Ongoing Evolution of Strand Composition in Bacterial Genomes

*Eduardo P. C. Rocha**† *and Antoine Danchin*†‡

*Atelier de BioInformatique, Université Paris VI, Paris, France †Unité REG, URA 2171, Institut Pasteur, Paris, France; and ‡HKU-Pasteur Research Centre, Pokfulam, Hong Kong

We tried to identify the substitutions involved in the establishment of replication strand bias, which has been recognized as an important evolutionary factor in the evolution of bacterial genomes. First, we analyzed the composition asymmetry of 28 complete bacterial genomes and used it to test the possibility that asymmetric deamination of cytosine might be at the origin of the bias. The model showed significant correlation to the data but left unexplained a significant portion of the variance and indicated a systematic underestimation of GC skews in comparison with TA skews. Second, we analyzed the substitutions acting on the genes from five fully sequenced *Chlamydia* genomes that had not suffered strand switch since speciation. This analysis showed that substitutions were not at equilibrium in *Chlamydia trachomatis* or in *C. muridarum* and that strand bias is still an on-going process in these genes. Third, we identified substitutions involved in the adaptation of genes that had switched strands after speciation. These genes adapted quickly to the skewed composition of the new strand, mostly due to C→T, A→G, and C→G asymmetric substitutions. This observation was reinforced by the analysis of genes that switched strands after divergence between *Bacillus subtilis* and *B. halodurans.* Finally, we propose a more extended model based on the analysis of the substitution asymmetries of *Chlamydia.* This model fits well with the data provided by bacterial genomes presenting strong strand bias.

Introduction

Mutational pressures leading to dramatic differences between the nucleotide compositions of genomes have long been recognized among bacteria (Sueoka 1962). These patterns produce heterogeneities in the chromosome, either because there is horizontal transfer from genomes with very different compositions (Ochman, Lawrence, and Groisman 2000) or because the mechanisms causing the bias act differentially in certain regions of the chromosome (Gautier 2000). Such is the case for replicating strand-associated biases. Under no-strand bias conditions, one would expect to find the equalities A = T and C = G in each strand of DNA (Lobry 1995). However, the analysis of complete bacterial genomes reveals an important asymmetry between the leading and the lagging replicating strands (Lobry 1996). These observations have been confirmed in many bacteria and have allowed for the determination of the putative origins of replication in a substantial number of bacterial genomes (Grigoriev 1998; Salzberg et al. 1998; Lopez et al. 1999; Rocha, Danchin, and Viari 1999*a*). Interestingly, all genomes present the same asymmetry, G is relatively more abundant than C in the genes coded in the leading strand and less abundant in the genes coded in the lagging strand, which is frequently accompanied by a larger abundance of T over A in the leading strand. These biases propagate into higher-order biases in a correlated way, thereby changing the relative frequencies of codons and amino acids of genes and corresponding proteins in each of the replicating strands (Perrière, Lobry, and Thioulouse 1996; McInerney 1998; Rocha, Danchin, and Viari 1999*a*).

Bacterial genes transcribed at high levels have a preference for positioning into the leading strand, presumably to minimize collisions between the replication fork and the transcription bubble (McLean, Wolfe, and Devine 1998). As a consequence, more genes are coded from the leading strand than from the lagging strand in most bacteria (in Gram-positive bacteria, this may go up to 80% of the genes in the leading strand) (Rocha et al. 2000). Since protein-coding sequences do not contain the same abundance of G and C, this bias accumulates with the replication bias (Sueoka 1999). Methods designed to deal with this problem revealed that replication strand bias was not due to the asymmetric distribution of genes in the bacterial chromosomes (Rocha, Danchin, and Viari 1999*a*; Mackiewicz et al. 1999).

Explanations for strand biases as a by-product of mechanisms other than replication also include transcription-coupled repair, codon usage bias, and oligonucleotide bias (Francino et al. 1996; Mrázek and Karlin 1998; Salzberg et al. 1998). The former are related to the asymmetrical distribution of highly expressed genes along the two DNA strands. However, the data on codon usage bias (Moszer, Rocha, and Danchin 1999) and on expression arrays (Tao et al. 1999) seem to indicate that only a reduced number of genes are highly expressed at exponential growth. Indeed, differences in transcription levels have not been found to constitute a major cause of replication-linked bias (Tillier and Collins 2000*a*). Finally, the contribution of signals such as the χ sequence to strand bias was found to be very small due to the fraction of the genome they occupy (Tillier and Collins 2000*a*).

Among the theories aimed at explaining strand bias based on the asymmetry of the replication bubble, the cytosine deamination theory enjoys the most attention

Abbreviations: DS, orthologous genes present in different replicating strands (i.e., leading versus lagging); K_a, nonsynonymous substitution rate; K_s, synonymous substitution rate; NS, genes without orthologs in the other species; SS, orthologous genes present in the same replicating strand.

Key words: replication, strand bias, mutation, genome analysis, sequence evolution.

Address for correspondence and reprints: Eduardo P. C. Rocha, Atelier de BioInformatique, Université Paris VI, 12 Rue Cuvier, 75005 Paris, France. E-mail: erocha@abi.snv.jussieu.fr.

Mol. Biol. Evol. 18(9):1789–1799. 2001
 ISSN: 0737-4038

(Frank and Lobry 1999). The deamination of cytosine in DNA occurs at significant rates in vivo and leads to the formation of uracil, which is excised by the action of uracil-DNA glycosylase (Lindahl 1993). The rate of cytosine deamination increases by a factor of 140 when the DNA is single-stranded (Beletskii and Bhagwat 1996). Methylation of cytosine is known to increase the rate of deamination by a factor of 4. In this case, the deamination produces a T, which cannot be corrected by the glycosylase (Coulondre et al. 1978). Since the leading strand is more exposed in the single-stranded state (in order to serve as template for the synthesis of the lagging strand) (Marians 1992), and C→T mutations would induce the formation of GC skews, cytosine deamination has been proposed to be at the basis of strand bias (Frank and Lobry 1999). This hypothesis has the advantage of explaining GC and TA skews (and larger GC skews in G+C-poor genomes) within a known mechanistic model and based on a well-established mutational hot spot.

Recently, a large number of studies have accounted for strand asymmetries (for reviews, see Francino and Ochman 1997; Frank and Lobry 1999). However, several important questions still remain unanswered: (1) Are the genomes that present strong strand bias at compositional equilibrium? (2) What are the major substitutions associated with the establishment of the bias? (3) Is there a simple specific function/mutation responsible for the bias? To tackle these questions, we benefited from the existence of the complete genome sequences of several very closely related bacteria, in particular five *Chlamydia* genomes (Stephens et al. 1998; Kalman et al. 1999; Read et al. 2000; Shirai et al. 2000) and two species of *Bacillus* (Kunst et al. 1997; Takami et al. 2000).

Asymmetrical changes are usually studied either by phylogenetic reconstruction of homologous sequences or by detection of deviations from the parity of bases in the genome. In the present work, we explored both types of methodologies, taking advantage of the very strong strand bias of *Chlamydia* and *Bacillus* genomes and their extensive homology.

Materials and Methods

Data

Sequence data for all complete bacterial genomes were retrieved from GenBank (http://www.ncbi.nlm.nih.gov). We analyzed the following complete genomes, using the annotations contained in their respective GenBank files: *Aquifex aeolicus* (Aae), *Bacillus halodurans* (Bha), *B. subtilis* (Bsu), *Borrelia burgdorferi* (Bbu), *Buchnera* sp. (Bsp), *Caulobacter crescentus* (Ccr), *Campylobacter jejuni* (Cje), *Chlamydia pneumoniae* CWL029 (CpnC), *C. pneumoniae* AR39 (CpnA), *C. pneumoniae* J130 (CpnJ), *C. trachomatis* serovar D (Ctr), *C. muridarum* (Cmu), *Deinococcus radiodurans* (two chromosomes) (Dra), *Escherichia coli* (Eco), *Haemophilus influenzae* (Hin), *Helicobacter pylori* 26695 (Hpy), *Mycoplasma genitalium* (Mge), *M. pneumoniae* (Mpn), *Mycobacterium tuberculosis* (Mtu), *Neisseria meningitidis* MC58 (Nme), *Pseudomonas aeruginosa* (Pae), *Rickettsia prowazekii* (Rpr), *Synechocystis* spp. C125 (Ssp), *Treponema pallidum* (Tpa), *Thermotoga maritima* (Tma), *Ureaplasma urealyticum* (Uur), *Vibrio cholerae* (two chromosomes) (Vch), and *Xylella fastidiosa* (Xfa).

Statistical Analysis of Skews

Identification of Biased Genomes Through Linear Discriminant Analysis

Simple cumulative GC and TA skews are sensitive to different populations of genes in the two replicating strands. Therefore, we used linear discriminant analysis to identify genomes with significant strand bias (Rocha, Danchin, and Viari 1999*a*). We considered a genome to contain a significant strand bias when the maximal accuracy of our method (percentage of true positives in the classification of leading- and lagging-strand genes) was better than that of the best of 10 random genomes of the same size and composition. In practice, this implies that a genome displays significant strand bias if the accuracy of the discrimination between the replicating strands is larger than 60%–75% (depending on genome length). Because different bacterial strains within a species are often very similar in sequence, and in order not to bias the results, we analyzed only one representative strain for each bacterial species.

Identifying Origins of Replication by GC Skew

The skews for a given sequence were defined as (Lobry 1996)

$$\text{GC skew} = (G - C)/(G + C), \quad (1)$$

$$\text{TA skew} = (T - A)/(T + A). \quad (2)$$

After the identification of the genomes presenting a significant strand bias by linear discriminant analysis, we computed cumulative GC skews with overlapping 10-kb windows (1-kb step) to identify the origin and terminus of replication. When unknown, the origin and terminus were defined as the maxima of the integrated curves such that the leading strand contained more G's and more genes than the lagging strand (Grigoriev 1998).

Genes' GC Skews

We computed GC and TA skews for the gene sequences using equations (1) and (2). To compare the differences in skews between the genes present in the different strands, we computed ΔGC and ΔTA skews. These quantities are defined as the difference between the average skews of the genes in the leading strand and the ones in the lagging strand. Considering N_{leading} and N_{lagging}, the numbers of genes in the leading and lagging strands, respectively, one obtains

$$\Delta\text{GC skew} = \frac{\sum_{i\in\text{genes leading}} \text{GC skew}_i}{N_{\text{leading}}} - \frac{\sum_{i\in\text{genes lagging}} \text{GC skew}_i}{N_{\text{lagging}}} \quad (3)$$

$$\Delta\text{TA skew} = \frac{\sum_{i\in\text{genes leading}} \text{TA skew}_i}{N_{\text{leading}}} - \frac{\sum_{i\in\text{genes lagging}} \text{TA skew}_i}{N_{\text{lagging}}}. \quad (4)$$

This eliminates the bias due to the larger population of genes in the leading strand, and it also normalizes the replication biases in terms of the average gene asymmetry in nucleotide composition.

Analysis of Similarity

Definition of Homologous Genes

Two genes were considered homologous if they coded for proteins similar both in sequence and in size. To identify homologous genes, we performed pairwise comparisons of all proteins of all proteome pairs, filtering potential hits with $P < 10^{-5}$ in BlastP and a maximal difference of protein lengths of 20%. Subsequently, we aligned the sequences using a variant of the classical dynamic programming algorithm for global alignment, where one counts 0-weight for gaps at both ends of the largest sequence using the BLOSUM62 matrix (Erickson and Sellers 1983). Finally, we retained pairs of proteins with more than 40% similarity.

Classification of Orthologous Genes

Two homologous genes were considered to be orthologous if they were each other's best matches in the respective genomes. We obtained 687 sets of five orthologous for the *Chlamydia* set and 2,123 sets of two orthologous for the *Bacilli* set. Using these sets, we analyzed the conservation of the gene organization between genomes by displaying a scatter-plot of the positions of orthologous genes in the different genomes. We further defined three classes of genes for *Chlamydia* and for *Bacillus*: genes present in all genomes in the same replicating strand (SS), genes present in different replicating strands (DS), and genes not present in other genomes according to our stringent orthology criteria (NS). For *Chlamydia,* this resulted in 49 DS and 638 SS genes, whereas for *Bacillus* we obtained 372 DS and 1751 SS genes. Naturally, the number of NS genes changed from species to species. For example, we obtained 1,977 NS genes for *B. subtilis* and 131 NS genes for *C. muridarum.*

Characterization of Orthologous Genes

The *Chlamydia* set was extremely interesting due to the small divergence between the five genomes. The similarity of the ribosomal 16S subunit was 93.5% between *C. trachomatis* and *C. pneumoniae,* 97.4% among *C. trachomatis* and *C. muridarum* genomes, and >99% among *C. pneumoniae* strains. This is in accordance with the phylogeny of *Chlamydia* that proposes a first speciation event between *C. pneumoniae* and the pair *C. trachomatis*/*C. muridarum* (Everett, Bush, and Andersen 1999). The divergence between the two *Bacillus* 16S subunits (94.4%) was intermediate to that between different species of *Chlamydia.* A further advantage of the *Chlamydia* set was the intermediate level of synteny between the elements: within *C. pneumoniae* and within the pair *C. trachomatis*/*C. muridarum,* there was complete conservation of gene order, whereas between these two sets there was less conservation. As a result, the classification of orthologous genes was valid for the genomes of *C. pneumoniae* on one side and for *C. trachomatis* and *C. muridarum* on the other. Also, the vast majority of NS genes were present either in all *C. pneumoniae* strains or in both *C. trachomatis* and *C. muridarum.* Since *Chlamydia* and *Bacillus* contain strand biases, one expects a strong and significant signal in these genomes. Finally, we can neglect the effects of differences in G+C content, because they were similar within the two groups of genomes (table 1).

Characterization of Differences in Alignments

Directed Changes

We made multiple alignments of protein and DNA sequences of SS and DS genes using CLUSTAL W with default parameters (Thompson, Higgins, and Gibson 1994). Multiple alignments of proteins were back-translated into DNA in order to determine rates of synonymous (K_s) and nonsynonymous (K_a) substitutions (Li 1997). Since these values are sensitive to low levels of sequence identity, we took into consideration only pairs of genes with $K_s < 2.0$ for the analysis of SS genes. We used this set of DNA alignments to analyze the spectrum of changes observed among the orthologous genes. For this analysis, we invoked parsimony by counting a change only when a column of the multiple alignments presented a nucleotide consensus minus one (e.g., one column with four G's and one C is a G→C change in the sequence with C). Since each alignment corresponded to a set of SS genes whose positions in the chromosome were known, we could separate the alignments into two subsets; one corresponding to all genes in the leading strand (346 genes) and the other corresponding to genes in the lagging strand (292 genes). Consequently, we analyzed the changes occurring in each strand separately. For these analyses, we used JaDis, a publicly available program designed to compute distances between nucleic acid sequences (http://pbil.univ-lyon1.fr/) (Gonçalves et al. 1999). We performed a similar analysis for DS genes. In both cases, the absolute values of substitutions from i to j were converted to relative substitution frequencies f_{ij} by dividing by the average number of nucleotides i in the sequence and expressing the result as a percentage among all types of nucleotide substitution (Gojobori, Li, and Graur 1982).

Table 1
Asymmetries Observed in the Complete Genomes of Bacteria Used in the Study

Species	%G+C in Genes	Lag/Lead[a] (nt)	Maximal Accuracy[b] (%)	ΔGC Skew	ΔTA Skew
Aae	43.5	—	NS	—	—
Bha	43.7	0.255	78	0.172	0.027
Bsu	43.3	0.330	78	0.143	0.014
Bbu	28.1	0.516	99	0.619	0.419
Bsp	26.3	0.778	81	0.147	0.095
Ccr	67.7	0.826	68	0.034	0.026
Cje	53.8	0.622	82	0.416	0.074
Cmu	41.6	0.891	98	0.484	0.071
CpnA	41.3	0.868	89	0.291	0.056
Ctr	41.6	0.891	97	0.439	0.056
Dra_1	67.0	—	NS	—	—
Dra_2	66.7	—	NS	—	—
Eco	51.2	0.836	70	0.081	0.025
Hin	38.8	0.921	69	0.152	0.036
Hpy	39.5	0.717	67	0.110	0.003
Mge	31.7	0.254	NS	—	—
Mpn	40.0	0.274	NS	—	—
Mtu	65.7	0.710	68	0.089	0.131
NmeA	53.2	0.859	73	0.219	0.160
Pae	67.2	0.758	75	0.106	0.265
Rpr	30.3	0.624	78	0.274	0.062
Ssp	47.7	—	NS	—	—
Tma	46.4	0.825	68	0.063	0.046
Tpa	52.8	0.544	90	0.335	0.190
Vch_1	48.3	0.665	80	0.171	0.051
Vch_2	47.6	0.706	83	0.167	0.050
Xfa	53.8	0.622	77	0.324	0.385
Uur	25.5	0.348	NS	—	—

[a] Ratio of coding nucleotides in the lagging strand to coding nucleotides in the leading strand.

[b] Maximal accuracy on the discrimination of replicating strands using the nucleotide composition of genes in a linear discriminant analysis. NS = nonsignificant accuracy.

Undirected Changes

Since the switch of replicating strand took place before the divergence *C. trachomatis*/*C. muridarum* or of *C. pneumoniae* strains, we used a complementary approach to analyze DS genes, using pairwise alignments to count mismatches. This analysis should account for the adaptation of genes since switching strands, whereas the analysis of DS genes with multiple alignments only accounts for adaptation since speciation. To limit our analysis to well-conserved proteins, we imposed a stricter threshold of similarity (>60% in protein sequence), which resulted in a set of proteins exhibiting an average similarity of 75%. These proteins were more constrained at the amino acid level, which renders pairwise alignments more reliable. For the two bacilli, we lacked a sufficiently close outgroup to determine the direction of mutations, and the same method was used: (1) We aligned the sequences and identified the mismatches. (2) Having arbitrarily chosen one of the two species as the reference species (RS), we cataloged all mismatches in classes $X_{RS}:Y_{nonRS}$ ($X \neq Y$).(3) We separated all DS genes into two categories: the genes that were in the leading strand in the reference species (lagging in the other genome), and those that were in the lagging strand in the reference species (leading in the other genome). (4) For each of these two categories, we computed the difference between $X_{RS}:Y_{nonRS}$ and $Y_{RS}:X_{nonRS}$, which indicates the asymmetry in terms of mismatches. The comparison of the asymmetries between the two categories indicates the asymmetry at the basis of the adaptation to the new strand.

It is important to emphasize the differences between this and the preceding analysis. Suppose that a gene is in the leading strand of the reference species. In the analysis of the multiple alignment of SS genes, a mismatch $C_{4,lead}:T_{1,lead}$ indicates a $C_{lead}\rightarrow T_{lead}$ substitution. In the pairwise alignment of DS genes, a mismatch $T_{lead}:C_{lag}$ cannot be interpreted as a $C_{lead}\rightarrow T_{lead}$ change. This mismatch may have been caused by C→T in the leading strand of the reference genome or by T→C in the lagging strand of the other genome (in the previous analysis, the probability of four $T_{lead}\rightarrow C_{lead}$ changes was neglected, referring to parsimony). Therefore, C : T mismatches in DS genes indicate either a $C_{lead}\rightarrow T_{lead}$ change or an $A_{lead}\rightarrow G_{lead}$ change. We indicate this ambiguity by C(A)→T(G).

Further Analysis of Alignments

To analyze the importance of multiple substitutions in our data set, we looked for changes in the SS multiple alignments between *C. trachomatis* and *C. muridarum* (which corresponded to a larger evolutionary distance than the one between *C. pneumoniae* strains). The average K_s was below one substitution per synonymous site, and the upper K_s value for the 95% confidence in-

terval was 0.84 (for nonsynonymous sites [K_a] it was 0.07).

Results

Strand Biases Among Bacteria

Prevalence of Strand Biases

Twenty-one out of 28 chromosomes (representing 26 species) of bacteria had significant strand biases (table 1). Exceptions were the Mycoplasmas, *D. radiodurans, A. aeolicus,* and *Synechocystis* sp., many of which have been described previously (Rocha, Danchin, and Viari 1999*a*). The largest strand bias (as measured by maximal discrimination) was observed in the genome of *Borrelia burgdorferi,* followed by *C. trachomatis, T. pallidum, C. pneumoniae,* and chromosome 2 of *V. cholerae.* Although many of these highly biased genomes presented low G+C contents, the correlation between accuracy and G+C content was not significantly different from 0 (Spearman's rho = -0.18, $P > 0.3$). In fact, the above-mentioned genomes lacking strand bias (with the exception of *D. radiodurans*) also had G+C contents well below 50%.

Correlation Between Discrimination and Skews

ΔGC skews of the genomes were closely correlated with the accuracy of discrimination between strands. This was less evident for the ΔTA skews. In fact, the Spearman's rank correlation between maximal discrimination and ΔGC skew was 0.77 ($P < 0.001$), but it was only 0.40 between maximal discrimination and ΔTA skew ($P < 0.1$). Because the cytosine deamination theory predicts proportionality between the two skews, we built a model for the theory and tested it with the data of biased genomes.

Testing the Cytosine Deamination Theory

The composition of a hypothetical unbiased gene in terms of the four bases is $N_{A,0}$, $N_{C,0}$, $N_{G,0}$, $N_{T,0}$, taken as the mean of the average composition of the leading- and lagging-strand genes. Suppose that the gene is in the leading strand and will suffer the corresponding asymmetry. The deamination theory predicts that strand bias will induce a C→T change with probability z for each C during a period t. Therefore, the composition of the gene remains unchanged for G and A, but not for C and T:

$$dN_A/dt = 0; \qquad dN_C/dt = -zN_C; \qquad dN_G/dt = 0;$$
$$dN_T/dt = zN_C. \tag{5}$$

At time t, the nucleotide frequencies of the gene become $N_{A,0}$, $e^{-zt}N_{C,0}$, $N_{G,0}$, $N_{T,0} + (1 - e^{-zt})N_{C,0}$, which corresponds to a change

$$\begin{cases} GC^0_{skew} = \dfrac{N_{G,0} - N_{C,0}}{(N_{G,0} + N_{C,0})} \\ TA^0_{skew} = \dfrac{N_{T,0} - N_{A,0}}{(N_{T,0} + N_{A,0})} \end{cases}$$
$$\rightarrow \begin{cases} GC^t_{skew} = \dfrac{N_{G,0} - e^{-zt}N_{C,0}}{(N_G + e^{-zt}N_{C,0})} \\ TA^t_{skew} = \dfrac{N_{T,0} + N_{C,0}(1 - e^{-zt}) - N_{A,0}}{(N_{T,0} + N_{C,0}(1 - e^{-zt}) + N_{A,0})}. \end{cases} \tag{6}$$

We can determine the observed values of GC^t and TA^t and test if they coincide with the expected ones. This results in two estimations of zt from equation (6) for each genome, one coming from TA skews and the other from GC skews, which can be compared by a nonparametric test (Wilcoxon test). For each genome, a different zt will apply, but within a genome, the pair should be identical. Thus, we are not comparing the values of the bias in different genomes, but checking the ability of C→T asymmetries to predict coherent values for ΔGC^t and ΔTA^t within each genome. Because the deamination theory is a mutational theory, the best fit of the model should correspond to the analysis of third positions of codons. Note that this model refers to C→T asymmetries, of which cytosine deamination is only one particular case.

The 21 bacterial chromosomes with strand bias showed a significant correlation between the two values of zt predicted by the two expressions in equation (6) using third positions of codons. In fact, after excluding *Buchnera* sp. (see below), the Spearman's rank correlation between the two terms was 0.83 ($P < 0.001$), suggesting an important contribution of C→T asymmetries to the establishment of the strand bias. However, a systematic underevaluation of ΔGC skew was also apparent ($P < 0.05$, Wilcoxon test). In 17 out of 20 chromosomes, the number of changes required to explain the ΔTA skews was larger than the one required to explain the ΔGC skew. The systematic underestimation of ΔGC skews suggests the existence of other sources of bias.

Preliminary regression analysis of the zt data indicated heteroscedasticity, with variance increasing with the values of zt. Standard procedures of regression analysis recommend a logarithmic transformation of the data in such cases (Zar 1996). The linear regression on the transformed data (excluding *Buchnera* sp. and *H. pylori*) resulted in a fit presenting a coefficient of determination of 0.68 ($P < 0.001$). Although the regression line fits the data well, it does not fit the expected result (fig. 1). This confirms the results of the Wilcoxon test, suggesting that a model based solely on C→T asymmetries underestimates GC skews.

Strand Bias Oscillations in *Chlamydia*

Chlamydia *SS Genes Are Not at Equilibrium*

We identified the recent substitution asymmetries in the SS genes of *Chlamydia* by analyzing the changes in *C. trachomatis, C. muridarum,* and *C. pneumoniae.*

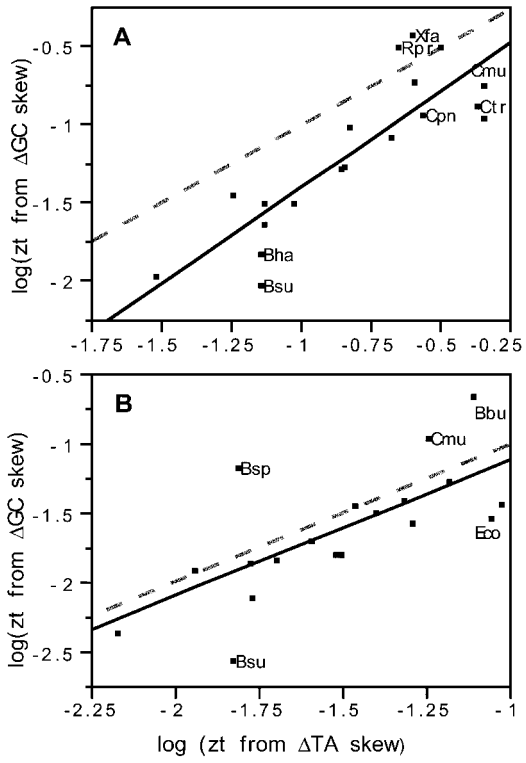

FIG. 1.—Regression analysis of *zt* values using ΔTA and ΔGC skews in 20 bacterial genomes. The dashed gray lines represent the expectations of the models. *A,* Model including only C→T asymmetries (eq. 6), where *Buchnera* sp. ASP and *Helicobacter pylori* were excluded as outliers (regression line: log Z(GC) = −0.17 (±0.16) + 1.23 (±0.18) log Z(TA)). *B,* Extended model, using equation (7), where *Mycobacterium tuberculosis* was excluded as outlier (regression line: log Z(GC) = −0.04 (±0.36) + 0.98 (±0.23) log Z(TA)).

For *C. pneumoniae,* we considered all changes observed in the genomes of the three strains to obtain larger counts. Because these strains diverged rather recently, the parameters of the substitution matrix had very large confidence intervals. Indeed, our analysis used only 0.1% of the positions in the multiple alignments for *C. pneumoniae,* compared with 6.0% for *C. trachomatis* and *C. muridarum.* The frequencies of substitutions for the other *Chlamydia* are presented in table 2. Most changes are not significantly unbalanced between replicating strands in *C. trachomatis* and *C. muridarum,* and none are statistically significant in *C. pneumoniae.* In *C. muridarum,* we observed that SS genes from both strands were getting richer in G and T and poorer in A and C ($P < 0.001$; Wilcoxon tests). In *C. trachomatis,* SS genes were getting richer in G and C and poorer in A and T ($P < 0.001$; Wilcoxon tests). When we compared the evolutions of the relative compositions in the two strands, we observed that the *C. muridarum* genes in the leading strand were getting richer in G and C and poorer in A and T ($P < 0.02$; Wilcoxon tests). In *C. trachomatis,* the leading strand genes were getting richer in T and G ($P < 0.01$; Wilcoxon tests) when compared with the lagging strand (differences for A and C are not statistically significant). As a consequence, there was no significant evolution of the GC and TA skews in the genes of the two strands for *C. muridarum,* and there was an increase in GC skew in *C. trachomatis* ($P < 0.01$; Wilcoxon test), with no significant differences for TA skews. For *C. pneumoniae,* the differences were not statistically significant. The most important asymmetries in substitution frequencies in *C. muridarum* and *C. trachomatis* were A→G–G→A (2.64% and 0.86%, respectively) and C→T–T→C (1.45% and −0.87%, respectively), but only the former were statistically significant ($P < 0.01$; Wilcoxon test).

Table 2
Relative Substitution Frequencies Observed for *Chlamydia trachomatis* (Ctr) and *Chlamydia muridarum* (Cmu) for SS Genes in the Leading and Lagging Strand

	A	C	G	T
Ctr				
Leading strand				
A	—	4.87	18.85	4.73
C	3.95	—	2.16	15.47
G	15.20	1.93	—	3.77
T	5.45	18.22	5.42	—
Lagging strand				
A	—	6.10	15.25	5.06
C	4.40	—	2.23	19.28
G	12.46	1.87	—	2.76
T	5.25	21.15	4.18	—
Cmu				
Leading strand				
A	—	3.63	18.73	4.94
C	5.13	—	1.92	18.14
G	15.19	1.92	—	4.54
T	5.05	16.50	4.31	—
Lagging strand				
A	—	4.73	14.07	5.16
C	5.69	—	2.05	20.58
G	13.17	2.02	—	3.80
T	5.38	20.39	2.97	—

The Fate of Inverted Genes and the Evolution of ΔGC Skews

Inversions in Chlamydia

Inverted genes in genomes with strand bias are expected to adapt fast to the composition of the new strand (Rocha, Danchin, and Viari 1999*a*; Tillier and Collins 2000*b*). Hence, we computed GC skews in third codon positions of DS genes in *C. pneumoniae* AR39 and *C. muridarum* (orthologous in different replicating strands) to test if they had evolved toward the bias of the new strand. We observed that all but two DS genes had GC skews typical of the new strand (fig. 2). We then proceeded to test if DS genes were distinguishable from SS and NS genes of the same current strand. Taking *C. muridarum* as a reference, we observed that GC skew at position 3 was significantly different between the three types of genes ($P < 0.001$; Kruskal-Wallis test), but, surprisingly, DS genes possessed a larger ΔGC skew (0.567) than SS (0.492) or NS (0.54) genes; i.e., the order of the bias was DS ≈ NS > SS ($P < 0.05$; Tukey-Kramer test). The same analysis applied to *C. pneumoniae* AR39 genes revealed similar results, but in

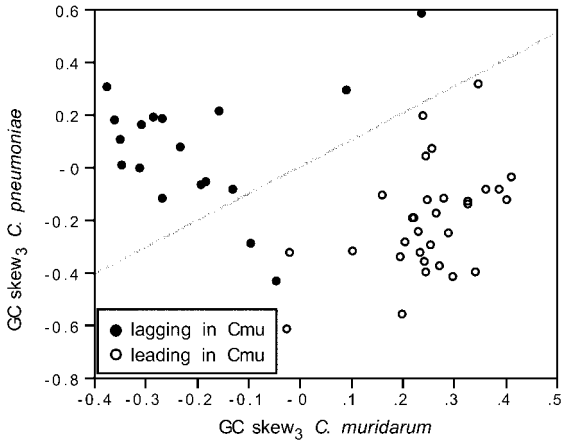

FIG. 2.—Distribution of ΔGC skews in the DS genes of leading and lagging strands of *Chlamydia muridarum* and *C. pneumoniae.*

this case the order was NS ~ DS > SS ($P < 0.01$; Tukey-Kramer test).

Inversions in Bacillus

A similar pattern was found among DS genes in *Bacillus.* Taking *B. subtilis* as a reference, we observed ΔGC skews in the order NS (0.171) ≈ DS (0.167) > SS (0.124) ($P < 0.05$; Tukey Kramer test), and similar results were obtained using *B. halodurans* as the reference. As in *Chlamydia,* the genes in *Bacillus* that have suffered an inversion have acquired the composition corresponding to the new host strand. Also, in both genera, DS genes exhibited biases larger than expected for their new strand.

Characterization of Mutations in Inverted Genes

Within each of the two *Chlamydia* monophyletic groups the genomes were collinear. Because strand switch took place before the speciation events, the analysis of 34 DS genes was first performed using only a pair of genomes. We chose *C. muridarum* and *C. pneumoniae* AR39 for simplicity (both sequences start at the origin of replication). DS genes suffered a strand switch since speciation of these two species, and this induced opposite replication biases (Tillier and Collins 2000*b*). Hence, the substitutions that took place in these genes provide clues for the establishment of the bias. We observed significant asymmetries in A(G)→C(T)–C(T)→A(G), in C→G, and especially in C(A)→T(G)–T(G)→C(A) (table 3). In this analysis, the directionality of changes could not be determined (see *Materials and Methods*).

These changes led to an increase in G and T in the leading strand and an increase in C and A in the lagging strand. Similar results were obtained for the DS genes in the two bacilli. For both *Bacillus* and *Chlamydia,* C and G increased in the respective strands at a faster rate than A and T. This is because C→G–G→C is asymmetric, but A→T–T→A is not. We analyzed the changes in the third codon positions for both sets in order to check if selective constraints at the amino acid level could be responsible for part of the signal. The differences were not statistically significant ($P > 0.1$; Wilcoxon test).

Table 3
Asymmetric Substitution Frequencies in DS Genes[a]

Mutation	*Chlamydia* Δ[b]	*Bacillus* Δ[b]
A(G)→C(T)–C(T)→A(G)	**−7.1%**	**−5.9%**
A→T–T→A	0.7%	−0.2%
C→G–G→C	**7.1%**	**5.2%**
C(A)→T(G)–T(G)→C(A)	**17.2%**	**11.9%**
Effectives	7,143/3,491	22,879/19,569

[a] Asymmetry in relative substitution frequencies between two bases *X* and *Y* comparing *C. muridarum* and *C. pneumoniae* (*B. subtilis* and *B. halodurans*) DS genes when the *C. muridarum* (*B. subtilis*) gene is on the leading or the lagging strand.

[b] Difference between relative substitution frequencies (bold values indicate significant differences, $P < 5\%$, Wilcoxon test). $N \rightarrow M - M \rightarrow N$ is given by $(f_{NM,\text{lead}} - f_{MN,\text{lead}}) - (f_{NM,\text{lag}} - f_{MN,\text{lag}})$. A(G)→C(T) indicates either an A→C change or a G→T change (see *Materials and Methods*).

Although the strand switch took place before speciation events within *C. pneumoniae* or between *C. trachomatis* and *C. muridarum,* one may suppose that some of the change occurred after speciation. In this case, the analysis of DS genes could be done using the multiple alignments, as for SS genes, with the advantage that the direction of the substitutions can be determined. We built the table of relative substitution frequencies for the DS genes (table 4), which revealed three significantly different frequencies of substitution: C→T–T→C (11.4%), A→G–G→A (8.4%). and C→G–G→C (2.2%) ($P < 0.01$; Wilcoxon tests). The A→C–C→A difference (2.4%) had the same magnitude as that of C→G-G→C, but it was not statistically significant ($P > 0.1$).

Refining the Model

We started by proposing a model based on C→T asymmetries, which were clearly insufficient to explain the data. The analysis of pairwise alignments of DS genes suggested that C→G had about half the importance of C(A)→T(G)–T(G)→C(A) in the establishment

Table 4
Relative Substitution Frequencies Observed for DS Genes in the Leading and Lagging Strands Using *Chlamydia muridarum* as the Reference Genome

	A	C	G	T
Leading strand				
A	—	4.35	20.20	5.14
C	5.58	—	3.36	21.94
G	10.87	1.43	—	2.61
T	5.42	14.87	4.25	—
Lagging strand				
A	—	5.02	13.62	4.72
C	3.82	—	1.68	20.66
G	12.64	1.98	—	2.82
T	4.17	25.04	3.84	—

of the asymmetry. The analysis of multiple alignments indicated that C→T–T→C was only slightly more important than A→G–G→A. Since these values were for *Chlamydia* and they may change for different species, we make the simplification that the three asymmetries all have the same relative importance. Hence, we can rewrite equation (5) as

$$dN_A/dt = -zN_A; \qquad dN_C/dt = -2zN_C;$$
$$dN_G/dt = zN_C + zN_A; \qquad dN_T/dt = zN_C. \qquad (7)$$

At time t, the nucleotide frequencies become $e^{-zt}N_{A,0}$, $e^{-2zt}N_{C,0}$, $N_{G,0} + (1 - e^{-2zt}/2)N_{C,0} + (1 - e^{-zt})N_{A,0}$, $N_{T,0} + (1 - e^{-2zt}/2)N_{C,0}$. Using, as previously, the genes in the leading strand to compute the observed values of GC^t and TA^t, we can test if these values coincide with the expected ones. The regression analysis of the logarithms of zt provided a good adjustment to the data ($R^2 = 0.52$, $P < 0.001$), and the regression line could not be distinguished from the expectation of the model ($P > 0.30$; t-tests for the slope and the intercept). Although a large variance remains unexplained, it seems that the use of a simple model including these three contributions with equal contributions for the establishment of the strand bias is able to correctly fit the data (fig. 1). The differences between observed and expected GC and TA skews were not statistically significant ($P > 0.1$; Wilcoxon test). However, one outlier, *M. tuberculosis*, had to be removed from the regression.

Genes Evolve at Different Rates Depending on Position and Type

Amounts of Changes in Genes in the Different Replicating Strands

The lagging-strand genes presented more changes among SS genes both for *Chlamydia* (6.6%, $P < 0.001$; Wilcoxon test) and for *Bacillus* (6.4%, $P < 0.001$).

Different Evolution Rates for NS, SS, and DS Genes

One would expect SS genes to exhibit higher GC skews, but we observed the opposite. Hence, we tested to determine if genes in different strands and belonging to different types (i.e., SS, DS) evolved at similar rates. Both in *Chlamydia* and in *Bacillus*, DS genes evolved significantly faster than SS genes ($P < 0.001$; Wilcoxon test). Between *C. trachomatis* and *C. muridarum*, similarity scores increase as follows: NS < DS < SS ($P < 0.01$; Tukey-Kramer test).

K_a/K_s Ratios

We investigated the K_a/K_s ratios of orthologs of *C. trachomatis* and *C. muridarum* and of orthologs of *C. muridarum* and *C. pneumoniae* AR39 (note that we removed all pairs for which $K_s > 2.0$; see *Materials and Methods*). The median K_a/K_s value for orthologs in *C. trachomatis*/*C. muridarum* was 0.07 (0.069 in the lagging strand and 0.073 in the leading strand) and 0.12 for orthologs in *C. muridarum*/*C. pneumoniae* AR39 (0.120 in the lagging strand and 0.117 in the leading strand). These differences between leading- and lagging-strand genes are not statistically significant. The comparison of SS and DS genes among bacilli revealed similar values of K_a/K_s for both sets (respectively, 0.12 and 0.14).

Discussion

Ongoing Strand Bias in Bacteria

Following previous observations (Mackiewicz et al. 1999; Rocha, Danchin, and Viari 1999*a*), we observed pervasiveness of strand bias in bacterial genomes, with GC skews correlating better with strand discrimination than TA skews. This suggests that strand bias should not have its origin only on C→T asymmetries, a hypothesis which is reinforced by the analyses of the models and of the asymmetric substitution rates. The extended model for the impact of asymmetry in gene evolution revealed a good fit to the data and one single outlier: *M. tuberculosis.* Nevertheless, *Buchnera* sp. is a borderline case (fig. 1), which may not be surprising given that it has the characteristics of an endocellular symbiont suffering a genome reduction that involves the lost of a considerable amount of repair mechanisms (Shigenobu et al. 2000). One should note that the species under comparison diverged a long time ago, have very different lifestyles, and contain proteomes heterogeneous in size and composition, different G+C contents, and very different sets of replication and repair machineries. All of these effects result in the considerable amount of variance that remains unexplained by the linear regression analyses.

One might expect that substitution frequencies in SS genes in stable genomes with strong GC skews such as *C. muridarum* and *C. trachomatis* might be at equilibrium. Instead, they were found to entail variations in the nucleotide compositions of the two strands. As a result, in *C. trachomatis*, there is a net increase in GC skew in SS genes (nonsignificant for *C. muridarum*). This means that, at least in this species, strand bias is still an ongoing process even in SS genes. Hence, it is not surprising that in both *Chlamydia* and *Bacillus*, DS genes have adapted to the composition of the new strand. It was previously reported that paralogs in different replicating strands of *B. burgdorferi* presented signs of adaptation to the respective strands (Lafay et al. 1999; Rocha, Danchin, and Viari 1999*a*), and a recent study of DS genes in *Chlamydia* demonstrated that they adapt fast to the new strand (Tillier and Collins 2000*b*). Our results confirm these previous works and provide some clues for the substitutions at the basis of the adaptation.

Our analysis of SS and DS genes used closely related sequences, but not enough to assure complete avoidance of multiple substitutions. To avoid a large number of multiple substitutions and to produce faithful alignments, we restricted our study to the most conserved proteins. These proteins have evolved less due to functional constraints. We are then observing substitutions for which the weight of selection may be important, which incorporates a bias, particularly due to

the codon usage bias and transcription-coupled repair of highly expressed genes. *Chlamydia* and *Bacillus* have different codon usage biases (Moszer, Rocha, and Danchin 1999; Romero, Zavala, and Musto 2000), but the analysis of the substitutions in the two groups is concordant. If codon usage bias was biasing the results in a very important way, this should not happen. Also, it has been found that biases associated with highly expressed genes do not significantly interfere with replication bias (Mackiewicz et al. 1999; Tillier and Collins 2000*a*), and one may expect this to not significantly change the results. The interference of some biases in the analyses of other biases is a very important question that unfortunately remains unsolved. Analyses of other close genomes with important strand bias may shed further light on this question.

Different Causes for a Simple Bias

The two major asymmetries in DS genes of *C. muridarum* and *C. trachomatis* since speciation are A→G–G→A and C→T–T→C. This is consistent with the observations of SS genes that also indicate that these two substitutions are the most asymmetric (although the latter is not statistically significant). It is also consistent with the analysis of DS genes using pairwise alignments in order to capture the substitutions during all the processes of adaptation. Unfortunately, the latter analysis does not allow one distinguish between C→T and A→G. Both asymmetries induce increases in GC and TA skews and lead to proportionality between the increases. C→T asymmetries may be assigned to preferential cytosine deamination in single-stranded DNA, although other hypotheses based on C→T asymmetries may also be compatible with the data. G : C→A : T transitions dominate the spectra of mutations of *E. coli*; however, most studies have focused on C→T mutations, not on A→G (Frank and Lobry 1999), and few data are available on A→G mutations.

Two other different substitution frequencies are asymmetric in the adaptation process of DS genes: C→G–G→C and A→C–C→A. Both were also observed in the analyses of pairwise alignments, but in this case A→C–C→A was indistinguishable from G→T (although G→T was not significant in the multiple-alignment approaches). Interestingly, C→G and G→C are among the most rare mutations observed in *E. coli* (Hutchinson 1996). However, the asymmetry is not necessarily correlated with the absolute number of substitutions. Indeed, also in our data set, C→G and G→C are systematically the most rare substitutions (tables 2 and 4). Nevertheless, it is puzzling to observe that the pairwise alignments indicate a stronger role for C→G–G→C in the adaptation to the strand than do the multiple alignments (7% and 2.2%, respectively). One may consider two explanations for this observation: (1) C→G asymmetries are more important in earlier phases of the adaptation, or (2) multiple substitutions in the pairwise alignments are biasing our results concerning these rare mutations (pairwise alignments represent a longer period of evolution). Although the A→C–C→A asymmetry plays no direct role in the establishment of the bias (it just converts TA skew in GC skew), one may suppose that part of the C→G substitutions in fact correspond to C→A→G multiple substitutions. Independent of its origin, the C→G asymmetry results in an increase in GC skew without an increase in TA skew. This may explain the systematic underevaluation of ΔGC skews by the model based only on C→T asymmetry: the contribution of C→G asymmetry would ''correct'' this effect (note that A→T is not asymmetric).

We have suggested elsewhere that genome shuffling would disturb strand bias (Rocha, Danchin, and Viari 1999*b*). However, one might also suppose mutation rates and repair efficiency to play an important role in the tuning of the process. As for mutation, using the example of cytosine deamination, the methylation of cytosine increases C→T mutations. In bacteria, such methylation is provided by restriction modification systems, which are constantly being acquired and lost by horizontal transfer (Jeltsch and Pingoud 1996). Therefore, mutation could be a cause of the change in bias with time. As for repair, several reports indicate that the mismatch repair system compensates for C→T substitution asymmetries, even when the cytosine is methylated (Jones, Wagner, and Radman 1987). A small change in the repair machinery could originate an increased capacity for repair and a consequent reduction in the asymmetry. The efficiency of mismatch repair does change along the evolutionary history of bacteria (Taddei et al. 1997; Sniegowski et al. 2000). It has also been shown that replicating strands exhibit different replication accuracy rates (Izuta, Roberts, and Kunkel 1995; Iwaki et al. 1996; Fijalkowska et al. 1998) and that repair might be involved in it (Radman 1998).

Since the analyses of DS genes were performed between genomes that diverged some time ago, we cannot assume that the substitutions we observe are devoid of selective constraints. Nevertheless, we observed that switched genes adapted fast to the new strand and that this adaptation resulted in a small K_a/K_s ratio. This suggests that amino acid bias is not at the origin of such biases, even if the adaptation process involves changing the amino acid content of the coded proteins (Perrière, Lobry, and Thioulouse 1996; Lafay et al. 1999). Effects of amino acid selection on strand bias have also been discarded by other analyses (Mackiewicz et al. 1999; Tillier and Collins 2000*b*). However, selective effects at the DNA level cannot be ruled out through this approach. In particular, one cannot exclude the possibility that strand bias is the result of selection for more efficient replication by the asymmetric replication bubble.

Conclusions

The existence of strand bias has important consequences for the study of bacterial molecular evolution. First, it indicates that in many bacteria the use of substitution matrices that do not take into account strand asymmetries provides a poor approximation of real data. Second, it indicates that a gene may suffer a process of accelerated evolution just through a change of replicat-

ing strand. This may also make the discrimination of paralogy from orthology difficult, especially in large functional families. Third, it provides an additional signal to use in the detection of horizontal transfer and genome rearrangements, but only for recent events. Fourth, it may elucidate still unknown particularities in the processes underlying DNA replication and repair.

Acknowledgments

Alain Viari played a very important role in earlier discussions of this work. Isabelle Gonçalves programmed in JaDis the function that allows the determination of mutations of the type consensus − 1 in the multiple alignments. We are grateful for comments and suggestions from Elisabeth Tillier, Carmen Gomes, and Isabelle Gonçalves on previous versions of the manuscript. The criticisms and suggestions of two anonymous referees constituted important contributions to this work.

LITERATURE CITED

BELETSKII, A., and A. S. BHAGWAT. 1996. Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. Proc. Natl. Acad. Sci. USA **93**:13919–13924.

COULONDRE, C., J. H. MILLER, P. J. FARABAUGH, and W. GILBERT. 1978. Molecular basis of base substitution hotspots in *Escherichia coli*. Nature **274**:775–780.

ERICKSON, B. W., and P. H. SELLERS. 1983. Recognition of patterns in genetic sequences. Pp. 55–91 *in* D. SANKOFF and J. B. KRUSKAL, eds. Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Addison-Wesley, Reading, Mass.

EVERETT, K. D., R. M. BUSH, and A. A. ANDERSEN. 1999. Emended description of the order *Chlamydiales*, proposal of *Parachlamydiaceae fam. nov.* and *Simkaniaceae fam. nov.*, each containing one monotypic genus, revised taxonomy of the family *Chlamydiaceae*, including a new genus and five new species, and standards for the identification of organisms. Int. J. Syst. Bacteriol. **49**:415–440.

FIJALKOWSKA, I. J., P. JONCZYK, M. M. TKACZYK, M. BIALOKORSKA, and R. M. SCHAAPER. 1998. Unequal fidelity of leading strand and lagging strand DNA replication on the *Escherichia coli* genome. Proc. Natl. Acad. Sci. USA **95**: 10020–10025.

FRANCINO, M. P., L. CHAO, M. A. RILEY, and H. OCHMAN. 1996. Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science **272**:107–109.

FRANCINO, M. P., and H. OCHMAN. 1997. Strand asymmetries in DNA evolution. Trends Genet. **13**:240–245.

FRANK, A. C., and J. R. LOBRY. 1999. Asymmetric patterns: a review of possible underlying mutational or selective mechanisms. Gene **238**:65–77.

GAUTIER, C. 2000. Compositional bias in DNA. Curr. Opin. Genet. Dev. **10**:656–661.

GOJOBORI, T., W.-H. LI, and D. GRAUR. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18**:360–369.

GONÇALVES, I., M. ROBINSON, G. PERRIERE, and D. MOUCHIROUD. 1999. JaDis: computing distances between nucleic acid sequences. Bioinformatics **15**:424–425.

GRIGORIEV, A. 1998. Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res. **26**:2286–2290.

HUTCHINSON, F. 1996. Mutagenesis. Pp. 2218–2235 *in* F. NEIDHARDT, R. CURTISS, J. L. INGRAHAM, E. C. LIN, K. B. LOW, B. MAGASANIK, W. S. REZNIKOFF, M. RILEY, M. SCHAECHTER, and H. E. UMBARGER, eds. *Escherichia coli* and *Salmonella*: cellular and molecular biology. ASM Press, Washington, D.C.

IWAKI, T., A. KAWAMURA, Y. ISHINO, K. KOHNO, Y. KANO, N. GOSHIMA, M. YARA, M. FURUSAWA, H. DOI, and F. IMAMOTO. 1996. Preferential replication-dependent mutagenesis in the lagging DNA strand in *Escherichia coli*. Mol. Gen. Genet. **251**:657–664.

IZUTA, S., J. D. ROBERTS, and T. A. KUNKEL. 1995. Replication error rates for G. dGTP, T.dGTP, and A.dGTP mispairs and evidence for differential proofreading by leading and lagging strand DNA replication complexes in human cells. J. Biol. Chem. **270**:2595–2600.

JELTSCH, A., and A. PINGOUD. 1996. Horizontal gene transfer contributes to the wide distribution and evolution of type II restriction-modification systems. J. Mol. Evol. **42**:91–96.

JONES, M., R. WAGNER, and M. RADMAN. 1987. Mismatch repair of deaminated 5-methyl-cytosine. J. Mol. Biol. **194**: 155–159.

KALMAN, S., W. MITCHELL, R. MARATHE, C. LAMMEL, J. FAN, R. W. HYMAN, L. OLINGER, J. GRIMWOOD, R. W. DAVIS, and R. S. STEPHENS. 1999. Comparative genomes of *Chlamydia pneumoniae* and *Chlamydia trachomatis*. Nat. Genet. **21**:385–389.

KUNST, F., N. OGASAWARA, I. MOSZER et al. (151 co-authors). 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. Nature **390**:249–256.

LAFAY, B., A. T. LLOYD, M. J. MCLEAN, K. M. DEVINE, P. M. SHARP, and K. H. WOLFE. 1999. Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res. **27**: 1642–1649.

LI, W.-H. 1997. Molecular evolution. Sinauer, Sunderland, Mass.

LINDAHL, T. 1993. Instability and decay of the primary structure of DNA. Nature **362**:709–715.

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

———. 1996. Asymetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. **13**:660–665.

LOPEZ, P., H. PHILIPPE, H. MYLLYKALLIO, and P. FORTERRE. 1999. Identification of putative chromosomal origins of replication in Archaea. Mol. Microbiol. **32**:883–886.

MCINERNEY, J. O. 1998. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc. Natl. Acad. Sci. USA **95**:10698–10703.

MACKIEWICZ, P., A. GIERLIK, M. KOWALCZUK, M. R. DUDEK, and S. CEBRAT. 1999. How does replication-associated mutational pressure influence amino acid composition of proteins? Genome Res. **9**:409–416.

MCLEAN, M. J., K. H. WOLFE, and K. M. DEVINE. 1998. Base composition skews, replication orientation and gene orientation in 12 prokaryote genomes. J. Mol. Evol. **47**:691–696.

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

MOSZER, I., E. P. C. ROCHA, and A. DANCHIN. 1999. Codon usage and lateral gene transfer in *Bacillus subtilis*. Curr. Opin. Microbiol. **2**:524–528.

MRÁZEK, J., and S. KARLIN. 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. USA **95**:3720–3725.

OCHMAN, H., J. G. LAWRENCE, and E. A. GROISMAN. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature **405**:299–304.

PERRIÈRE, G., J. R. LOBRY, and J. THIOULOUSE. 1996. Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acid sequences. CABIOS **12**:519–524.

RADMAN, M. 1998. DNA replication: one strand may be more equal. Proc. Natl. Acad. Sci. USA **95**:9718–9719.

READ, T. D., R. C. BRUNHAM, C. SHEN et al. (25 co-authors). 2000. Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. Nucleic Acids Res. **28**: 1397–1406.

ROCHA, E. P. C., A. DANCHIN, and A. VIARI. 1999*a*. Universal replication bias in bacteria. Mol. Microbiol. **32**:11–16.

———. 1999*b*. Functional and evolutionary roles of long repeats in prokaryotes. Res. Microbiol. **150**:725–733.

ROCHA, E. P. C., P. GUERDOUX-JAMET, I. MOSZER, A. VIARI, and A. DANCHIN. 2000. Implication of gene distribution in the bacterial chromosome for the bacterial cell factory. J. Biotechnol. **78**:209–219.

ROMERO, H., A. ZAVALA, and H. MUSTO. 2000. Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. Nucleic Acids Res. **28**:2084–2090.

SALZBERG, S. L., A. J. SALZBERG, A. R. KERLAVAGE, and J.-F. TOMB. 1998. Skewed oligomers and origins of replication. Gene **217**:57–67.

SHIGENOBU, S., H. WATANABE, M. HATTORI, Y. SAKAKI, and H. ISHIKAWA. 2000. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera sp.* APS. Nature **407**:81–86.

SHIRAI, M., H. HIRAKAWA, M. KIMOTO et al. (77 co-authors). 2000. Comparison of whole genome sequences of *Chlamydia pneumoniae* J138 from Japan and CWL029 from USA. Nucleic Acids Res. **28**:2311–2314.

SNIEGOWSKI, P. D., P. J. GERRISH, T. JOHNSON, and A. SHAVER. 2000. The evolution of mutation rates: separating causes from consequences. Bioessays **22**:1057–1066.

STEPHENS, R. S., S. KALMAN, C. LAMMEL et al. (12 co-authors). 1998. Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis.* Science **282**: 754–759.

SUEOKA, N. 1962. On the genetic basis of variation and heterogeneity of DNA base composition. Proc. Natl. Acad. Sci. USA **48**:582–591.

———. 1999. Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A=T and G=C. J. Mol. Evol. **49**:49–62.

TADDEI, F., I. MATIC, B. GODELLE, and M. RADMAN. 1997. To be a mutator, or how pathogenic and commensal bacteria can evolve rapidly. Trends Microbiol. **5**:427–429.

TAKAMI, H., K. NAKASONE, Y. TAKAKI et al. (12 co-authors). 2000. Complete genome sequence of the alkaliphilic bacterium *Bacillus halodurans* and genomic sequence comparison with Bacillus subtilis. Nucleic Acids Res. **28**:4317–4331.

TAO, H., C. BAUSCH, C. RICHMOND, F. R. BLATTNER, and T. CONWAY. 1999. Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. J. Bacteriol. **181**:6425–6440.

THOMPSON, J. D., D. G. HIGGINS, and T. J. GIBSON. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673–4680.

TILLIER, E. R., and R. A. COLLINS. 2000*a*. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. J. Mol. Evol. **50**:249–257.

———. 2000*b*. Replication orientation affects the rate and direction of bacterial gene evolution. J. Mol. Evol. **51**:459–463.

ZAR, J. H. 1996. Biostatistical analysis. 3rd edition. Prentice Hall, Upper Saddle River, N.J.

HOWARD OCHMAN, reviewing editor

Accepted June 5, 2001