Volume 16 Number 20 1988 Nucleic Acids Research

Arrangement of Dam methylation sites (GATC) in the *Escherichia coli* chromosome

Frederic Barras+ and M.G.Marinus*

Department of Pharmacology, University of Massachusetts Medical School, Worcester, MA 01655, USA

Received June 16, 1988; Revised and Accepted August 31, 1988

ABSTRACT

The occurrence of GATC (Dam-recognition) sites in available E. coli DNA sequences (representing about 2% of the chromosome) has been determined by a simple numerical analysis. Our approach was to analyze the nucleotide composition of nine large sequenced DNA stretches ("cantles") in order to identify patterns of GATC distribution and to rationalize such patterns in biological/ structural terms. The following observations were made:(i) In addition to oriC, GATC-rich regions are present in numerous locations. (ii) There is a wide variation in GATC frequency both between and within DNA cantles which led to the identification of a void-cluster pattern of GATC arrangement. The distance between two GATCs was never greater than 2 kb. (iii) GATC sites are found more frequently in translated regions than (in decreasing order) non-coding or non-translated regions. In particular, rRNA and tRNA encoding genes exhibit the lowest GATC content.

INTRODUCTION

In Escherichia coli, the 5'-GATC-3' palindromic sequence is methylated by the dam gene product. Modification takes place at the N6 position of the adenine residue and is thought to occur shortly after the passage of the replication fork. The existence of both methylated and unmethylated forms acts as a regulatory signal for directing the mismatch repair machinery (1). Potential roles for the methylation of GATC sites have been invoked in various aspects of the cell cycle, such as replication initiation, recombination and gene expression (2,3).

The hypothesis that GATC sites are involved in the replication initiation process comes partly from the observation that 11 of these sites are located within the minimal oriC region (4). Since the minimal origin is 250 bp in size, this number of GATCs has been interpreted as being unusually high compared to what would be expected from a random occurrence, i.e., 1 GATC in 256

bp. However, this assumption is not valid since it has unambiguously been shown that the nucleotide composition of the E. coli genome is non-random (5). This non-randomness prevents most statistical analyses from achieving sound predictions. Recently, Phillips et al.,(6,7) tested several statistical methods and concluded that a Markov chain is the most accurate in predicting oligonucleotide frequency. This allowed these authors to show that codon usage strongly influences oligonucleotide composition of the E. coli genome. However, this method failed to predict accurately the frequency of occurrence of 45 of the 256 possible tetranucleotides, among which was GATC. The existence of a counterselection acting specifically at the tetranucleotide level was postulated (6,7,8).

The role played by the cluster of GATC sites at oriC is unclear. It has been proposed that they regulate the formation of secondary structures and influence the binding of ancillary proteins for the initiation of chromosome replication (9,10,11). Recently, Ogden et al.,(12) have shown that hemimethylated, but not fully methylated, oriC DNA binds to an outer membrane fraction of E. coli. This suggests that methylation of GATC sites in oriC is important in efficient segregation of chromosomes. Other GATC clusters have been identified within the dnaA (13) and metL genes (8) but their function is unknown.

Hemimethylated GATC sites allow the mismatch repair system to recognize and act upon the newly synthesized strand. These sites are recognized by an enzyme, MutH, that initiates the repair process by nicking DNA 5′ to GATC (14). Recent studies have revealed the existence of a correlation between the number of GATC sites within a DNA duplex and the efficiency of the repair system to correct a mismatch (15). In addition it has been shown that the distance between a mismatch and a GATC site should not be more than 1 kb for efficient repair to take place (16). Since mismatch repair is thought to work in both directions from a given GATC, a DNA region that is flanked by two GATCs 2 kb apart should be protected from mutation induction.

The methylation of GATC has also been shown to affect gene expression and transposition (2). In both cases GATCs were located within DNA regions such as promoters and transposon ends.

The state of methylation can thereby influence the interaction between DNA and the relevant proteins such as RNA polymerase and transposase (3).

It appears, therefore, that both the organization and location of GATC sites could superimpose signals in addition to adenine methylation. As a first step toward the characterization of this additional level of GATC-mediated regulation we asked three questions. (i) Does the clustering of GATC sites at oriC represent a specific and exclusive feature of that chromosomal region? (ii) Does the distribution of GATC sites in the chromosome exhibit a particular pattern? (iii) What are the factors, if any, that influence GATC distribution? We answered these questions by developing a numerical analysis that used a set of large stretches of sequenced DNA. We have termed each of these a cantle (meaning a slice or cut). A cantle is several kb of contiguous nucleotides of chromosomal DNA containing genes, their regulatory elements and/or non-translated and non-coding regions. (The opposite approach is to use a large number of short non-contiguous sequences, "cantlings", for analysis). We have also searched for recurring patterns within DNA cantles. Our findings allowed us to rationalize the present distribution of GATC sites within the context of evolution of the E. coli chromosome.

MATERIALS AND METHODS

GenBank,release 48, was obtained from BBN Labs,Inc. It was converted to an MS-DOS usable format by Dr. John Sloan of the Computing Center at the University of Massachusetts Medical School. The programs of Conrad and Mount (17) and Mount and Conrad (18), version 3.8, were used for DNA sequence analysis. These programs were obtained from Dr. D.W. Mount at the University of Arizona. DNA sequences were translated into amino acids using the translation programs of DNASTAR.

The GenBank files used and our abbreviations for them (in parentheses) were as follows: ECOGLTA(GLT); ECOUNC(UNC); ECOTRP (TRP); ECOLAC(LAC); ECOACE(ACE); ECONRDA(NRD); ECOTHRI.NF(THR); ECORPL.PO(RPL) and ECORGNB(RGN). The genes encoded by each of these cantles is shown in Figure 1. The size and G+C content of each cantle is given in Tables 1 and 2 respectively. The thyA-

argA sequence was kindly provided by Drs A. Taylor (F. Hutchinson Cancer Research Center,Seattle,USA) and P. Emmerson (Univ. of Newcastle,UK).

RESULTS AND DISCUSSION

Frequency of occurrence of GATC and other tetranucleotides within DNA cantles.

The mean value of GATC occurrence, 1/241 nucleotides (nt), as reported by Phillips et al.,(6,7) comes from pooling a large

Table 1. GATC content of the nine DNA cantles

Cantle	Size (nt)	Coding	Non-coding	GATC Number Total	Coding	Non-coding
GLT	13,063	10,554 (81%)	2,509	66 (1/198)	56 (1/188)	10 (1/250)
UNC	7,540	6,654 (88%)	886	40 (1/188)	38 (1/175)	2 (1/443)
TRP	7,330	6,558 (90%)	772	27 (1/271)	25 (1/262)	2 (1/386)
LAC	7,477	6,020 (81%)	1,457	26 (1/288)	20 (1/301)	6 (1/243)
ACE	7,740	6,686 (86%)	1,054	40 (1/194)	40 (1/167)	0
NRD	8,554	3,460 (40%)	5,094	47 (1/182)	30 (1/115)	17 (1/300)
THR	7,784	6,651 (84%)	1,233	38 (1/205)	36 (1/185)	2 (1/616)
RPL	12,337	10,245 (83%)	2,092	56 (1/220)	54 (1/190)	2 (1/1,046)
RGN	7,508	4,638 (62%)	2,870	17 (1/442)	8 (1/580)	9 (1/319)
TOTAL	79,333	61,366	17,967	357 (1/222)	307 (1/200)	50 (1/360)

The frequency of occurrence values (in parentheses) represent the number of GATC sites in a given cantle (columns 4, 5, or 6) divided by the size of that cantle (columns 1, 2 or 3). Coding and non-coding regions were those assigned in GenBank files. The values expressed as percentages (column 2) represent the proportion of coding region within each cantle. The genes located within these cantles are shown in Figure 1.

number of non-contiguous DNA cantlings. This value is in agreement with experimental data showing that the limit digest product, after incubation of E. coli DNA with restriction endonucleases which cut at GATCs, is about 300 nt (19,20). In contrast to the approach of Phillips et al.,(6,7), we decided to look at the frequency of occurrence of GATC within different large cantles. We chose nine E. coli sequences of contiguous nucleotides in GenBank which were more than 7,000 nt in size; the shortest and the longest cantles were 7,330 nt and 13,063 nt, respectively (Table 1). The GATCs were counted and frequency of occurrence values deduced (Table 1; column 4). The values obtained ranged from 1/182 nt (NRD) to 1/442 nt (RGN). In six cantles the values were higher than that observed by Phillips et al.,(6,7). However, when the cantles were pooled the mean value of occurrence came closer to that of Phillips et al.,(6,7), i.e.,1/222 nt. These results suggested that the distribution of the GATCs is asymmetrical within the E. coli chromosome.

As a control, within each DNA cantle we analyzed the occurrence of the 24 tetranucleotides that were a permutation of the four nucleotides (A, C, G, T). In all nine cantles a clear hierarchy was observed and there was more than a 100-fold difference between the most and the least abundant tetranucleotide (Table 2). For instance RPL contains 123 CTGAs and only one CTAG. We then ranked the 24 tetranucleotides according to their frequency of occurrence. When cantles were compared, the relative position of a given tetranucleotide was approximately the same with the exception of that found within RGN (data not shown). Likewise we positioned all 24 tetranucleotides around an F value, specific for each cantle and representing the mean value of occurrence of a (A, C, T, G) containing tetranucleotide (Table 2). We observed that the tetranucleotides listed from CTGA to TGAC can be considered as a high density group while those listed from TACG to CTAG were always part of a low density group (Table 2). GATC was always part of the high density group except in RGN. This cantle behaved differently from all others since the tetranucleotides were more evenly distributed and both GATC and ATCG frequencies were reduced (see below) while that of CTAG was increased (Table 2). The unusual features of RGN highlights how

Table 2. Occurrence of the 24 (A, C, T, G) containing tetranucleotides within selected DNA cantles.

	GLT	UNC	TRP	LAC	ACE	NRD	THR	RPL	RGN
CTGA	++	++	++	++	++	+	++	++	++
	(80)				(72)		(53)	(123)	(47)
ACTG	+	++	++	+	++	+	++	++	++
							(45)	(87)	(44)
ATCG	+	++	+	+	+	+	+	+	-
	(71)	(53)			(46)				(16)
TCAG	-	+	+	++	+	++	+	+	+
				(43)		(56)			
GATC	+	+	+	+	+	+	+	+	-
TGAC	+	+	+	-	+	+	+	+	+
CGAT	+	-	++	++	+	++	+	-	-
			(41)	(49)		(54)			
CGTA	+	+	-	-	+	+	+	++	+
TCGA	+	+	-	-	+	+	+	+	-
ACGT	+	++	-	-	+	+	+	+	-
		(49)							
GCAT	+	-	+	++	-	+	+	-	-
GTCA	-	-	+	+	+	+	+	-	-
TGCA	+	+	+	+	-	+	-	-	+
CAGT	-	-	+	++	-	+	-	-	-
AGCT	-	+	-	-	+	-	-	+	+
ATGC	+	-	++	-	-	-	-	-	+
			(41)						
GTAC	-	-	-	-	-	-	+	+	-
TACG	-	-	-	+	-	-	-	-	-
CATG	-	-	-	+	-	--	-	-	-
						(17)			
GCTA	--	-	-	-	-	-	+	-	-
	(20)								
GACT	-	-	-	--	-	-	-	-	-
				(12)					
AGTC	-	--	-	-	-	-	--	--	-
							(14)	(20)	
TAGC	--	--	--	--	--	-	-	--	-
		(11)	(8)	(12)	(10)				
CTAG	--	--	--	--	--	--	--	--	--
	(1)	(0)	(3)	(1)	(0)	(3)	(3)	(1)	(10)
T	1139	716	604	606	701	840	683	1275	615
F	47.5	30	25	25	29	35	28	53	26
G+C	52	52	53	53	53	50	51	51	51

Arithmetical signs indicate the position of a tetranucleotide vis-a-vis the F value. If X is the number of a given tetranucleotide within a given cantle, then signs are assigned as follows : ++ when X > 0.75 F; + when 0.5 F < X < 0.75 F; - when 0.25 F < X < 0.5 F; -- when X < 0.25 F. Numbers of the two most and two least frequent tetranucleotides are indicated for each cantle. The letter T stands for the total number of (A, C, T, G) containing tetranucleotides within a given cantle and F = T/24 represents the average value of occurrence of such a tetranucleotide within that cantle. G+C content of each cantle is shown in the bottom line of the Table.

statistical methods for predicting the number of tetranucleotides within a given chromosomal region can be misleading.

Overall, both the values and rankings of tetranucleotides obtained were consistent with those reported by Phillips et al., (6,7) while the DNA samples used are quite different. We found the number of GATCs to vary between cantles, yet exhibiting among the 24 tetranucleotides a steady relative frequency throughout eight of the nine DNA cantles. This suggests the existence of a positive selection that allows the number of GATCs to be consistently kept among the most frequently found (A, C, T, G) containing tetranucleotides. This positive influence could originate at different oligonucleotide levels (ie.,di-,tri- or tetra-). For instance, we noted that GA and TC dinucleotides were mostly present in abundant tetranucleotides (Table 2). Likewise we noted that GATC and ATCG which contain the ATC trinucleotide, behaved similarly throughout all cantles, including within RGN where a sharp drop in frequency was observed for both (17 GATCs and 16 ATCGs). Given that RGN contains only untranslated genes (Figure 1), this suggests that the over-abundance of these two tetranucleotides in other cantles is related to codon usage (see below). It was also noted that of the 24 tetranucleotides, eight are palindromes of which GATC is one. These are evenly split in the high versus low groups and hence, there was not obvious selection for or against palindromes per se.

One of the most striking observations was the scarcity of the CTAG tetranucleotide in the eight cantles. This has already been noticed in E. coli (6,7) and a wide array of other bacterial genomes (21). In this regard it is noteworthy that the only restriction enzyme known to cut at CTAG sites is produced by an archaebacterium, Methanococcus aeolicus (22) and a survey of the few available archeabacterial DNA sequences seem not to indicate a comparable lack of CTAG (unpublished data).

Identification of voids and clusters of GATC sites within E. coli genes

Different frequencies of GATC occurrence were found within the nine cantles. One explanation is that GATCs are distributed in voids and clusters along the E. coli chromosome. In this

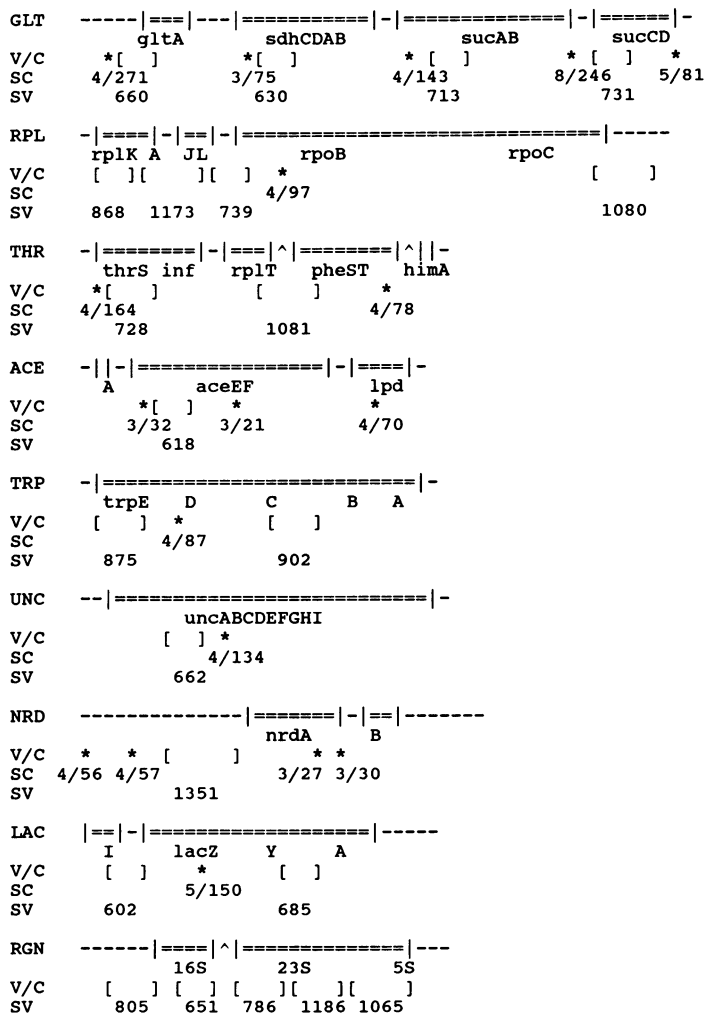

Figure 1. Genetic organization and structural features of the nine DNA cantles

DNA cantle designation are abreviations of GenBank files. The location and name of the genes are indicated. Symbols * and [ ] represent a GATC cluster and a void, respectively. Blank spaces indicate the absence of a void or a cluster. Letters V, C, SV, SC stand for void, cluster, size of void and size of cluster, respectively. Lines SC and SV contain the numerical values characterizing cluster and void structures, respectively. The ^ symbol was used for designating the location of a tRNA encoding gene (see THR and RGN). The figure is approximately to scale with a - representing 250 bp. Genes are indicated by double dashes (=).

regard the NRD and RGN cantles would be part of a cluster and a void, respectively. We further hypothesized that such voids and clusters should also be found within the cantles themselves. We therefore undertook an analysis of the distribution of the GATC sites within each cantle. Limits for a void and a cluster were arbitrarily defined as follows. A void was a region of DNA which harbored no GATC site and was at least 600 nt in size; this corresponds to a three-fold reduction in frequency value as compared with the expected mean value. A cluster was a DNA region exhibiting either three GATCs separated from each other by less than 30 nt or at least four GATCs each separated by less than 45 nt; this corresponds to more than a seven-fold increase in frequency value as compared with the expected mean value.

From the analysis of all cantles, two observations were significant (Figure 1). First, voids and/or clusters were found in almost all cantles. The identified voids ranged from 600 nt (our lower limit for a void) up to 1351 nt (NRD). Similarly,

Table 3. Identification of voids and clusters within different *E. coli* genes.

Gene(s)	Void (nt)	Limits	Cluster A	Cluster B	Limits
gltX			9	(1/51)	227-686
malP			9	(1/40)	1437-1795
btuB	1202	1-1203			
hisT	1346	824-2170			
hsdS	1344	2-1346			
supBE	>1100				
tyrT	1618	268-1946			
rpmH/dnaA	>632	?-632	9	(1/25)	632-854
	780	854-1634	9	(1/29)	1634-1892
guaBA	1083	1597-2680	4	(1/58)	886-1118
metL	600	1065-1665	10	(1/40)	1665-2071
pbpB	1236	992-2228	6	(1/55)	249-578
rpsP	690	1237-1927	5	(1/29)	1927-2071
rplS	1137	2982-4119	4	(1/42)	2402-2572
xylB	1118	1787-2905	4	(1/47)	1600-1787

Cluster and void are defined in the text. Numerical values defining void and/or cluster limits are taken from GenBank file numbering. Columns A and B are the number of GATCs involved in a cluster and for the GATC frequency value within that cluster, respectively. No flanking DNA sequence was available for setting the boundaries of the void identified within the supBE region, thus refered to as larger than 1100 nt. As a reference the oriC region contains 11 GATC sites within 250 nt,i.e.,a 1/23 frequency.

with the exception of RGN, all cantles contained at least one cluster. These varied in the number of GATC sites involved (from three to eight) and were generally present within the coding regions. Second, the frequency of associated voids and clusters (i.e.,a cluster followed by a void followed by another cluster) was high. This was true in GLT, NRD, ACE, UNC and partly in RPL, THR and TRP. Gap without an accompanying cluster was found in RPL, RGN and TRP. In RPL, however, the identified void is on the edge of the cantle and the presence of a nearby cluster remains a possibility.

Given that the size of the DNA cantles constituted the sole basis of their selection for analysis, the recurring presence of clusters and voids suggested these structures are widespread within the chromosome. In order to test this prediction we searched all E. coli sequences (excluding the cantles) in the GenBank release for the identification of other void-cluster and associated void and cluster structures (Table 3). Associated void-cluster arrangements were found within the dnaA, metL, htpR, rpsP and xylB regions (Table 3). The rpmH-dnaA gene region, for example, is composed of a void, the size of which could not be estimated but is greater than 632 nt, followed by the cluster present in the dnaA promoter, then another void of 780 nt was followed by a second cluster containing the 3' region of the dnaA gene (Table 3).

Three clusters were found (in gltX, malP and metL) which contained as many GATCs as exhibited by the rpmH-dnaA region although the frequency value was lower, 1/25 vs. 1/40. In contrast, a cluster with frequency similar to that of rpmH-dnaA was identified within the rpsP gene but exhibited only five sites versus nine.

Two further observations were noted. First, two of the largest GATC free regions harbor tRNA encoding genes, namely tyrT and supBE (Table 3). This, added to the fact that five voids were identified within a cantle (RGN) containing a tRNA as well as the 16S, 23S and 5S encoding genes, strongly suggests that GATC sites have been selected against within chromosomal regions containing untranslated genes.

Second, the largest void identified was 1618 nt (in the tyrT

gene). Since the mismatch repair system acts bidirectionally for a distance of about one kb (16) this suggests that the whole E. coli chromosome is suceptible to Dam-directed mismatch repair.

Influence of codon usage upon GATC site distribution and frequency

Phillips et al.,(6,7) observed that, in contrast to most of the tetranucleotides, no satisfying correlation could be found between codon usage and GATC frequency. Since the number of observed GATCs was lower than expected, these authors suggested that an unidentified negative selection was acting at the level of the GATC tetranucleotide. In contrast, by comparing both GATC and ATCG distribution throughout the nine DNA cantles, we speculate that codon usage positively influenced the occurrence of GATC sites. The GATC tetranucleotide can occur in four codons depending upon the reading frame: GAU and AUC, which encode aspartate and isoleucine or the pair NGA, UCX which encode arginine (AGA, CGA) or glycine (GGA) and serine (UCX). By examining all known genes in the eight cantles (RGN was not included) we found that of 299 GATCs, 296 were in the GAU and AUC reading frame. Of these 296, 41% were GAU and the remainder were AUC.

The possible influence of codon usage was addressed in the eight DNA cantles by asking how often were GAU and AUC codons used and how many of them correspond to GATC in DNA. Of the two codons for aspartate, GAU was found to be prefered in both TRP and LAC (Table 5, col. 2). Although the use of GAU as a codon was reduced in the five remaining cantles, it was still substantial (approximately 40%). Conversely, the use of AUC as an isoleucine codon was reduced in both LAC and TRP but preferred in the other cantles. The proportion of GAU and AUC codons in GATC sequences was 20% to 30% (Table 5). These values indicate that there is no specific selection for GATC derived GAU and AUC codons. This is consistent with the finding that both GA and TC dinucleotides recur at the same average frequency within the E. coli chromosome (5,23).

A reduced number of (G)AUC codon sequences was noted in both TRP and LAC. This suggested a reason for the low number of GATCs in these cantles. Note that in both ACE and RPL, the reduced number of GAU codons was balanced by the highly used AUC, thereby

Table 4. Distribution of GATC sites within coding and non-coding regions.

DNA cantle	X	Y	R=X/Y
GLT	4.2	5.6	1.3
UNC	7.5	19	2.5
TRP	8.5	12.5	1.5
LAC	4.1	3.3	0.8
ACE	6.3	-	-
NRD	0.7	1.8	2.6
THR	5.3	18	3.4
RPL	4.9	27	5.5
RGN	1.6	0.9	0.6

X is the ratio between coding and non-coding regions lengths. Y is the ratio between number of GATCs within coding and non-coding regions. An R value higher than 1 indicates that the number of GATCs is higher in coding than in non-coding regions.

resulting in a high number of GATCs (Table 1). Thus, codon usage can account for the assymetric distribution of GATCs in coding regions (Table 1). This conclusion neither supports nor rules out the existence of a specific negative selection acting at the tetranucleotide level as suggested by Phillips et al.,(6,7). It is, however, important to mention that the divergences observed by Phillips et al., (6,7) between expected and actual numbers of GATCs could have been artificially increased by the fact that the rrn genes (i.e., our RGN cantle) contributed for approximately 10% of their sample size. Such a proportion might have introduced a bias into the statistical predictions.

Distribution of GATC in coding and non-coding chromosomal regions

The average frequency of occurrence of GATCs in coding

Table 5. Influence of codon usage on GATC frequency.

DNA cantle	D	GAU / D	GAU(C) / GAU	I	AUC / I	(G)AUC / AUC
GLT	207	50%	27%	196	64%	25%
UNC	106	39%	22%	149	63%	30%
TRP	115	63%	25%	113	38%	16%
LAC	96	64%	20%	107	36%	15%
ACE	120	38%	18%	115	76%	28%
NRD	66	42%	36%	76	58%	41%
THR	131	49%	28%	130	49%	26%
RPL	205	35%	30%	207	79%	20%

Aspartate (D) and isoleucine (I) residues were counted after translation in each gene of all cantles. GAU(C) and (G)AUC represent GAU and AUC codons with a 3'C residue and a 5'G residue, respectively.

versus non-coding regions within each of the nine cantles is shown in Table 1. Non-coding indicates the intergenic regions and does not include non-translated (tRNA, rRNA) genes. A comparison of these values suggested that the number of GATCs is higher within coding than non-coding regions. This was directly shown by the use of a coefficient, R, the value of which is more than 1 whenever the relative number of GATCs is higher within coding portions (Table 4). Such was the case in all but two cantles, namely RGN and LAC. When the nine cantles were pooled, the mean values of occurrence were 1/200 nt and 1/360 nt within the coding (61,366 nt) and non coding (17,967 nt) regions, respectively. These results clearly indicate a correlation between the location of GATC sites and the organization of the DNA in coding and non-coding cantles.

This arrangement was also found when a similar analysis was performed with all E. coli DNA sequences (other than the cantles) in the GenBank release. In this case the values of occurrence frequency were found to be 1/196 nt and 1/292 nt within coding (55,359 nt) and non-coding (10,510 nt) regions, respectively (data not shown).

As noted above, the LAC and RGN cantles appeared to be exceptions to the general pattern of more GATCs in coding regions. In LAC, five of the six GATC sites, which were presumably within a non-coding region, were located 3' to the lacA stop codon within a 1113 nt sequence (Figure 1), giving a 1/213 nt frequency value. The remaining GATC was found within the lacY-lacA intercistronic region. Considering the size of the total untranslated regions of the lac operon, this gives a 1/313 nt frequency value. The two values, 1/213 nt and 1/313 nt, are very close to that characterizing the coding and non-coding regions, respectively. Therefore it is probable that most, if not all, of the five GATC sites are actually within the coding region of an as yet unidentified gene. Such a prediction is consistent with the identification of an open reading frame 3' to lacA that would contain three of the five GATC sites (24).

In RGN the mean value of GATC occurrence obtained within the non-coding regions compared favorably with those obtained in other cantles (Table 1). In contrast, the coding region charac-

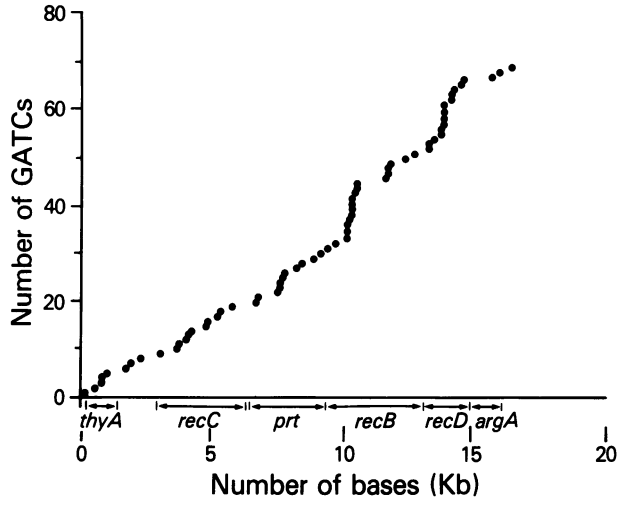

Figure 2. Distribution of GATC sites within the thyA-recD-ptr-recB-recC-argA region. Each dot represents a GATC.

terizing value was the lowest found in any cantle (Table 1). This meant that the overall low content of GATC in that cantle is due to the lack of GATC within the coding parts. From an examination of the five GATC sites present within the 16S rrnB gene and their location within the secondary structure of the encoded rRNA (25) the following points were noted: (i) Two GAUCs are found at positions 14 and 1529 of E. coli which are conserved among the three "kingdoms" (eubacteria, archaebacteria and eukaryotes) (ii) One GAUC is conserved in 90% of the eubacterial catalog (position 1355) and (iii) each of the five GAUCs lie within one helical element in the RNA secondary structure. This strongly suggests that GATC sites present within the 16S rrnB gene are selected with regard to their contribution to the secondary structure of the cognate rRNA.

On the other hand, the presence of palindromic GATC sequences at specific locations in the rRNA molecule, other than those described above, might hamper secondary structure formation. If so,other palindromic tetranucleotides, in addition to GATC, should be under-represented in these genes. Indeed, Phillips et al.,(6,7) found that among the under-represented tetranucleotides there was an unexpectedly large number of palindromic ones.

Use of a tenth DNA cantle

After this work was completed, the argA gene nucleotide sequence was published (26). The addition of this sequence to that of the contiguous genes, made it the largest available continuous nucleotide sequence, 17 kb, of E. coli DNA (Fig. 2). It was therefore of interest to ask whether the main features identified within the nine cantles above would be present in this new cantle. Figure 2 shows the increase in the number of GATC sites as a function of the number of nucleotides analyzed. Several voids were identified including two of 1110 nt and 1074 nt. Two of the coupled void-cluster structures are the most dense so far identified in the E. coli chromosome: seven GATCs in 116 nt (1/16.5) in the recB gene and seven in 76 nt (1/10.8) in the recD gene. The relative number of GATCs in the coding region (1/216) was higher than that of non-coding region (1/511). Thus, clearly all expected structural features are exhibited by this new cantle, thereby strengthening the conclusions deduced from the study of the first nine DNA cantles.

Concluding Comments

In this study, we showed that (i) GATC rich regions are present in numerous locations within the E. coli chromosome, (ii) regions of more than 1 kb in size that lack GATC sites are readily found, (iii) there is a correlation between codon usage and GATC frequency and distribution and (iv) untranslated genes have a low GATC content. In addition we postulated the existence of (i) associated void-cluster structures and (ii) a correlation between the location of GATC sites and the organization of the DNA in both coding and non-coding regions. Furthermore, these observations and postulates were found to apply to the longest DNA cantle to date.

The existence of GATC clusters distinct from those at oriC and dnaA, raises the question of the specific role GATC plays within the chromosomal replication initiation process. It might be significant that potential DnaA binding sites (i.e., sequences homologous to the DnaA binding sites in oriC with 7/9 or 8/9 matching scores (4)) were identified in the vicinity of these GATC clusters (unpublished data). It is possible, therefore, that these clusters constitute either ancient or secondary replication

origins (27) or, alternatively, these may be regions which can also bind to the E. coli outer membrane to ensure chromosome segregation (12).

The existence of voids between GATCs will have to be considered when experimental systems have precisely defined the DNA size over which a GATC site can efficiently direct the action of mismatch repair. The lack of voids of more than 2 kb in size suggests that the whole E. coli genome is susceptible to Dam-directed mismatch repair. It is, however, likely that changes in nucleotide composition render different DNA regions more or less susceptible to mismatch occurrence and hence more or less demanding to the repair system. It will be of interest to compare mutability rates and nucleotide sequences of regions that exhibit a similar pattern of GATC distribution.

The clustering of GATCs can be considered in regard to the functioning of the Dam-directed mismatch repair system. Besides the Dam enzyme, at least one other component of the repair machinery, MutH protein, recognizes and acts at GATC sites (14). Clustering GATC sites within a DNA region may have a synergistic effect on MutH nicking activity resulting in increased repair efficiency.

Both codon usage and an unidentified specific negative selection have been postulated to influence the GATC frequency (6,7). We have confirmed the importance played by codon usage in both distribution and frequency of GATCs. Moreover, even though codon usage varies within E. coli genes, the fact that GATC occurrence depends upon two codons allows compensatory effects to take place. This ensures representation of GATC sites throughout the E. coli genome. An unexpected observation was the sharp decrease in frequency of GATC in both the untranslated genes and non-coding regions. Since methylation of GATC can hamper the contact between DNA and proteins, it seems likely that GATC sites have been selected against in regions that control gene expression.

From an evolutionary perspective, we speculate that the occurrence of GATC was first dictated by its potential influence upon RNA (and DNA ?) secondary structure; second,by the rules of codon usage and third, by the regulatory roles methylation pro-

vided them with. A corollary is that the selection of GATC as a signal directing mismatch repair originates from the mode codon usage had distributed these tetranucleotides within the E. coli chromosome. A prediction then is that those bacterial species that contain a Dam-directed mismatch repair should have similar codon usage.

The choice of dealing with independent DNA cantles allowed us (i) to observe patterns common to all cantles (ii) to predict such patterns within other DNA cantles and (iii) to point out the unusual behavior of one of the cantles, RGN. The impending complete nucleotide sequence of the whole E. coli genome will render statistically based predictions of lesser use. In contrast a method such as the one described here should be more adaptable since it allows one both to formulate and verify hypotheses while using a particular sequence as a guide for exploring genome structure.

ACKNOWLEDGMENTS

We are indebted to Dr. Martine Crasnier for her continued encouragement as well as her thoughtful advice and suggestions. Stimulatory discussions with Dr. Antoine Danchin led to the initiation of this study. Dr. R. Lew provided statistical advice. We thank Drs. Marc Chippaux, R. Ivarie and T. Bickle for comments on the manuscript, Arlene Semerjian for assistance with the computer analysis and Dr. John Sloan for assistance with the computer files. Dr. J. Jiricny shared unpublished results and Drs. A. Taylor and P. Emmerson provided the thyA-argA sequence. This work was supported by grant GM30330 from the National Institutes of Health.

*To whom correspondence should be addressed

+Present address: LCB, CNRS, 31 Chemin Joseph Aiguier, 13009 Marseille, France

REFERENCES

1. Pukkila,P.,Peterson,J.,Herman,G.,Modrich,P. and Meselson,M. (1983) Genetics 104, 571-582.
2. Marinus,M.G. (1987) Ann. Rev. Genet. 21, 113-131.
3. Sternberg,N. (1985) J. Bacteriol. 164, 490-493.
4. Zyskind,J.W. and Smith,D.W. (1986) Cell 46, 489-490.
5. Nussinov,R. (1987) J. Theor. Biol. 125, 219-235.

6. Phillips,G.J.,Arnold,J. and Ivarie,R. (1987) Nucl. Acids Res. 15, 2611-2626.
7. Phillips,G.J.,Arnold,J. and Ivarie,R. (1987) Nucl. Acids Res. 15, 2627-2638.
8. McClelland,M. (1985) J. Mol. Evol. 21, 317-322.
9. Smith,D.W.,Garland,A.M.,Herman,G.,Enns,R.E.,Baker,T.A. and Zyskind,J.W. (1985) EMBO J. 4, 1319-1326.
10. Messer,W.,Bellekes,U. and Lother,H. (1985) EMBO J. 4, 1327-1332.
11. Russel,D.W. and Zinder,N.D. (1987) Cell 50, 1071-1079.
12. Ogden,G.B.,Pratt,M.J. and Schaechter,M. (1988) Cell 54,127-135.
13. Braun,R. and Wright,A. (1986) Mol. Gen. Genet. 202, 246-250.
14. Welsh,K.M.,Lu,A-L.,Clark,S. and Modrich,P. (1987) J. Biol. Chem. 262, 15264-15629.
15. Lu,A-L. (1987) J. Bacteriol. 169, 1254-1259.
16. Bruni,R.,Martin,D. and Jiricny,J. (1988) Nucl. Acids Res. 16, 4875-4890.
17. Conrad,B. and Mount,D.W. (1982) Nucl. Acids Res. 10, 31-38.
18. Mount,D.W. and Conrad,B. (1984) Nucl. Acids Res. 12,part 2, 819-824.
19. Lacks,S.A. and Greenberg,B. (1977) J. Mol. Biol. 114, 153-168.
20. Szyf,M.,Gruenbaun,Y.,Urieli-Shoval,S, and Razin,A. (1982) Nucl. Acid Res. 10, 7247-7259.
21. McClelland,M.,Jones,R.,Patel,Y. and Nelson, M. (1987) Nucl. Acids Res. 15, 5985-6505.
22. Schmid,K.,Thomm,M.,Laminet,A.,Laue,F.G.,Kessler,C.,Stetter, K.O. and Schmitt,R. (1984) Nucl. Acids Res. 12, 2619-2628.
23. Nussinov,R. (1984) Nucl. Acid Res. 12, 1749-1763.
24. Heidiger,M.A.,Johnson,D.F.,Nierlich,D.P. and Zabin,I. (1985) Proc. Natl. Acad. Sci. USA, 82, 6414-6418.
25. Woese,C.R. (1987) Microbiol. Rev. 51, 221-271.
26. Brown,K.,Finch,P.W.,Hickson,I.D. and Emmerson,P.T. (1987) Nucl. Acids Res. 15, 10586.
27. De Massy,B.,Fayet,O. and Kogoma,T. (1984) J. Mol. Biol. 178, 227-236.