell into two distinct groups: (a) dCyd-dGuo and dThd-dAdo dinucleotides were inserted once per nine tracts and were inserted with greater frequency as tract length increased, whereas (b) dThd-dGuo and dCyd-dAdo were inserted 3-fold more often, viz., once per three tracts, and were inserted uniformly at three tracts per dinucleotide. Current knowledge suggests that dCyd-dAdo and dThd-dGuo dinucleotides serve only a codon function, whereas dCyd-dGuo and dThd-dAdo each serve an additional function, viz., methylation sites (Razin & Riggs, 1980; Ehrlich & Wang, 1981; Doerfler, 1983; Cooney et al., 1984; Pollack et al., 1984) and signals for polypeptide chain termination (Lehninger, 1970), respectively.

Doerfler (1983) accepted the causal link between DNA methylation and the regulation of gene expression but observed that we urgently need to understand rules for the positioning of 5-Me-Cyt relative to other regulatory signals. Several observations on nearest neighbors among coding sequences, Table IV, support the idea that regulatory 5-Me-Cyt may be located at 3′ termini of certain pyrimidine tracts. For instance, dCyd-dGuo and dThd-dAdo dinucleotides show similar average insertion frequencies, Table IV, but dThd-dAdo as part of the polypeptide termination signal (Lehninger, 1970) only can function at 3′ termini of tracts. If half the values in Table IV were assumed to represent methylated dCyd-dGuo moieties (Razin & Friedman, 1981; Doerfler, 1983), 5-Me-Cyt frequencies among coding sequences would approach values actually determined for mammalian DNA, viz., 19–26 pyrimidine tracts per 5-Me-Cyt.

Clearly, the role of 5-Me-Cyt in regulatory mechanisms controlling gene expression remains to be elucidated. Our data showing the remarkable constancy for placement of 5-Me-Cyt at 3′ termini of pyrimidine tracts from several DNAs suggest a role for this placement in DNA regulation.

Registry No. dC-dG, 15178-66-2; dT-dA, 19192-40-6; dC-dA, 4624-07-1; dT-dG, 4251-20-1; 5-Me-Cyt, 554-01-8; thymine, 65-71-4; cytosine, 71-30-7.

References

Adams, R. L. P., & Eason, R (1984) *Nucleic Acids Res. 12*, 5869–5877.

Bird, A. P. (1980) *Nucleic Acids Res. 8*, 1499–1504.

Boehm, T. L. J., & Drahovsky, D. (1983) *J. Natl. Cancer Inst. (U.S.) 71*, 429–433.

Browne, M. J., & Burdon, R. H. (1977) *Nucleic Acids Res. 4*, 1025–1037.

Bruno, G. A., & Christian, J. E. (1961) *Anal. Chem. 33*, 1216–1218.

Bush, E. T. (1963) *Anal. Chem. 35*, 1024–1029.

Carrier, W. L., & Setlow, R. B. (1971) *Methods Enzymol. 21*, 230–237.

Cerny, R., Cerna, E., & Spencer, J. H. (1969) *J. Mol. Biol. 46*, 145–156.

Cooney, C. A., Matthews, H. R., & Bradbury, E. M. (1984) *Nucleic Acids Res. 12*, 1501–1515.

Doerfler, W. (1983) *Annu. Rev. Biochem. 52*, 93–124.

Ehrlich, M., & Wang, R. Y.-H. (1981) *Science (Washington, D.C.) 212*, 1350–1357.

Fitch, W. M. (1977) *J. Mol. Biol. 109*, 151–171.

Higgins, G. M., & Anderson, R. M. (1931) *Arch. Pathol. 12*, 186–202.

Jaenisch, R., & Jahner, D. (1984) *Biochim. Biophys. Acta 782*, 1–9.

Kizer, D. E., Ringer, D. P., & Howell, B. A. (1983) *Biochim. Biophys. Acta 740*, 402–409.

Kuo, K. C., McCune, R. A., Gehrke, C. W., Midgett, R., & Ehrlich, M. (1980) *Nucleic Acids Res. 8*, 4763–4776.

Lehninger, A. L. (1970) *Biochemistry*, pp 713–728, Worth, New York.

Lieberman, M. W., Rutman, J. Z., & Farber, E. (1971) *Biochim. Biophys. Acta 247*, 497–501.

Natrella, M. G. (1963) *Experimental Statistics, National Bureau of Standards Handbook 91*, U.S. Government Printing Office, Washington, DC.

Pollack, Y., Kasir, J., Shemer, R., Metzger, S., & Szyf, M. (1984) *Nucleic Acids Res. 12*, 4811–4824.

Razin, A., & Riggs, A. D. (1980) *Science (Washington, D.C.) 210*, 604–610.

Razin, A., & Friedman, J. (1981) *Prog. Nucleic Acid Res. Mol. Biol. 25*, 33–52.

Razin, A., Sedat, J. W., & Sinsheimer, R. L. (1970) *J. Mol. Biol. 53*, 251–259.

Russell, G. J., Walker, P. M. B., Elton, R. A., & Subak-Sharpe, J. H. (1976) *J. Mol. Biol. 108*, 1–23.

Shapiro, H. S. (1976) in *Handbook of Biochemistry and Molecular Biology* (Fasman, G. D., Ed.) Vol. 2, pp 241–311, CRC Press, Cleveland, OH.

Shapiro, H. S., & Chargaff, E. (1960) *Biochim. Biophys. Acta 39*, 68–82.

Shapiro, H. S., Rudner, R., Miura, K.-I., & Chargaff, E. (1965) *Nature (London) 205*, 1068–1070.

Smith, T. F., Waterman, M. S., & Sadler, J. R. (1983) *Nucleic Acids Res. 11*, 2205–2220.

Sneider, T. W., & Stone, K. (1971) *J. Biol. Chem. 246*, 4774–4783.

Spencer, J. H., & Chargaff, E. (1963) *Biochim. Biophys. Acta 68*, 18–27.

Spencer, J. H., Cape, R. E., Marks, A., & Mushynski, W. E. (1969) *Can. J. Biochem. 47*, 329–337.

Teebor, G. W., & Becker, F. F. (1971) *Cancer Res. 31*, 1–3.

Tykocinski, M. L., & Max, E. E. (1984) *Nucleic Acids Res. 12*, 4385–4396.

Vanyushin, B. F., Tkacheva, S. G., & Belozersky, A. N. (1970) *Nature (London) 225*, 948–949.

Vanyushin, B. F., Mazin, A. L., Vasilyev, V. K., & Belozersky, A. N. (1973) *Biochim. Biophys. Acta 299*, 397–403.

von Hippel, P. H., Bear, D. G., Morgan, W. D., & McSwiggen, J. A. (1984) *Annu. Rev. Biochem. 53*, 389–446.

Williams, J. R. (1978) *Anal. Biochem. 86*, 339–340.