Asymmetric Substitution Patterns in the Two DNA Strands of Bacteria

J.R. Lobry

Laboratoire de Biométrie, Université Claude Bernard

Analyses of the genomes of three prokaryotes, *Escherichia coli, Bacillus subtilis,* and *Haemophilus influenzae,* revealed a new type of genomic compartmentalization of base frequencies. There was a departure from intrastrand equifrequency between A and T or between C and G, showing that the substitution patterns of the two strands of DNA were asymmetric. The positions of the boundaries between these compartments were found to coincide with the origin and terminus of chromosome replication, and there were more A-T and C-G deviations in intergenic regions and third codon positions, suggesting that a mutational bias was responsible for this asymmetry. The strand asymmetry was found to be due to a difference in base compositions of transcripts in the leading and lagging strands. This difference is sufficient to affect codon usage, but it is small compared to the effects of gene expressivity and amino-acid composition.

Introduction

The differences in the way leading and lagging strands of DNA are replicated (Nossal 1983; Marians 1992; Baker and Wickner 1992), at least in vitro, could result in strand-dependent mutation patterns. In the absence of any selection bias between the strands, differences in patterns of replication error lead to differences in substitution patterns. The persistence of asymmetric substitution patterns should yield equilibrium frequencies for the four nucleotides that differ from those expected under no–strand-bias conditions. This difference may be the basis for a very simple test for asymmetry in substitution patterns.

The null hypothesis for the detection of unequal substitution patterns between the two strands of DNA, the no–strand-bias condition (Sueoka 1995), states that there is no bias between the two DNA strands for either mutation or selection. Under no–strand-bias conditions, the equilibrium point is such that the base frequencies in each strand are always [A]=[T] and [C]=[G], regardless of the initial state of the DNA sequence, or the details of the substitution patterns (Sueoka 1995, Lobry 1995). Any significant deviation from the intrastrand [A]=[T] or [C]=[G] relationships is an indication that there is asymmetry in the substitution patterns of the two DNA strands.

Previous attempts to detect unequal mutation patterns in the two strands of DNA from the β-globin region of primates (Wu and Maeda 1987) were unsuccessful (Wu 1991; Bulmer 1991), mainly because the location of the origins and termini of replication in the data sets analyzed were not known. Significant violations of the [A]=[T] or [C]=[G] relationships have been reported for the human fetal globin gene region, the mammalian viruses SV40 and polyomavirus strain A-2 (Smithies et al. 1981), but without biological interpretation. The deviation from the [C]=[G] rule for SV40 was interpreted (Filipski 1990) as evidence for asymmetric mutation pressure because of a polarity switch at the origin of replication. However, the SV40 genome is very rich in coding sequences, and asymmetric selective pressure was not ruled out. Mitochondrial DNA from *Homo sapiens, Gallus gallus, Asterina pectinifera,* and *Ascaris suum* were found to have intergenic A + C or G + T contents on one strand that were statistically different from 50% (Jermiin et al. 1995). Since the intrastrand [A]=[T] and [C]=[G] rules imply that A + C = G + T = 50%, the null hypothesis for detecting inequality in substitution patterns must also have been rejected.

The very long contiguous stretches of DNA sequence from three bacterial species, *Escherichia coli, Haemophilus influenzae,* and *Bacillus subtilis,* are well suited to an examination of strand asymmetry in substitution patterns.

Material and Methods

Variations in base frequencies along sequences were studied using a nonoverlapping moving window and by plotting three indices of base frequency: G + C content, (G + C)/(A + T + C + G), deviation from [A]=[T], (A − T)/(A + T), and deviation from [C]=[G], (C − G)/(C + G). These three indices are, by design, pairwise-independent and summarize relative base frequencies without loss of information. Because of their very low frequency, ambiguous bases were not taken into account. The standard deviation (SD) for the three indices are given by

$$\mathrm{SD}[(G + C)/(A + T + C + G)] = \frac{1}{N}\sqrt{\frac{SW}{N}} \quad (1)$$

$$\mathrm{SD}[(A - T)/(A + T)] = \frac{2}{W}\sqrt{\frac{AT}{W}} \quad (2)$$

$$\mathrm{SD}[(C - G)/(C + G)] = \frac{2}{S}\sqrt{\frac{CG}{S}} \quad (3)$$

where W = A + T, S = C + G, and N = A + T + C + G.

In the following, the total number of bases is so large that the normal distribution approximation can be used.

Key words: DNA evolution model, DNA replication and repair, substitution patterns, substitution asymmetry, mutational bias, chirochore.

Address for correspondence and reprints: J.R. Lobry, Laboratoire de Biométrie, CNRS URA 2055, Université Claude Bernard, 43 Bd. du 11 Novembre 1918, F-69622 Villeurbanne cedex, France. E-mail: lobry@biomserv.univ-lyon1.fr.

Mol. Biol. Evol. 13(5):660–665. 1996

Therefore, 95% confidence intervals for index values were computed as means ± 1.96 SDs.

Coding sequences were analysed by computing the deviations from the base composition of the strand corresponding to the mRNA, not for the transcription template strand. Coding sequences of less than 100 codons were excluded to minimize the influence of stochastic variations in the base composition of short sequences. Standard deviations were deduced from the sample variances, and confidence intervals for mean values in coding sequence groups were computed to provide comparisons that took into account between-gene variability.

Escherichia coli

The sequence examined was a 1,616,174-base fragment from 67.4 min to 4.1 min on the genetic map of the *E. coli* chromosome (Sofia et al. 1994; Burland et al. 1993; Daniels et al. 1992; Plunkett et al. 1993; Blattner et al. 1993; Yura et al. 1992). The sequence is available in the EMBL/DDBJ/GenBank database as nine overlapping segments (U18997, U00039, L10328, M87049, L19201, U00006, U14003, D10483, D26562) and as a single contiguous sequence at the URL: ftp://ecoliftp.genetics.wisc.edu/pub/sequence/m16j.seq. The fragment represents 34% of the *E. coli* chromosome; it has been completely and continuously sequenced, except for 407 ambiguous bases (0.025% of total data). The fragment includes the origin of replication between positions 706,480 and 706,711. The strand analysed here is the 5′→3′ strand clockwise on the genetic map. The three indices were computed with a window size of 10 kilobases (kb).

Haemophilus influenzae

The sequence was a 1,830,137-base fragment representing the entire *H. influenzae* chromosome that has been completely and continuously sequenced, except for 47 ambiguous bases (0.003% of total data). The fragment includes a putative replication origin near base 603,000 and a putative replication terminus near base 1,518,000 (Fleischmann et al. 1995). The sequence is available in the EMBL/DDBJ/GenBank database as 163 overlapping segments (L42023) and as a single contiguous sequence at the URL: ftp://ftp.tigr.org/pub/h_influenzae/GHI.1con.Z. The strand analyzed here was the same as in the above-mentioned contiguous sequence; in this strand *rrnA* is oriented 5′→3′. A window size of 10 kb was used.

Bacillus subtilis

The sequence was a 193,394-base fragment representing 5% of the *B. subtilis* chromosome that has been completely and continuously sequenced (Ogasawara et al. 1994; Jeong et al. 1993; Boor et al. 1995), except for five ambiguous bases (0.003% of total data). The fragment includes the replication origin near base 65,500. The sequence is available in the EMBL/DDBJ/GenBank database as five overlapping segments (D26185, D13303, D50303, L24376, L43593) and as a single contiguous sequence (BS0310) in the NRSub nonredundant database (Perriere et al. 1996) at the URL: http://ddbjs4h.genes.nig.ac.jp/cgi-bin/acnuc-search-ac?query=BS0310. The strand analyzed here was the 5′→3′ strand clockwise on the genetic map. A window size of 1 kb was used.

Results

Escherichia coli

There is no spatial structure for the G + C content in *E. coli* (fig. 1*a*), i.e., no isochore-like structure (Bernardi 1989). The slight enrichment in G + C content reported by Deschavanne and Filipski (1995) for genes close to the replication origin did not appear to be a strong feature on this scale. There was a small A-T deviation and a striking systematic deviation from [C]=[G] that changed its sign precisely at the origin of chromosome replication. The switch in C-G deviation was very significant (table 1), and the switch in A-T deviation was also statistically significant, but the difference was smaller than for the C-G switch and difficult to appreciate in figure 1. Hence, there is an inequality in the substitution patterns of the two strands of DNA that switch polarity at the replication origin.

There were increases in the switch intensities for both the A-T and C-G deviations when all the coding sequences and genes for rRNAs and tRNAs were removed (table 1). As the selective pressure should be lower in intergenic regions, this suggests that the substitution pattern asymmetry is due to mutation, but there is some doubt as to whether these intergenic regions are completely selection-free. When only nucleotides in coding sequences were considered, the switch intensity was lower and concentrated in third and first codon positions (table 1). This is also in favor of mutational bias.

Fifty-four percent of the positions in coding sequences were in the leading strand, i.e., transcribed divergently from the origin. This may introduce a selective bias and also a mutation bias because of differences in DNA repair in the transcribed and the nontranscribed strands. The effect of this uneven distribution was removed by computing the A-T and C-G deviations in leading and lagging coding sequences separately. The A-T and C-G deviations were found to be different for leading and lagging coding sequences, and the difference was higher in third and first codon positions (table 2).

If the differences in the A-T and C-G deviations between lagging and leading coding sequences were due to a selective effect on codon usage or amino-acid composition, the same phenomenon should not be found in untranslated transcribed regions. The positions of these regions are less well defined than for coding sequences. For this study, a nucleotide was assumed to be in a untranslated transcribed region if it was surrounded by two coding sequences with the same orientation and separated by less than 50 nucleotides (1,923 nucleotides were included in the lagging group and 2,877 in the leading group). The A-T deviation was found to be +12.5% ± 5.0 in the leading group and +11.8% ± 6.0 in lagging group, and the C-G deviation was found to be −15.4% ± 5.2 in the leading group and −9.5% ±

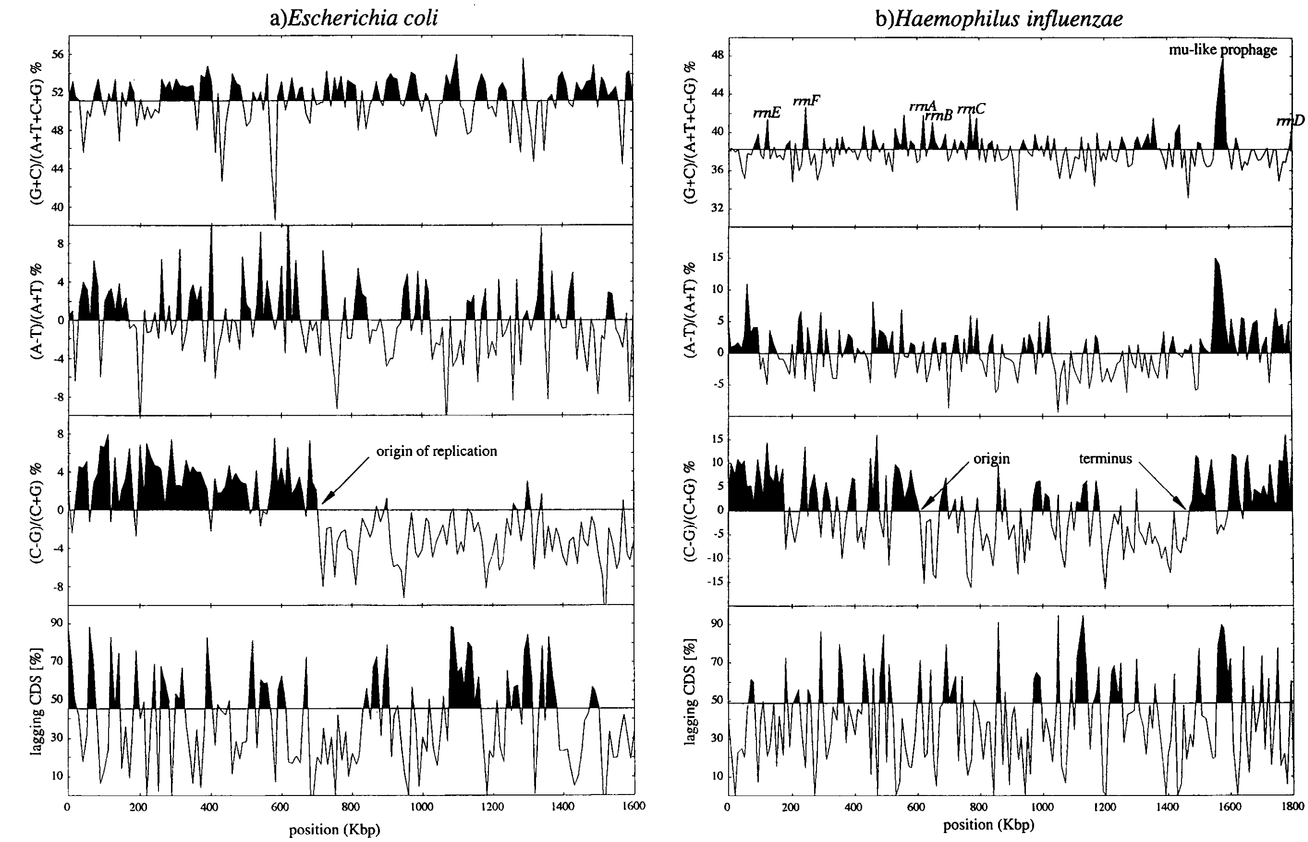
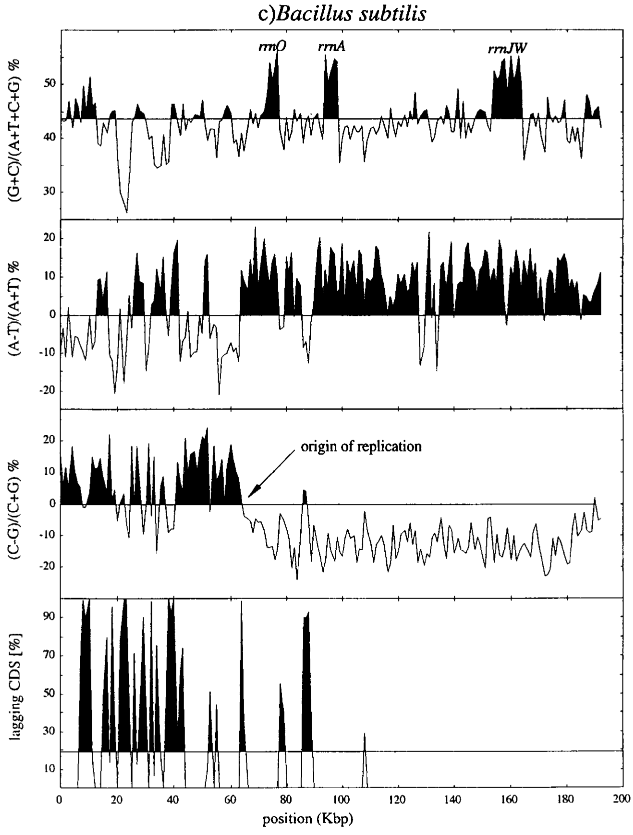


FIG. 1.—Variations in three nucleotide base frequency indices and frequency of positions in lagging coding sequences along (*a*) a fragment of the *E. coli* chromosome, (*b*) the *H. influenzae* chromosome, and (*c*) a fragment of the *B. subtilis* chromosome. The horizontal line for A-T and C-G deviations is drawn at the value zero, corresponding to what is expected under no–strand-bias conditions. The horizontal line for the G + C content and for the frequency of position in lagging coding sequences is drawn at the mean value. The A-T and C-G deviations are usually dominated by the deviation resulting from leading coding sequences (table 2). When there is a local abundance of lagging coding sequences, the general trend is reversed, yielding reverse asymmetry spikes.



6.6 in the lagging group. The number of remaining sites in these regions was too small for the difference to be significant, but the trend was consistent with the findings for coding sequences.

The relative importance of the difference in the compositions of leading and lagging coding sequences must be placed in context. Correspondence analysis (Hill 1974) of codon usage in the 1333 coding sequences with over 100 codons showed that the two first factors (25.4% of initial variability) were gene expressivity, the third factor separated the integral membrane protein group (6.7% variabilty), and the fourth factor (3.75% variability) was the difference between leading and lagging sequences. The deviation is small on the scale of the gene, so that a C-G deviation of −5% on a 1000-nt gene, starting from [C]=[G], is produced by only two or three C→G events per 100 C or G nucleotides, corresponding to about 12 substitutions in the gene.

Haemophilus influenzae

The *H. influenzae* genome (fig. 1*b*) contains no isochore-like structure, but there is some local enrichment

Table 1
Switches in the C-G and A-T Deviations at the Origin of Replication in Three Bacterial Genomes

	(C − G)/(C + G) (%)			(A − T)/(A + T) (%)		
	Before (5′)	After (3′)	Δ	Before (5′)	After (3′)	Δ
Escherichia coli:						
Global	+3.0 ± 0.3	−3.3 ± 0.3	6.6 ± 0.6	+0.8 ± 0.3	−0.7 ± 0.3	1.5 ± 0.6
Intergenic	+4.1 ± 1.1	−3.8 ± 0.8	7.9 ± 1.9	+1.3 ± 0.9	−1.3 ± 0.7	1.6 ± 1.6
CDS	+2.8 ± 0.3	−2.9 ± 0.3	5.7 ± 0.6	+0.8 ± 0.4	−0.8 ± 0.3	1.6 ± 0.7
Codon pos. I	+3.6 ± 0.6	−3.2 ± 0.5	6.8 ± 1.1	−1.5 ± 0.7	+0.4 ± 0.6	1.9 ± 1.3
Codon pos. II	−0.2 ± 0.7	+0.1 ± 0.6	0.3 ± 1.3	−0.1 ± 0.6	−0.1 ± 0.5	0.0 ± 1.1
Codon pos. III	+4.1 ± 0.6	−4.6 ± 0.5	8.7 ± 1.1	+4.3 ± 0.7	−3.2 ± 0.6	7.5 ± 1.3
Haemophilus influenzae:						
Global	+4.1 ± 0.3	−3.1 ± 0.3	7.2 ± 0.6	+1.6 ± 0.3	−1.0 ± 0.3	2.6 ± 0.6
Intergenic	+4.7 ± 0.9	−5.0 ± 0.9	9.7 ± 1.8	+1.4 ± 0.6	−1.9 ± 0.6	3.3 ± 1.2
CDS	+3.6 ± 0.4	−2.4 ± 0.4	6.0 ± 0.8	+1.9 ± 0.3	−1.0 ± 0.3	2.9 ± 0.6
Codon pos. I	+3.7 ± 0.5	−2.3 ± 0.6	6.0 ± 1.1	+0.8 ± 0.6	−1.2 ± 0.6	2.0 ± 1.2
Codon pos. II	−0.9 ± 0.7	0.0 ± 0.6	0.9 ± 1.3	+1.2 ± 0.5	−0.6 ± 0.5	1.8 ± 1.0
Codon pos. III	+9.0 ± 0.7	−5.8 ± 0.7	14.8 ± 1.4	+3.7 ± 0.5	−1.3 ± 0.5	5.0 ± 1.0
Bacillus subtilis:						
Global	+7.8 ± 1.2	−12.1 ± 0.8	19.9 ± 2.0	−2.3 ± 1.0	+8.9 ± 0.7	11.2 ± 1.7
Intergenic	+9.8 ± 2.8	−11.1 ± 2.7	20.9 ± 5.5	−4.3 ± 2.2	+1.9 ± 2.1	6.2 ± 4.3
CDS	+7.3 ± 1.3	−11.4 ± 1.0	18.7 ± 2.3	−1.8 ± 1.1	+9.6 ± 0.8	11.4 ± 1.9
Codon pos. I	+8.8 ± 2.0	−29.0 ± 1.4	37.8 ± 3.4	−6.3 ± 2.1	+29.3 ± 1.6	35.6 ± 3.7
Codon pos. II	+0.1 ± 2.5	+11.8 ± 1.9	11.9 ± 4.4	−3.6 ± 1.8	+6.1 ± 1.4	9.7 ± 3.2
Codon pos. III	+11.6 ± 2.3	−7.6 ± 1.7	19.2 ± 4.0	+4.1 ± 2.0	−1.6 ± 1.4	5.7 ± 3.4

NOTE.—All deviations were computed from the base compositions in the DNA strands defined in Material and Methods. Intergenic: all coding sequences and genes for rRNAs and tRNA were removed. CDS: only positions in coding sequences were used. Codon position: only bases in the given codon position were taken into account. The importance of the switches is given by Δ, the distance between the values before and after the replication origin.

in G + C in ribosomal RNA operons, and a cryptic mu-like prophage (Fleischman et al. 1995). There was a significant switch in the deviations from [C]=[G] and [A]=[T] at the origin and terminus of replication (table 1). This indicated that there is also an asymmetry in substitution patterns in *H. influenzae.*

When all coding sequences and genes for rRNAs and tRNAs were removed, there was an increase in switch intensity for both the A-T and C-G deviations (table 1). When only nucleotides in coding sequences were considered, the switch intensity was lowered and concentrated in third and first codon positions (table 1). These findings suggest that the asymmetry in substitution patterns is due to mutational bias.

There was an uneven distribution of coding sequences, with 52% of positions in leading coding sequences. The A-T and C-G deviations were found to be different for leading and lagging coding sequences, and the difference was higher in third codon position (table 2).

The difference was still visible in untranslated transcribed regions (2,341 nucleotides in lagging sequences and 3,621 in leading sequences were examined), with an A-T deviation of +10.5% ± 4.9 in the lagging group and +15.1% ± 1.9 in the leading group, and a C-G deviation of −13.4% ± 7.0 in the lagging group and −27.2% ± 5.7 in the leading group.

Correspondence analysis of codon usage in the 1505 sequences of more than 100 codons showed that the first factor (12% of initial variability) was gene expressivity, the second and third factors (13.3% variability) discriminated integral membrane protein coding sequences and the mu-like G + C rich coding sequences, while the fourth factor (4.8% variability) indicated a difference between lagging and leading coding sequences.

Bacillus subtilis

B. subtilis (fig. 1*c*) contained no isochore-like structure, only local G + C enrichment in ribosomal operons. There were significant switches in the A-T and C-G deviations at the replication origin (table 1), showing that there was an asymmetric substitution pattern.

The number of nucleotides in the analyzed sequence was too small for a detailed analysis. When all coding sequences and genes for rRNAs and tRNAs were removed, there were only 26,916 nucleotides left, so that the result for intergenic regions in table 1 should be regarded with caution. The results for coding sequence regions (table 1) were not straigthforward because (1) there was a highly uneven distribution, with 82% of positions in leading coding sequences, (2) most of the lagging coding sequences were before the replication origin, (3) integral membrane protein genes were concentrated in the lagging coding sequence: they accounted for 26% of the lagging group and only 6.5% of the leading group.

When leading and lagging coding sequences were analyzed separately, the A-T and C-G deviations were found to be different (table 2); but the difference in A-

Table 2
Differences in the C-G and A-T Deviations in Leading and Lagging Coding Sequences in Three Bacterial Genomes

	(C − G)/(C + G) (%)			(A − T)/(A + T) (%)		
	Lagging	Leading	Δ	Lagging	Leading	Δ
Escherichia coli:						
All positions	−2.8 ± 0.5	−7.7 ± 0.4	4.9 ± 0.9	+1.1 ± 0.9	−0.7 ± 0.8	1.8 ± 1.7
Codon pos. I	−15.9 ± 1.1	−19.7 ± 0.9	3.8 ± 2.0	+24.8 ± 1.0	+23.5 ± 1.1	1.3 ± 2.1
Codon pos. II ...	+11.3 ± 1.0	+10.0 ± 1.0	1.3 ± 2.0	−3.7 ± 1.5	−3.1 ± 1.3	0.6 ± 2.8
Codon pos. III ...	+0.6 ± 1.1	−7.7 ± 1.0	8.3 ± 2.1	−15.3 ± 1.0	−19.5 ± 1.0	4.2 ± 2.0
Haemophilus influenzae:						
All positions	−5.9 ± 0.6	−11.4 ± 0.5	5.5 ± 1.1	+4.1 ± 0.7	+1.2 ± 0.7	2.9 ± 1.4
Codon pos. I	−26.7 ± 1.0	−30.3 ± 1.0	3.6 ± 2.0	+17.4 ± 0.9	+14.8 ± 0.9	2.6 ± 1.8
Codon pos. II ...	+14.8 ± 1.0	+14.4 ± 0.9	0.4 ± 1.9	+2.7 ± 1.3	+1.2 ± 1.1	1.5 ± 2.4
Codon pos. III ...	+4.4 ± 1.5	−10.1 ± 1.6	14.5 ± 3.1	−3.9 ± 0.8	−8.3 ± 0.7	4.4 ± 1.5
Bacillus subtilis:						
All positions	−3.8 ± 3.0	−13.7 ± 1.2	9.9 ± 4.2	+4.7 ± 4.1	+10.1 ± 1.8	5.4 ± 5.9
Codon pos. I	−21.7 ± 5.0	−31.0 ± 2.2	9.3 ± 7.2	+21.9 ± 4.6	+31.1 ± 2.1	9.2 ± 6.7
Codon pos. II ...	+16.2 ± 5.9	+11.0 ± 2.6	5.2 ± 8.5	−1.9 ± 8.1	+6.1 ± 2.6	8.0 ± 10.7
Codon pos. III ...	+1.4 ± 6.4	−10.8 ± 3.2	12.2 ± 9.6	−3.0 ± 3.4	−1.7 ± 2.4	1.3 ± 5.8

NOTE.—The importance of the differences between the two groups is given by Δ, the distance between the values in the two groups.

T deviation was not significant when integral membrane proteins were removed.

Discussion

All three bacterial genomes examined showed asymmetric substitution patterns. The unequal substitution patterns divide the chromosome into two segments. As these segments are defined by a symmetry break (a non-"racemic" proportion of one base and its complementary base), they could be termed *chirochores* by analogy with the isochores (Bernardi 1989), which are defined by a homogeneous G + C content.

There is some evidence that the asymmetry of substitution patterns is due to mutational bias, with dissimilar mutation rates in the two strands, rather than selective bias. The deviation switches at the origin and terminus of chromosome replication suggest a link with the replication and repair system; the relative increase in bias in intergenic regions and third codon positions show that relaxation of selective pressure increases bias, as might be expected with mutational bias.

However, it is clear that selective interpretation of chirochores cannot be ruled out, for it is always possible to develop selective hypotheses. For example, during the revision of the manuscript, a subtle selectionist interpretation was suggested by an anonymous referee. Chi-sites are preferentially found in the leading strand in *E. coli* (Sofia et al. 1994; Burland et al. 1993; Plunkett et al. 1993; Blattner et al. 1993), and this is interpreted as the result of selective pressure for better processing, by the RecBCD recombinational pathway, of collapsed replication forks due to single-strand interruption in template DNA (Kuzminov 1995). The Chi-site is an octamer 5′ GCTGGTGG 3′ that would occur only every 65.5 kb with an equifrequency base distribution. Since Chi-sites contain much more G than C, the C-G deviation may result from selective pressure to increase occurrences of Chi-sites with the right orientation. This does not seem to be the case, because a C-G deviation of −5.0% would produce a Chi-site every 54 kb, and this is not sufficient to explain the observed one site per 5 kb in *E. coli.*

According to the mutational bias hypothesis, the asymmetry could be a consequence of asymmetries in replication or repair. Replication is asymmetric in vitro, but the situation in vivo is less clear, since both strands of the duplex are replicated discontinuously (Okazaki et al. 1986; Ogawa and Okazaki 1980; Wang and Smith 1989). If the mechanisms of DNA replication are symmetric, then only repair is left to account for the asymmetry. The difference between leading and lagging transcripts suggests a link with transcription. There is an asymmetric interaction between transcription and repair, with a preferential removal of cyclobutane dimers from the transcription template strand (Mellon and Hanawalt 1989; Koehler et al. 1991; Hanawalt 1991; Lommel et al. 1995). The consequence would be more pyrimidines than purines in the transcription template strand; this type of enrichment has been reported for several bacteria, bacteriophages, and higher organisms (Szybalski et al. 1966). In coding sequences, the consequence would be more purines than pyrimidines, yielding a negative C-G deviation and a positive A-T deviation, as observed here. However, the correlation that might be expected between deviations and gene expressivity because there is no selective removal of pyrimidine dimers without transcription (Mellon and Hanawalt 1989) does not occur. Moreover, this preferential repair alone does not explain the difference in deviation between leading and lagging transcripts.

Acknowledgments

Thanks are due to Christian Gautier, Manolo Gouy, Dan Graur, Jacques Ninio, Guy Perrière, Andrei Kuzminov, and Noboru Sueoka. The interpretation of results was greatly improved by stimulating comments from two anonymous referees and the reviewing editor. This work was supported in part by "Groupement de Recherches et d'Etudes sur les Génomes."

LITERATURE CITED

BAKER, T. A., and S. H. WICKNER. 1992. Genetics and enzymology of DNA replication in *Escherichia coli.* Annu. Rev. Genet. **26**:447–477.

BERNARDI, G. 1989. The isochore organization of the human genome. Annu. Rev. Genet. **23**:637–661.

BLATTNER, F. R., V. BURLAND, G. PLUNKETT, H. J. SOFIA, and D. L. DANIELS. 1993. Analysis of the *Escherichia coli* genome. IV. DNA sequence of the region from 89.2 to 92.8 minutes. Nucl. Acids. Res. **21**:5408–5417.

BOOR, K. J., M. L. DUNCAN, and C. W. PRICE. 1995. Genetic and transcriptional organization of the beta subunit of *Bacillus subtilis* RNA polymerase. J. Biol. Chem. **270**:20329–20336.

BULMER, M. 1991. Strand symmetry of mutation rates in the β-globin region. J. Mol. Evol. **33**:305–310.

BURLAND, V., G. PLUNKETT, D. L. DANIELS, and F. R. BLATTNER. 1993. DNA sequence and analysis of 136 kilobases of the *Escherichia coli* genome: organizational symmetry around the origin of replication. Genomics **16**:551–561.

DANIELS, D. L., G. PLUNKETT, V. BURLAND, and F. R. BLATTNER. 1992. Analysis of the *Escherichia coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes. Science **257**:771–778.

DESCHAVANNE, P., and J. FILIPSKI. 1995. Correlation of GC content with replication timing and repair mechanisms in weakly expressed *E. coli* genes. Nucl. Acids. Res. **23**:1350–1353.

FILIPSKI, J. 1990. Evolution of DNA sequence, contributions of mutational bias and selection to the origin of chromosomal compartments. Pp. 1–54 *in* G. OLE, ed. Advances in mutagenesis research 2. Springer-Verlag, Berlin.

FLEISCHMANN, R. D., M. D. ADAMS, O. WHITE et al. (40 coauthors). 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269**:496–512.

HANAWALT, P. C. 1991. Heterogeneity of DNA repair at the gene level. Mutat. Res. **247**:203–211.

HILL, M. O. 1974. Correspondence analysis: a neglected multivariate method. Appl. Statis. **23**:340–354.

JEONG, S. M., H. YOSHIKAWA, and H. TAKAHASHI. 1993. Isolation and characterization of the *secE* homologue gene of *Bacillus subtilis.* Mol. Microbiol. **10**:133–142.

JERMIIN, L. S., D. GRAUR, and R. H. CROZIER. 1995. Evidence from analyses of intergenic regions for strand-specific directional mutation pressure in metazoan mitochondrial DNA. Mol. Biol. Evol. **12**:558–563.

KOEHLER, D. R., S. S. AWADALLAH, and B. W. GLICKMAN. 1991. Sites of preferential induction of cyclubutane pyrimidine dimers in the nontranscribed strand of *lacI* correspond with sites of UV-induced mutation in *Escherichia coli.* J. Biol. Chem. **266**:11766–11773.

KUZMINOV, A. 1995. Collapse and repair of replication forks in *Escherichia coli.* Mol. Microbiol. **16**:373–384.

LOBRY, J. R. 1995. Properties of a general model of DNA evolution under no–strand-bias conditions. J. Mol. Evol. **40**: 326–330.

LOMMEL, L., C. CARSWELL-CRUMPTON, and P. C. HANAWALT. 1995. Preferential repair of the transcribed DNA strand in the dihydrofolate reductase gene throughout the cell cycle in UV-irradiated human cells. Mutat. Res. **336**:181–192.

MARIANS, K. J. 1992. Prokaryotic DNA replication. Annu. Rev. Biochem. **61**:673–719.

MELLON, I., and P. C. HANAWALT. 1989. Induction of the *Escherichia coli* lactose operon selectively increases repair of its transcribed DNA strand. Nature **342**:95–98.

NOSSAL, N. G. 1983. Prokaryotic DNA replication systems. Annu. Rev. Biochem. **52**:581–615.

OGASAWARA, N., S. NAKAI, and H. YOSHIKAWA. 1994. Systematic sequencing of the 180 kilobase region of the *Bacillus subtilis* chromosome containing the replication origin. DNA Res. **1**:1–14.

OGAWA, T., and T. OKAZAKI. 1980. Discontinuous DNA replication. Annu. Rev. Biochem. **49**:421–457.

OKAZAKI, R., T. OKAZAKI, K. SAKABE, K. SUGIMOTO, R. KAINUMA, A. SUGINO, and N. IWATSUKI. 1968. In vivo mechanism of DNA chain growth. Cold. Spring Harbor Symp. Quant. Biol. **33**:129–143.

PERRIERE, G., I. MOSZER, and T. GOJOBORI. 1996. NRSub: a non-redundant data base for *Bacillus subtilis.* Nucl. Acids Res. **24**:41–45.

PLUNKETT, G., V. BURLAND, D. L. DANIELS, and F. R. BLATTNER. 1993. Analysis of the *Escherichia coli* genome. III. DNA sequence of the region from 87.2 to 89.2 minutes. Nucl. Acids Res. **21**:3391–3398.

SMITHIES, O., W. R. ENGELS, J. R. DEVEREUX, J. L. SLIGHTOM, and S.-H. SHEN. 1981. Base substitutions, length differences and DNA strand asymmetries in the human $^{G}\gamma$ and $^{A}\gamma$ fetal globin gene region. Cell **26**:345–353.

SOFIA, H. J., V. BURLAND, D. L. DANIELS, G. PLUNKETT, and F. R. BLATTNER. 1994. Analysis of the *Escherichia coli* genome. V. DNA sequence of the region from 76.0 to 81.5 minutes. Nucl. Acids Res. **22**:2576–2586.

SUEOKA, N. 1995. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. J. Mol. Evol. **40**:318–325.

SZYBALSKI, W., H. KUBINSKI, and P. SHELDRICK. 1966. Pyrimidine clusters on the transcribing strand of DNA and their possible role in the initiation of RNA synthesis. Cold. Spring Harbor Symp. Quant. Biol. **31**:123–127.

WANG, T.-C. V., and K. C. SMITH. 1989. Discontinuous DNA replication in a *lig-7* strain of *Escherichia coli* is not the result of mismatch repair, nucleotide-excision repair, or the base-excision repair of DNA uracil. Biochem. Biophys. Res. Comm. **165**:685–688.

WU, C.-I., and N. MAEDA. 1987. Inequality in mutation rates of the two strands of DNA. Nature **327**:169–170.

WU, C.-I. 1991. DNA strand asymmetry. Nature **352**:114–114.

YURA, T., H. MORI, H. NAGAI, T. NAGATA, A. ISHIHAMA, N. FUJITA, K. ISONO, K. MIZOBUCHI, and A. NAKATA. 1992. Systematic sequencing of the *Escherichia coli* genome: analysis of the 0–2.4 min region. Nucl. Acids. Res. **20**: 3305–3308.

PAUL M. SHARP, reviewing editor

Accepted February 1, 1996