J Mol Evol (1994) 39:340–346

Journal of Molecular Evolution

Compositional Heterogeneity of the *Escherichia coli* Genome: A Role for VSP Repair?

Gabriel Gutiérrez,[1] Josep Casadesús,[1] José L. Oliver,[2] Antonio Marín[1]

[1] Departamento de Genética, Universidad de Sevilla, Apartado 1095, E-41080 Sevilla, Spain
[2] Departamento de Genética, Universidad de Granada, Campus de Fuentenueva, E-18071 Granada, Spain

Received: 18 January 1994 / Accepted: 25 March 1994

Abstract. *E. coli* genes that contain a high frequency of the tetranucleotide CTAG are also rich in the tetramers CTTG, CCTA, CCAA, TTGG, TAGG, and CAAG (group-I tetramers). Conversely, *E. coli* genes lacking CTAG are rich in the tetranucleotides CCTG, CCAG, CTGG, and CAGG (group-II tetramers). These two gene samples differ also in codon usage, amino acid composition, frequency of Dcm sites, and contrast vocabularies. Group-I tetramers have in common that they are depleted by very-short-patch repair (VSP), while group-II tetramers are favored by VSP activity. The VSP system repairs G:T mismatches to G:C, thereby increasing the overall G+C content of the genome; for this reason the CTAG-rich sample has a lower G+C content than the CTAG-poor sample. This compositional heterogeneity can be tentatively explained by a low level of VSP activity on the CTAG-rich sample. A negative correlation is found between the frequency of group-I tetramers and the level of gene expression, as measured by the Codon Adaptation Index (CAI). A possible link between the rate of VSP activity and the level of gene expression is considered.

Key words: *E. coli* genome — VSP repair — CTAG tetranucleotide — G+C content — Contrast vocabularies — CAI

Introduction

Classic studies by DNA centrifugation showed that bacterial genomes differ broadly in base composition, namely, in G+C content, and for some time it was accepted that bacterial DNA pieces isolated from a given species had the same base ratio or showed narrow variation (Sueoka 1959; Rolfe and Meselson 1959). However, when a considerable number of sequence data have been available, intragenomic heterogeneity in G+C content has been found in prokaryotes (Nomura et al. 1987; Médigue et al. 1991b; D'Onofrio and Bernardi 1992; Sueoka 1992).

While there is a widely accepted explanation for bacterial intergenomic diversity in G+C content based on directional mutation pressure (Sueoka 1962; Jukes and Bhushan 1986; Muto and Osawa 1987), the mechanisms responsible for intragenomic variation have received less attention. Intragenomic G+C heterogeneity among bacterial genes has been mostly attributed to silent-site G+C content variation, a feature that can be largely explained by the uneven usage of synonymous codons, which is in turn related to the level of gene expression (Gouy and Gautier 1982; Sharp 1990; Sharp and Lloyd 1993). However, the possibility that directional mutation pressure affecting G+C content may have distinct effects on different regions of a single bacterial genome has not been ruled out (Nomura et al. 1987; Sueoka 1992). The possibility that intragenomic heterogeneity might reflect horizontal gene transfer has been also considered (Médigue et al 1991a,b).

This paper describes unequivocal examples of compositional variation within the *E. coli* genome. In our

Correspondence to: A. Marín

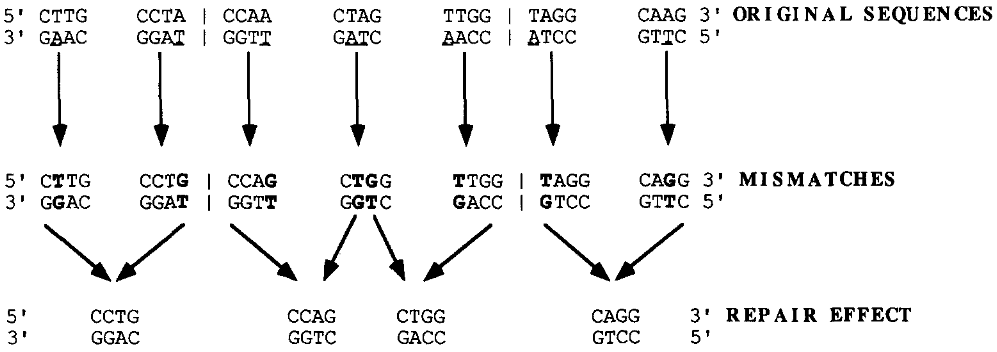
5' CTTG CCTA | CCAA CTAG TTGG | TAGG CAAG 3' ORIGINAL SEQUENCES
3' GAAC GGAT | GGTT GATC AACC | ATCC GTTC 5'
5' CTTG CCTG | CCAG CTGG TTGG | TAGG CAGG 3' MISMATCHES
3' GGAC GGAT | GGTT GGTC GACC | GTCC GTTC 5'
5' CCTG CCAG CTGG CAGG 3' REPAIR EFFECT
3' GGAC GGTC GACC GTCC 5'

Fig. 1. Tetranucleotides involved in the mutational effect of VSP. CTAG is included twice because the VSP repair system has two ways of repairing it. Mismatches are in **bold.**

model, the differences in base composition found among different DNA segments are tentatively explained by the differential activity of the base mismatch correction process called *very-short-patch* (VSP) repair (Lieb 1991). This methyl-independent process corrects T:G mismatches to C:G when found in the strings NTWGG/N′GW′CC and CTWGN/GGW′CN′; since it changes a T:G mismatch to a C:G even when T was the original correct base, it can produce T-to-C substitutions at some sites. Thus the VSP repair system, once viewed as a mechanism for mutation avoidance, may also play a role as a source of spontaneous mutation (Bhagwat and McClelland 1992; Merkl et al. 1992). The mutagenesis process, with regard to tetranucleotide strings, is summarized in Fig. 1.

The VSP repair process results in the underrepresentation of CTAG, CTTG, CCTA, CCAA, TTGG, TAGG, and CAAG (henceforth referred to as group-I tetramers) and the overrepresentation of CCTG, CCAG, CTGG, and CAGG (henceforth, group-II tetramers). Acting over an evolutionary time scale, VSP repair could be considered a main force shaping the nucleotide composition of *E. coli* genome: It actually provides an insightful explanation for its oligonucleotide composition, particularly for the extreme scarcity of the tetranucleotide CTAG (Bhagwat and McClelland 1992; Merkl et al. 1992; Burge et al. 1992).

Our search for compositional heterogeneity within the *E. coli* genome was based on the rationale that differential activity of VSP repair can be inferred from the frequencies of the oligonucleotides that are potential VSP targets. Because of its palindromic symmetry, CTAG is unique in that it can be doubly processed by VSP leading to CCAG and CTGG. Therefore, the frequency of CTAG tetramers within a segment of DNA can reflect the degree of activity of VSP repair on that DNA segment. Since VSP activity depletes the genome of CTAG, the regions where the observed frequency of CTAG is high must have been less affected by VSP repair than the regions where CTAG is rare or absent.

On these grounds, we have analyzed two samples of the *E. coli* genome selected according to their CTAG content: One sample contains DNA segments which are rich in CTAG; the other contains DNA segments lacking CTAG. We have found that these two samples differ in G+C content, codon usage, and frequencies of di- to pentanucleotides, thus providing a clear-cut example of intragenomic compositional heterogeneity. Finally, we speculate on the possible biological significance of VSP differential activity, since a tentative correlation can be established between the level of gene expression and the activity of VSP repair.

Data and Methods

DNA sequences were retrieved from GenBank Release 73.0 and analyzed with the UWGCG Sequence Analysis Software Package 7.1 (Devereux et al. 1984).

Two samples of entries longer than 1,000 bp and containing no plasmid sequences have been analyzed. One of the samples is CTAG-rich while the other is CTAG-poor.

The CTAG-rich sample contains 35 entries which do not share homologies and contain the CTAG tetranucleotide at frequencies higher than 0.07%; the average, overall frequency of this tetranucleotide in the *E. coli* genome is 0.02% (Bhagwat and McClelland 1992). The entry names of these sequences, together with their accession numbers and a short description of each sequence, are given in Table 1.

The CTAG-poor sample contains 82 entries randomly selected from a list of entries lacking CTAG. All the entries contain the query "complete cds" in the definition row.

FORTRAN programs were developed and executed on a VAX system. Measures of the Codon Adaptation Index (CAI) were performed according to Sharp and Li (1987); comparison of "contrast vocabularies" was as defined by Pietrokovski et al. (1990). We used the FORTRAN program of Lébart and Fenelon (1975) to carry out a factorial correspondence analysis on the frequencies of tetramers of groups I and II, excluding CTAG in the samples analyzed; CTAG was excluded because it had been used to stratify the samples. Statistical tests were carried out by using the program P3D of the BMDP Statistical Software Package (Dixon and Brown 1979).

Results

Tetranucleotide Frequencies

As a test for differential activity of VSP repair, we compared the frequencies of group-I (except CTAG) and

Table 1. Entries from Genbank contained in the CTAG-rich sample[a]

Entry name	Accession number	Basic description
eco1721dna	X61367	Gram-negative bacteria Transposon Tn1721; *tetA, tetR, tnpA, tnpR* genes
eco21sul1	X15371	Transposon Tn21 *sulI* gene, 3′ conserved region of integron
eco571mr	M74821	Restriction endonuclease (*eco57IR*) and methyltransferase (*eco57IM*) genes
eco67rtdm	M55249	Retron Ec67 DNA encoding reverse transcriptase and Dam methylase functions
ecoappyaa	M24530	Transcriptional regulatory protein gene (*appY*)
ecobglts	D00626	B glutamate carrier (*gltS*) gene
ecocfaia	M55661	CFA/I fimbrial operon (*cfaA, cfaB, cfaC, cfaE, cfaD*) genes
ecoclaa	M64113	Retronphage (phi)R73, partial sequence
ecocs3p	X16944	Genes involved in synthesis of CS3 pili
ecodsdaa	M19035	D-serine deaminase activator (*dsdC*) gene
ecoendx	M26404	*EcoRII* endonuclease gene
ecoepecae	M58154	Attaching and effacing (*eae*) gene
	M34051	
ecofimbe	X03923	Genes *fimB, fimE,* and N-terminus of *fimA* (type 1 fimbriae)
ecofanab	X05797	Genes *fanA* and *fanB* involved in biogenesis of K99 fimbriae
ecohsdd	V00287	Specificity gene of EcoD restriction enzyme (*hsdS*)
ecois2is30	X62680	Insertion sequences of IS2 and IS30
ecokanra	M84115	Kanamycin resistance protein (*neo*) gene, putative
ecolit	M19634	*lit* gene encoding a bacteriophage T4 late gene expression blocking protein (gplit)
	M80599	
ecolposacr	M86935	Lipopolysaccharide core biosynthesis protein operon (*rfaQ, rfaP, rfaG, rfaS, rfaB, rfaI,* and *rfaJ*) genes
econeua	J05023	CMP-*N*-acetylneuraminic acid synthetase (*neuA*) gene
econeuc	M84026	Protein p7 (*neuC*) gene
econeukps	M76370	*neuE* gene, 3′ end; glycosyl transferase (*neuS*) gene and *kps* gene, 3′ end
ecoompt1	X06903	*ompT* gene for outer membrane protein
	V00316	
	J01662	
ecophof	X06652	gene *phoE* encoding the phosphate-limitation-inducible outer membrane pore protein
ecoproret	Z12832	*proA* and *ret* genes encoding gamma-glutamylphosphate reductase and reverse transcriptase
ecopss	M58699	Phosphatidylserine synthase (*pss*) gene
ecorfa	M95398	Lipopolysaccharide core synthesis (*rfaY, rfaZ, rfaL,* and *rfaK*) genes
ecorfada	M33577	*rfaD* gene
ecorgnb	J01695	rRNA operon (*rrnB*) coding for $tRNA^{Glu}$-2, 5S, 16S, and 23S rRNA
ecosat2	X51546	SAT-2 gene for streptothricin-acetyltransferase
ecotn1000	X60200	Transposon Tn1000 (γδ) *tnpR* and *tnpA* genes for resolvase and transposase
ecotn4522	M17618	Tn4521 right junction
ecotn7tns	X17693	Transposon Tn7 transposition genes *tnsA, B, C, D,* and *E*
ecotnpa	Y00502	Tn2501 *tnpA* gene for transposase
ecotrpz	M38366	Intercistronic DNA in the 5′ boundary of the tryptophan (*trp*) operon

[a] The 585 bases at the 3′ end from Ecompt1 and the 150 at the 5′ end from Econeuc were deleted since they overlapped with Ecoappyaa and Econeua, respectively.

group-II tetramers in our two samples (CTAG-rich and CTAG-poor). On long-term evolution, it can be expected that the mutagenesis process driven by VSP will deplete the *E. coli* genome of group-I tetramers, while group-II tetramers are predicted to accumulate; these expectations are fulfilled by the observation that group-I tetramers are present at less-than-expected frequencies in the overall genome, while the opposite occurs to group-II tetramers (Phillips et al. 1987a,b; Bhagwat and McClelland 1992; Merkl et al. 1992).

As shown in Table 2, the observed averages of group-I tetramers are consistently lower in the CTAG-poor than in the CTAG-rich sample, while the opposite occurs with group-II tetramers. Figure 2 shows the results of the factorial correspondence analysis applied on a data matrix containing the tetranucleotide frequencies, excluding CTAG. The main observation is that entries belonging to the CTAG-poor sample are clustered in the lower half of the plot while entries corresponding to the CTAG-rich sample are scattered in the upper half.

Frequency of Dcm Sites

If the CTAG-rich sample undergoes low VSP activity, the frequency of Dcm sites (CCWGG) should be lower in the CTAG-rich sample than in the CTAG-poor sample. In the absence of VSP activity, mutations affecting the second C at Dcm sites (e.g., T:G mismatches) will not be corrected, thus decreasing the frequency of such sites. This prediction was fully confirmed: The frequency of Dcm sites turned out to be 0.149% in the CTAG-rich

Table 2. Comparisons of the averaged tetranucleotide frequencies in the CTAG-rich and CTAG-poor samples[a]

	CTAG-poor sample	CTAG-rich sample	*P*
CTAG	0.000	0.178	
Group I tetramers			
CTTG	0.221	0.312	*
CCTA	0.089	0.162	**
CCAA	0.285	0.300	NS
TTGG	0.298	0.366	*
TAGG	0.091	0.154	*
CAAG	0.229	0.325	**
Group II tetramers			
CCTG	0.551	0.398	**
CCAG	0.456	0.385	NS
CTGG	0.993	0.563	**
CAGG	0.543	0.383	**

[a] P = statistical significance of t-tests. (*) $P < 0.05$; (**) $P < 0.001$; NS = nonsignificant.

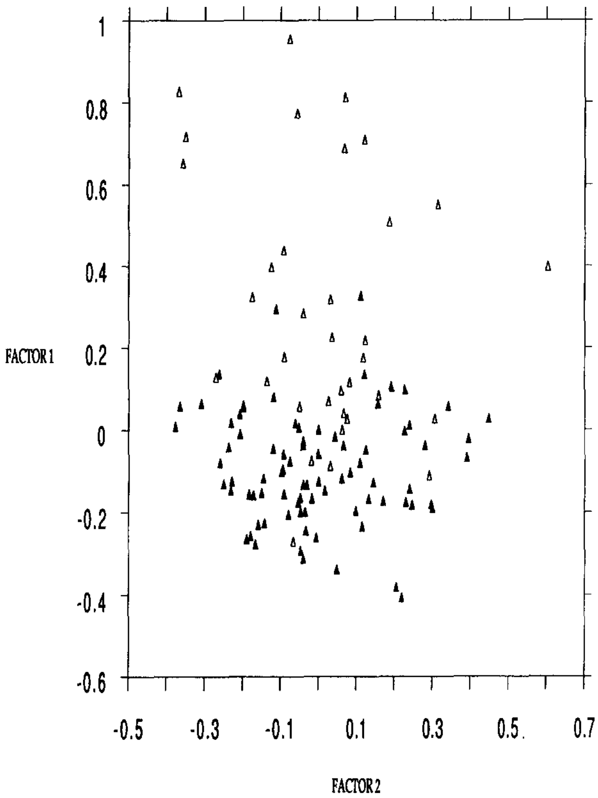

Fig. 2. Correspondence analysis on tetranucleotide frequencies (except CTAG) in the CTAG-rich (△) and CTAG-poor sample (▲). Factor 1 denotes the 28.63% variability and factor 2 the 15.80% in tetranucleotide frequencies found for these samples.

sample and 0.252% in the CTAG-poor sample. At this point it is important to recall that Dcm site distribution does not exhibit any identifiable regular pattern on the *E. coli* chromosome (Gómez-Eichelmann and Ramírez-Santos 1993).

Table 3. Similarity values of contrast vocabularies (above diagonal) and vocabulary correlation values (below diagonal) for dinucleotides (up left), trinucleotides (up right), tetranucleotides (down left), and pentanucleotides (down right)[a]

	CTAG-rich	CTAG-poor	Control	Outgroup
CTAG-rich	—	0.72	0.70	0.44
CTAG-poor	0.88/0.88 0.68/0.44	—	0.90	0.34
Control	0.86/0.88 0.69/0.37	0.99/0.98 0.91/0.73	—	0.35
Outgroup	0.70/0.27 0.68/0.13	0.48/0.35 0.47/0.04	0.47/0.38 0.50/0.04	—

[a] The outgroup is represented by an assembly of 65 entries (162,672 bases) containing complete coding sequences of *S. cerevisiae* retrieved from the Genbank. In order to substantiate the differences between the CTAG-rich and poor samples, we have added another sample called "control" (56 entries, none overlapping with the CTAG-poor sample). This sample was selected on the same criteria used to select the CTAG-poor sample.

Contrast Vocabularies

The contrast vocabularies for each sample were calculated by the method of Trifonov and co-workers (Pietrokovski et al. 1990). This method is especially suitable for comparing two unrelated DNA sequences without a drastic dependence on sequence length. The contrast vocabularies take into account the differences between observed and expected frequencies of di-, tri-, tetra-, and pentanucleotides; these differences are called "contrast values." The expected frequency of each oligomer is computed from Markov chains. For every oligonucleotide length a correlation coefficient is obtained; a similarity value between the two vocabularies under study is then calculated as the average value of the correlation coefficients. This method has proved to be useful to describe different kinds of sequences and to estimate the relations between them in terms of "relevant" oligonucleotides or "words" (Pietrokovski and Trifonov 1992).

Correlation coefficients for di-, tri-, tetra-, and pentanucleotides and their averages (similarity value) are given in Table 3. It can be seen that the higher differences between the CTAG-poor and the CTAG-rich samples correspond to tetra- and pentanucleotides, thus suggesting a major role for VSP activity; however, differences affecting shorter oligonucleotides are also observed. As a reference we have compared the CTAG-rich and CTAG-poor samples to an outgroup represented by DNA sequences from *Saccharomyces cerevisiae*, where no homolog of the VSP repair system has been detected (Bhagwat and McClelland 1992). The vocabulary of the outgroup is more similar to that of the CTAG-rich sample, as expected in the absence of VSP activity.

Nucleotide Composition and G+C Content

The average nucleotide frequencies by codon position in the sense strands of protein-coding genes in the two sam-

Table 4. Nucleotide composition of the protein coding genes from CTAG-rich and from CTAG-poor samples by codon position

		Codon position			
		I	II	III	Total
A	rich	0.30	0.33	0.26	0.30
	poor	0.24	0.28	0.17	0.23
C	rich	0.20	0.21	0.19	0.20
	poor	0.24	0.23	0.28	0.25
G	rich	0.30	0.17	0.22	0.23
	poor	0.36	0.19	0.30	0.28
T	rich	0.19	0.29	0.33	0.27
	poor	0.15	0.30	0.25	0.24
G+C	rich	0.50	0.38	0.41	0.43
	poor	0.60	0.42	0.58	0.53

ples are given in Table 4. The frequency of C and G nucleotides is always lower in the CTAG-rich sample. As expected, most of the differences in base composition are accounted for by third codon positions and to a gradually lesser extent by first and second positions. The latter two will result in amino acid frequency differences at the protein sequence level. (See below). These results are one expected consequence of the mutagenic role of VSP, which should promote a number of T-to-C transitions, thus increasing G+C content. Thus it is not surprising that G+C is consequently higher in the CTAG-poor sample (53.26%) than in the CTAG-rich sample (43.22%).

Codon Usage and Protein Composition

Differences in nucleotide composition between the two samples must be reflected (1) in the choice between synonymous codons and (2) in the overall amino acid composition of the proteins encoded. To illustrate the effect of VSP activity on these sequence features, let us consider a transition substituting CTTG to CCTG; depending on the reading frame we can expect the following effects:

1. CTT (leu) → CCT (pro) amino acid substitution
2. TTG (leu) → CTG (leu) silent substitution (first base)
3. NCT → NCC silent substitution (third base)

VSP can actually affect all codons except two synonymous codon groups (the histidine duet CAY and the arginine quartet CGN; data not shown). A general prediction is that the frequency of C-ending codons should be higher in the CTAG-poor sample than in the CTAG-rich sample, which the opposite should occur with T-ending codons. This prediction was fully confirmed, as seen from the nucleotide compositions at third codon sites shown in Table 4. A related aspect is the differential usage of the quartets and duets encoding leucine (CUN and UUR) and arginine (CGN and AGR). The number of arginine residues encoded by the duet in the CTAG-poor sample is one-sixth that found in the CTAG-rich sample. In the case of leucine, the equivalent figure is about one-half (Table 5).

Table 5. Percentage of leucine and arginine residues encoded by the quartet and duet codon groups in the samples studied

	CTAG-poor	CTAG-rich
Leu4	78.64	61.02
Leu2	21.36	38.98
Arg4	95.96	71.56
Arg2	04.04	28.44

Another interesting difference in codon usage concerns termination codons. TAG is less frequently used in the CTAG-poor sample (4% of stop codons) than in the CTAG-rich sample (20%) (Table 6).

Codon Adaptation Index (CAI)

The codon usage tables obtained for the CTAG-rich and the CTAG-poor samples were compared to the codon usage of highly and lowly expressed genes of *E. coli* (Sharp et al. 1988). These comparisons showed that the codon choices of genes from the CTAG-rich sample were more similar to those of lowly expressed genes than the codon choices from the CTAG-poor sample. Since this result suggested a possible link between tetranucleotide composition and expression level, we represent the distribution of CTAG-rich and CTAG-poor genes according to their Codon Adaptation Index, CAI (Sharp and Li (1987). Figure 3 shows that the distribution found for the CTAG-rich sample is narrower and more left-skewed than that of the CTAG-poor sample; their respective average CAIs are 0.167 and 0.314. Furthermore, the correlation coefficients between the CAI value and the frequency of group-I tetramers and of group-II tetramers, computed over all the genes analyzed, are always positive between CAI and group-II tetramers ($r = 0.44$, $P < 0.001$) and negative between CAI and group-I tetramers ($r = -0.45$, $P < 0.001$). Thus, the overall conclusion is that a lower level of expression can be expected for genes included in the CTAG-rich sample.

Discussion

The existence of intragenomic compositional heterogeneity in prokaryotes was anticipated by Bernardi et al. (1985) and later confirmed by Nomura et al. (1987), Médigue et al. (1991a), and D'Onofrio and Bernardi (1992). It is also well known that *E. coli* genes can be split into two groups according to their codon usage, which is strongly correlated with their level of expres-

Table 6. Amino acid composition per 1,000 codons in the CTAG-rich and CTAG-poor samples

	CTAG-poor	CTAG-rich
Total	44,905	25,987
Amino acids		
Gly	78.59	58.79
Glu	58.19	57.37
Asp	50.24	55.99
Val	71.73	55.95
Ala	99.17	71.58
Arg	56.23	59.26
Ser	55.39	73.07
Lys	43.04	64.46
Asn	37.08	51.49
Met	29.95	19.32
Ile	58.25	70.18
Thr	50.82	53.68
Trp	15.39	13.31
Cys	11.71	11.66
Tyr	28.73	39.75
Leu	104.07	101.41
Phe	39.18	41.52
Gln	42.75	39.02
His	21.80	23.44
Pro	44.22	35.90
TGA	0.94	0.35
TAG	0.11	0.54
TAA	1.80	1.31

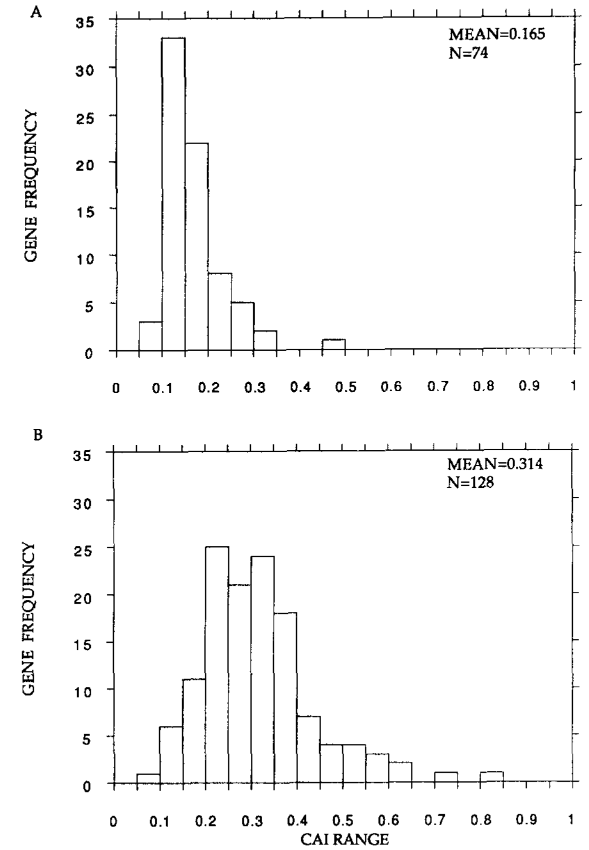

Fig. 3. Distribution of CAI values for the CTAG-rich (**A**) and CTAG-poor (**B**) samples. The CAI average value and the number of genes used are indicated in the plots.

sion (Gouy and Gautier 1982). More recently, codon usage in *E. coli* genes has been reanalyzed (Médigue et al. 1991b). These authors have defined a third class in addition to those of highly and lowly expressed genes; this class III includes genes that encode a variety of products, showing codon choices that do not reflect the average distribution of specific tRNA availability; moreover, class II coding sequences behave somewhat differently from core-metabolism genes in counterselection for CTAG and other palindromic sites. These observations led to the suggestion that they may have been acquired by horizontal transfer.

Our classification of *E. coli* genes was based on the frequency of the tetranucleotide CTAG. The resulting classes (CTAG-rich and CTAG-poor) differ in tetramer composition, codon usage, amino acid composition, frequency of Dcm sites, and contrast vocabularies. From the point of view of codon usage, genes included in class III of Médigue et al. (1991b) mostly belong to our CTAG-rich sample.

Codon adaptation indexes indicated that the codon choices of the CTAG-rich sample were similar to those of poorly expressed genes, suggesting a correlation between tetranucleotide composition and gene expression. For the purpose of this paper, the most relevant observation was that a compositionally well-characterized group of *E. coli* genes exists where traces of VSP repair activity are scarce. Somehow surprisingly, VSP activity seems to be hindered in genes that are not actively expressed. Thus VSP may act as a source of intragenomic heterogeneity in *E. coli* by inducing C-T transitions, but certain poorly expressed genes may be immune to this effect. The molecular mechanisms that might prevent VSP repair from operating on poorly expressed genes are presently unknown.

The existence of a VSP-induced compositional bias within the *E. coli* genome does not exclude the possibility that certain compositional differences may be explained by horizontal transfer, as suggested by Médigue et al. (1991b). However, many cases exist where horizontal transfer seems unlikely. An example may be that of the *dsdC* (included in Table 1) and *dsdA* genes of the serine deaminase operon. These genes are clustered in the same chromosomal region although they undergo divergent transcription (McFall 1987). The frequency of group-I tetramers (including CTAG) is 1.73% in the highly transcribed, structural gene *dsdA* and 2.53% in the lowly transcribed, regulator gene *dsdC*. Other examples of lowly expressed genes contained in the CTAG-rich sample are *appY* (entry name ECOAPPYAA; Atlung et al. 1989) and *lit* (entry name ECOLIT; Kao and Snyder 1988).

A particularly interesting example, although not included in Table 1 because it does not reach the limit of CTAG frequency required in our sample, is that of the macromolecular synthesis (MMS) operon (reviewed by

Lupski and Godson 1984). This operon contains *E. coli* genes that are involved in the initiation of translation (*rpsU,* encoding ribosomal protein S21), DNA replication (*dnaG,* encoding a primase), and transcription (*rpoD,* encoding sigma-70). The MMS operon appears to be under very complex regulatory control, including regulation at the gene level by strategically positioning transcriptional and translational control signals. Thus, the amount of *dnaG* primase is maintained at a very low level in the cell due to the presence of an RNA polymerase transcription terminator between the *rpsU* and *dnaG* genes; moreover, *dnaG* contains an unusual ribosome binding site and shows a poorly adapted codon usage. Interestingly, the frequency of group-I tetramers in the *dnaG* region (which contains the unique CTAG of the operon) is much higher (1.78%) than in the flanking genes (0.47% at *rpsU* and 0.98% at *rpoD*). The corresponding CAI values are 0.20 (*dnaG*), 0.71 (*rpsU*), and 0.53 (*rpoD*).

In conclusion, the main finding of this paper is that DNA sequences in the CTAG-rich sample, some of which apparently arrived in the *E. coli* genome by horizontal transfer, have, in addition to poorly adapted codon usage, the signs of not having been exposed to VSP repair. Although the link between VSP repair and gene expression level is not proved in this paper, it can be considered in the light of a ''new way of thinking about genome evolution, involving regulation of mutational input *via* patterns of damage, repair, recombination, replication and transcriptional activity'' (Holmquist and Filipski 1994).

References

Atlung T, Nielsen A, Hansen FG (1989) Isolation, characterization and nucleotide sequence of *appY,* a regulatory gene for growth-phase-dependent gene expression in *Escherichia coli.* J Bacteriol 171: 1683–1691

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M, Rodier F (1985) The mosaic genome of warm-blooded vertebrates. Science 228:953–958

Bhagwat AS, McClelland M (1992) DNA mismatch correction by very short patch repair may have altered the abundance of oligonucleotides in the *E. coli* genome. Nucleic Acids Res 20:1663–1668

Burge C, Campbell AM, Karlin S (1992) Over- and under-representation of short oligonucleotides in DNA sequences. Proc Natl Acad Sci USA 89:1358–1362

Devereux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 12: 387–395

Dixon WJ, Brown MB (1979) BMDP-79 biomedical computer programs P series. University of California Press, Berkeley

D'Onofrio G, Bernardi G (1992) A universal compositional correlation among codon positions. Gene 110:81–88

Gómez-Eichelmann MC, Ramírez-Santos J (1993) Methylated cytosine at Dcm (CC-A/T-GG) sites in *Escherichia coli:* possible function and evolutionary implications. J Mol Evol 37:11–24

Gouy M, Gautier C (1982) Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 10:7055–7074

Holmquist GP, Filipski J (1994) Organization of mutations along the genome: a prime determinant of genome evolution. TREE 9:65–69

Jukes TH, Bhushan V (1986) Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. J Mol Evol 24:39–44

Kao C, Snyder L (1988) The *lit* gene product which blocks bacteriophage T4 late gene expression is a membrane protein encoded by a cryptic DNA element, e14. J Bacteriol 170:2056–2062

Lieb M (1991) Spontaneous mutation at a 5-methylcytosine hotspot is prevented by very short patch (VSP) mismatch repair. Genetics 128:23–27

Lébart L, Fenelon JP (1975) Statistique et informatique appliquées. Dunod, Paris

Lupski JR, Godson GN (1984) The *rpsU-dnaG-rpoD* macromolecular synthesis operon of *E. coli.* Cell 39:251–252

McFall E (1987) The D-serine deaminase operon. In: Neidhardt FC (ed) *Escherichia coli* and *Salmonella typhimurium:* cellular and molecular biology. American Society of Microbiology, Washington, DC, p 1520

Médigue C, Viari A, Hénaut A, Danchin A (1991a) *Escherichia coli* molecular genetic map (1500 kbp): update II. Mol Microbiol 5: 2629–2640

Médigue C, Rouxel T, Vigier P, Hénaut A, Danchin A (1991b) Evidence for horizontal gene transfer in *Escherichia coli* speciation. J Mol Biol 222:851–856

Merkl R, Kröger M, Rice P, Fritz HJ (1992) Statistical evaluation and biological interpretation of non-random abundance in the *E. coli* K-12 genome of tetra- and pentanucleotide sequences related to VSP DNA mismatch repair. Nucleic Acids Res 20:1657–1662

Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. Proc Natl Acad Sci USA 84:166–169

Nomura M, Sor F, Yamagishi M, Lawson M (1987) Heterogeneity of GC content within a single bacterial genome and its implications for evolution. Cold Spring Harbor Symp Quant Biol 52:658–663

Phillips GJ, Arnold J, Ivarie R (1987a) Mono- through hexanucleotide composition of the *Escherichia coli* genome: a Markov chain analysis. Nucleic Acids Res 15:2611–2626

Phillips GJ, Arnold J, Ivarie R (1987b) The effect of codon usage on the oligonucleotide composition of the *E. coli* genome and identification of over- and underrepresented sequences by Markov chain analysis. Nucleic Acids Res 15:2627–2638

Pietrokovski S, Hirshon J, Trifonov EN (1990) Linguistic measure of taxonomic and functional relatedness of nucleotide sequences. J Biomol Struct Dyn 7:1251–1268

Pietrokovski S, Trifonov EN (1992) Imported sequences in the mitochondrial yeast genome identified by nucleotide linguistics. Gene 122:129–137

Rolfe R, Meselson M (1959) The relative homogeneity of microbial DNA. Proc Natl Acad Sci USA 45:1039–1043

Sharp PM, Li WH (1987) The codon adaptation index: a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295

Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F (1988) Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens;* a review of the considerable within-species diversity. Nucleic Acids Res 16:8207–8211

Sharp PM (1990) Processes of genome evolution reflected by base frequency differences among *Serratia marcescens* genes. Mol Microbiol 4:119–122

Sharp PM, Lloyd AT (1993) Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure. Nucleic Acids Res 21:179–183

Sueoka N (1959) A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. Proc Natl Acad Sci USA 45:1480–1490

Sueoka N (1962) On the genetic basis of variation and heterogeneity of DNA base composition. Proc Natl Acad Sci USA 48:582–592

Sueoka N (1992) Directional mutation pressure, selection constraints, and genetic equilibria. J Mol Evol 34:95–114