J Mol Evol (2000) 50:249–257
DOI: 10.1007/s002399910029

JOURNAL OF MOLECULAR EVOLUTION

The Contributions of Replication Orientation, Gene Direction, and Signal Sequences to Base-Composition Asymmetries in Bacterial Genomes

Elisabeth R.M. Tillier, Richard A. Collins

Department of Molecular and Medical Genetics, University of Toronto, 1 King's College Circle, Toronto, Canada M5S 1A8

Received: 25 June 1999 / Accepted: 15 October 1999

Abstract. Asymmetries in base composition between the leading and the lagging strands have been observed previously in many prokaryotic genomes. Since a majority of genes is encoded on the leading strand in these genomes, previous analyses have not been able to determine the relative contribution to the base composition skews of replication processes and transcriptional and/or translational forces. Using qualitative graphical presentations and quantitative statistical analyses (analysis of variance), we have found that a significant proportion of the GC and AT skews can be attributed to replication orientation, i.e., the sequence of a gene is influenced by whether it is encoded on the leading or lagging strand. This effect of replication orientation on skews is independent of, and can be opposite in sign to, the effects of transcriptional or translational processes, such as selection for codon usage, amino acid preferences, expression levels (inferred from codon adaptation index), or potential short signal sequences (e.g., chi sequences). Mutational differences between the leading and the lagging strands are the most likely explanation for a significant proportion of the base composition skew in these bacterial genomes. The finding that base composition skews due to replication orientation are independent of those due to selection for function of the encoded protein may complicate the interpretation of phylogenetic relationships, conserved positions in nucleotide or amino acid sequence alignments, and codon usage patterns.

Key words: Base composition — Skews — Replication orientation — Gene direction — Prokaryote genomes — Bacterial genomes — Analysis of variance

Introduction

The genome sequences of several bacteria and viruses have revealed asymmetries in base composition and gene direction (Lobry 1996a; McClean et al. 1998; Mrazek and Karlin 1998; Freeman et al. 1998; Blattner et al. 1997). For example, approximately 75% of the genes of *Bacillus subtilis* are encoded on the DNA strand that is the template for the leading strand in replication (Kunst et al. 1997). This bias can surpass 80% in the *Mycoplasma* (Fraser et al. 1995). The leading strand itself has an excess of G nucleotides relative to C nucleotides and of T nucleotides relative to A nucleotides, as measured by the GC and AT skews given by (G − C)/(G + C) and (A − T)/(A + T), respectively. There does not appear to be a relationship between base composition skews and G + C content (Lobry 1996a; McClean et al. 1998). GC and AT skews were also found for *Escherichia coli*, *B. subtilis*, *Mycoplasma genitalium*, *Mycoplasma pneumonia*, *Haemophilus influenzae*, *Borrelia burgdorferi*, *Helicobacter pylori*, *Treponema pallidum*, and *Rickettsia prowazekii* (Lobry 1996a; McClean et al. 1998; Mrazek and Karlin 1998; Blattner et al. 1997; Andersson et al. 1998). Base composition skews have more recently been found in *Chlamydia trachomatis* and *Chlamydia pneumoniae* (this study).

Four main explanations have been proposed that can

Correspondence to: Elisabeth R.M. Tillier; *e-mail:* e.tillier@utoronto.ca

account for the skews in base composition. Two of these essentially attribute the skew to asymmetries in the gene direction. First, biases in amino acid preference and third-position codon choice (translational selection) can automatically generate skew in the nucleotide composition if the coding strands of genes are not equally distributed between the two strands of the chromosome. Second, mutational events during transcription (mutational bias) could generate a skew in two ways. There is evidence that the nontranscribed strand, which is single-stranded during the course of transcription, is more prone to deamination than the transcribed strand (Francino et al. 1996; Francino and Ochman 1997). Deamination leads to a net loss of C nucleotides on that strand (as well as a gain in T). Additionally, the nontranscribed strand does not benefit from transcription-coupled repair of deamination events, which occurs on the transcribed strand (Beletskii and Bhagwat 1996). The third explanation for base composition skews involves mutational events during replication, which affect the leading and lagging strands differently. An asymmetry can result if one strand incorporates more mutations of a particular type during replication or if one strand is more efficiently repaired (Lobry 1996b,c). The fourth explanation involves the different distribution of signal sequences on each strand (Lobry 1996a; Mrázek and Karlin 1998; Blattner et al. 1997; Salzberg et al. 1998). For example, chi sequences, which are G and T rich and involved in recombination, are found most often on the leading strand (Mrázek and Karlin 1998). In addition, Okazaki primer sequences could also contribute to a bias between the leading and the lagging strands if these sequences were biased (Blattner et al. 1997) or other signal sequences could also contribute (Salzberg et al. 1998).

Several of the mechanisms proposed may be acting to contribute to the strand compositional asymmetry. We examined the completely sequenced bacterial genomes using statistical methods to determine the individual contributions of replication, transcription, and signal sequences to base composition asymmetry.

Methods

Sequences

The annotated sequences of all publicly available, completely sequenced genomes were obtained from the NCBI Entrez genomes web site for the following bacteria: *E. coli* K-12 (Blattner et al. 1997), *B. subtilis* (Kunst et al. 1997), *M. genitalium* (Fraser et al. 1995), *M. pneumonia* (Himmelreich et al. 1996), *H. influenzae* Rd. (Fleishmann et al. 1995), *B. burgdorferi* (Fraser et al. 1997), *H. pylori* (Tomb et al. 1997), *T. pallidum* (Fraser et al. 1998), *C. trachomatis* (Stephens et al. 1998), *C. pneumoniae* (Kalman et al. 1999), *R. prowazekii* (Andersson et al. 1998), *Aquifex aeolicus* (Deckert et al. 1998), and *Synechocystis* sp. PCC6803 (Kaneko et al. 1996). Sequences were also obtained for the archeabacteria *Methanobacterium thermoautotrophicum* (Smith et al. 1997), *Archeoglobus fulgidus* (Klenk et al. 1997), *Methanococcus jannaschii* (Bult et al. 1996), and *Pyrococcus horikoshii* (Kawarabayasi et al. 1998). Genes were not differentiated as to whether they were known or putative.

Cumulative Graphs

To visualize the base composition bias in the bacterial genomes, we plotted the skew on a gene-by-gene basis and cumulatively along one strand of the chromosome. The starting point is that of the sequence as given in the database. Protein coding sequences (CDS) and non-CDS regions larger than 100 bases were analyzed. Figure 1 shows the cumulative graphs for several bacterial species of the GC and AT skews calculated at each of the three positions of the codons and in the non-CDS regions. To normalize for differences in gene length, the GC and AT skews were multiplied by the number of nucleotides involved in calculating the skews and divided by the total number of nucleotides over the sequence length considered. Also plotted is the accumulation of gene direction. Gene orientation was assigned +1 if the strand considered was the coding (nontranscribed) strand and −1 if the strand was the noncoding (transcribed) strand. For the plots, the gene direction was multiplied by the length of the gene and divided by the sum of lengths for all genes considered.

ANOVA of GC and AT Skews

All nucleotide positions in the genome were analyzed except when two genes overlapped, in which case only the first encountered was considered. The numbers of genes considered are given in Table 1. For non-CDS regions, only those longer than 100 nucleotides were included to reduce error due to small samples.

The origin and termination of replication have been either experimentally identified [*E. coli* (Marsh and Worcel 1977), *B. subtilis* (Ogasawara et al. 1984), *B. Burgdorferi* (Picardeau et al. 1999)] or identified by the position of the dnaA gene (Ogasawara and Yoshikawa 1992; Marczynski and Shapiro 1993), a change of sign in the overall GC skew (Lobry 1996b,c), or the asymmetry of octamers (Salzberg et al. 1998). The completely sequenced genomes for which the position or positions of the origin of replication have not yet been determined with any certainty were not analyzed (the archeabacterial genomes *A. fulgidus, M. jannaschii, M. autotrophicum,* and *P. horikoshii* and the bacterial genomes *A. aeolicus* and *Synechocystis*).

We considered the GC and AT skews along one strand of the chromosome on a gene-by-gene basis. For the CDS regions, each codon position was analyzed separately. Each gene along the strand of the chromosome was assigned a gene direction t (−1 if the strand considered is noncoding or +1 if it is the coding strand) and a replication orientation r(−1 if it lies on the lagging strand or +1 if on the leading strand). Two-factor (r and t, defined above), two-level (−1 or +1 for each r and t) ANOVAs were performed on these quantities. The ANOVA is unbalanced because of unequal sample sizes for the factor levels, which is due to the excess of genes on the leading strand. The analyses were performed using the computer program Minitab v. 12 (Minitab Inc., State College, PA) using the general linear model (GLM):

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk} \qquad (1)$$

where y_{ijk} is the value of the response variable (the skew) for the kth gene that has an observed level i of r and level j of t; μ is the hypothetical grand mean of the values of the skew for all the genes, irrespective of their gene or replication orientations; α_i is the hypothetical simple effect of level i of variable t ($t = -1$ for $i = 1$ and $t = +1$ for $i = 2$) on the mean of the values of the skew (i.e., on the mean of the values of the y_{ijk}); β_j is the hypothetical simple effect of level j of variable r($r = -1$ for $j = 1$ and $r = +1$ for $j = 2$) on the mean of the

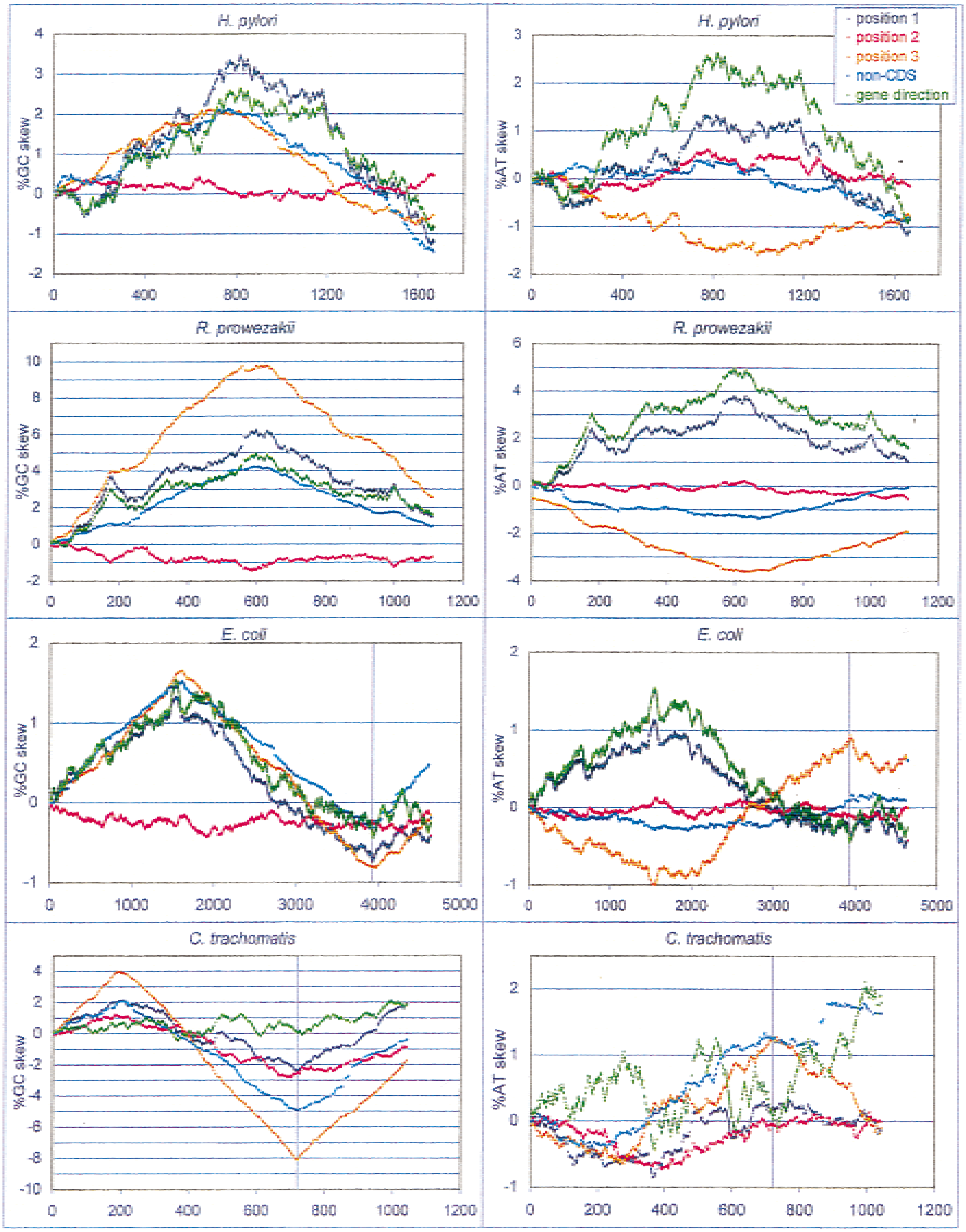

Fig. 1. Cumulative GC and AT graphs. The cumulative GC and AT skews (percentages) are plotted along one strand of the chromosome. The *X* axis indicates the position from the start of the sequence in thousands of bases. There is one plot for each position of the codon: blue for position 1, pink for position 2, orange for position 3, and light blue for non-CDS regions. The green plot corresponds to the cumulative gene direction (see text), which is divided by 6 to be on the same scale as the others. The origin of replication is approximately at base 1 in *H. pylori* and *R. prowazekii* and is indicated by a *vertical line* in *E. coli* and *C. trachomatis.*

values of the skew; γ_{ij} is the hypothetical interaction effect of concurrently giving an entity level *i* of variable *t* and level *j* of variable *r* on the mean of the values of the skew (independent of either of the above two simple effects); and ε_{ijk} is an error term reflecting the difference between the sum of the preceding four terms and the value of y_{ijk}.

Because of the complementary nature of DNA strands and the symmetrical nature of bidirectional replication and because there are approximately the same number of genes on both halves of the chromosomes, we would expect that there is no overall skew ($\mu = 0$). Also, we would expect that the skews attributable to transcription would be

Table 1. Contributions of transcription and replication orientations to GC and AT skews[a]

Genome	No. genes	% leading	Coefficient	GC1		GC2		GC3		GCS		AT1		AT2		AT3		ATS	
C. pneumonia	1052	54.0	α	17.7	*	−18.2	*	−3.0	*			12.7	*	−0.9		−9.7	*		
			β	**7.0**	*****	**6.6**	*****	**14.2**	*****	**8.8**	*****	**−2.9**	*****	**−0.6**		**−3.1**	*****	**−4.0**	*****
H. influenzae	1709	54.7	α	0.7		0.4		0.9				0.6		0.5		0.3			
			β	**3.6**	*****	**3.1**	*****	**3.2**	*****	**3.4**	*****	**−1.2**	*****	**−0.9**	*****	**−1.1**	*****	**−0.4**	
E. coli	4285	55.0	α	18.2	*	−10.7	*	3.9	*			23.4	*	−2.6	*	−17.0			
			β	**1.7**	*****	**1.0**	*****	**4.7**	*****	**4.2**	*****	**−0.5**		**0.2**		**−1.3**	*****	**−1.0**	*****
C. trachomatis	895	55.4	α	20.1	*	−18.5	*	3.4	*			9.3	*	−2.6	*	−10.6	*		
			β	**10.0**	*****	**8.1**	*****	**21.2**	*****	**13.3**	*****	**−2.0**	*****	**0.6**		**−2.4**	*****	**−2.1**	*****
H. pylori	1566	56.8	α	32.5	*	−8.0	*	−1.6	*			21.2	*	5.3	*	−8.7	*		
			β	**2.3**	*****	**1.9**	*****	**5.5**	*****	**4.8**	*****	**−0.1**		**0.2**		**−0.2**		**1.0**	
M. tuberculosis	3917	58.7	α	22.7	*	−14.8	*	−6.1	*			17.0	*	−9.1	*	−9.2	*		
			β	**1.6**	*****	**0.5**		**4.2**	*****	**3.1**	*****	**−1.3**	*****	**−0.6**		**−6.6**	*****	**−2.4**	*****
R. prowazekii	834	61.4	α	32.9	*	−17.0	*	14.6	*			28.6	*	1.8	*	−4.7	*		
			β	**2.8**	*****	**2.2**	*****	**13.7**	*****	**6.6**	*****	**−0.8**		**0.3**		**−2.9**	*****	**−2.9**	*****
T. pallidum	1031	65.3	α	18.0	*	−9.9	*	10.4	*			9.1	*	−6.7	*	−18.7	*		
			β	**4.8**	*****	**4.6**	*****	**16.6**	*****	**13.0**	*****	**−4.0**	*****	**−1.5**	*****	**−9.4**	*****	**−7.8**	*****
B. burgdorferi	849	66.2	α	38.5	*	−14.8	*	3.7	*			29.4	*	2.6	*	−2.2	*		
			β	**8.3**	*****	**9.4**	*****	**30.8**	*****	**21.5**	*****	**−8.3**	*****	**−3.9**	*****	**−20.2**	*****	**−11.1**	*****
B. subtilis	4099	73.6	α	27.6	*	−16.3	*	3.1	*			25.0	*	2.5	*	−1.7	*		
			β	**2.5**	*****	**1.9**	*****	**6.8**	*****	**8.7**	*****	**1.6**	*****	**1.6**	*****	**−0.7**	*****	**2.8**	*****
M. pneumoniae	675	79.9	α	23.1	*	−13.2	*	−9.3	*			23.2	*	6.4	*	−9.5	*		
			β	**1.4**		**−1.6**		**0.8**		**2.4**	*****	**0.3**		**0.8**		**−1.0**		**1.3**	
M. genitalium	350	82.2	α	28.2	*	−12.9	*	−4.8	*			28.6	*	5.3	*	−9.4	*		
			β	**0.7**		**−1.2**		**1.2**		**0.8**		**−0.6**		**−0.4**		**−1.3**		**1.2**	

[a] For the 12 genomes, the table presents the number of genes included in the analysis and the percentage of those genes lying on the leading strand. GC1, GC2, GC3, and GCS correspond to the percentage GC skew at first, second, and third codon positions and non-CDS (spacer) sequences, respectively. AT1, AT2, AT3, and ATS indicate the corresponding AT skews. The estimated coefficients in the regression of the GC and AT skews [Eq. (2)] are given for gene direction (α) and replication orientation (β). An asterisk indicates statistical significance at the 0.01 level.

of the same magnitude when considering the noncoding or coding strands ($\alpha_1 = -\alpha_2$) and that the skews attributable to transcription would be of the same magnitude when considering the leading or lagging strands ($\beta_1 = -\beta_2$). We do find that these values are not statistically different from this expectation (data not shown). We can thus describe the skews as fitting the equation

$$y_{ijk} = \alpha t_i + \beta r_j + \gamma(r_i t_j) + \varepsilon_{ijk} \qquad (2)$$

where $\alpha = -\alpha_1$, $\beta = -\beta_1$, $t = -1$ if $i = 1$, $t = 1$ if $i = 2$, $r = -1$ if $j = 1$, $r = -1$ if $j = 2$. The values of the parameters and their significance were obtained using the GLM in Eq. (1), and Eq. (2) is given only to simplify the explanation and discussion of the effects of gene direction and replication orientation on the base composition skews. α and β are estimates of the respective effects of gene direction and replication orientation on the base composition skews.

Signal Sequences

The potential chi signal sequence (5′ GCTGGTGG 3′) and its complement were removed from the original sequence by replacing the eight bases with X's. In a separate analysis, we eliminated all potentially skewed octamers identified by Salzberg et al. (1998) by replacing the octamers with eight X's in the sequence. Nucleotides coded as X were ignored in the calculation of the base composition skews.

Codon Adaptation Index

The codon adaptation index (CAI) measures the degree to which a gene uses the same codons as those in highly expressed genes and is thought to be a measure of the degree to which a gene is expressed (Sharp and Li 1987). The relative use of each codon in highly expressed genes has been tabulated for *E. coli, B. subtilis, H. influenzae,* and *M. tuberculosis* (Sharp and Li 1987; Shields and Sharp 1987; Pan et al. 1998). These were used to calculate the CAI values for all the genes considered in these organisms.

Results and Discussion

Cumulative Plots

The cumulative plots allow for a fast visual examination of the base composition skew at the different codon positions of the genome (Fig. 1) (see Freeman et al. 1998; Grigoriev 1998; Cebrat et al. 1999; Mackiewicz et al. 1999). Plotting the skew cumulatively along the genome shows a change in the direction of the slope with a change in sign of the skew and the quantity and quality of the skew can be assessed from the V or inverted-V shape of the curve. Cumulative plots have the effect of reducing the noise in the data and make inversions in the polarity of the skew readily apparent, without the need to average over large sequence windows. The genomes presented in Fig. 1 are some of the most recently available, as well as the most-studied organism, *E. coli.* These plots and equivalent ones for additional genomes can be found at the URL: http://medgen4285.med.utoronto.ca/medgen/collins.htm.

Most genomes examined show some degree of base composition asymmetry. No asymmetries were readily visible in the graphs of the cyanobacterium *Synechocystis* sp., the bacterium *A. aeolicus,* or the archaebacteria *A. fulgidus* and *M. jannaschii.* The archaebacterium *P. horikoshii* does present a trend in its AT skew at third codon positions, but this is not reflected in any of the other curves.

From the plots, we can see that the three positions of the codons can behave quite differently. The skews are most pronounced at the first and third positions of codons, indicating amino acid biases and biases dependent on codon preference. The plots corresponding to non-CDS regions also show GC and, in some cases, AT skews. Skews in non-CDS sequences indicate that amino acid and codon choice cannot be entirely responsible for the skews. Transcriptional explanations for skews in non-CDS regions cannot be eliminated because many of these regions are transcribed, even if not translated.

A plot of the cumulative gene direction along the genome in these graphs (green lines) reveals that there is an excess of genes on the leading strand causing a gene direction bias in these bacteria. Interestingly, the two *Chlamydia* species do not show a gene direction bias in such a plot because there are many stretches of genes on the lagging strand, even though overall these species have a slight excess of genes on the leading strand. That the gene direction bias often follows base composition skews is evident for many of the curves. For example, in *R. prowazekii,* every rise and fall of the AT skew at the first codon position (blue line) mirror the rise and fall of the gene direction curve. This indicates that the sign of the AT skew at that position is dependent upon gene direction. Another example is the third position GC skew in *C. trachomatis* (orange line). It presents a very strong and regular pattern, with a V at the origin of replication and an inverted V about halfway around the chromosome. This pattern is not apparent in the gene direction curve, indicating that the skew is better explained by replication orientation than gene direction in that case.

ANOVA

The cumulative plots in Fig. 1 provided a qualitative overview of the GC and AT skews, but the graphical analysis is not enough to ascertain to what degree these skews are correlated with replication orientation or gene direction. Linear discriminant analysis was used by Rocha et al. (1999a) and Perriere et al. (1996) to consider the effect of replication on base composition. Here, ANOVA analyses were used to quantify the extent of the individual effects of replication and gene direction on the skews and allowed us to determine the level of statistical significance of these effects.

Table 1 gives the estimated values of the coefficients α and β in Eq. (2), which provide a measure of the extent to which the skews correlate with either gene direction or replication orientation, respectively. The GC and AT skews at first, second, and third codon positions and non-CDS (spacer) sequences were analyzed separately. Those values of α and β followed by an asterisk are significantly different from zero at the $p = 0.01$ level or better. The coefficient γ in Eq. (2), which measures the potential additive effects of replication orientation and gene direction, was very small and usually not statistically significant (data not shown). In most cases the values of the α and β coefficients were significant, indicating that the replication and gene direction effects each explain a significant portion of the GC and AT skews. Both gene direction and replication orientation independently affect the base composition skews. These effects can be considered separately (below).

Gene Direction

Most bacteria show skews correlating with gene direction (Table 1). These skews vary in sign and intensity in the different bacterial sequences and among codon positions. AT skews with gene direction are mostly significant but can vary in sign at the first two positions and are usually negative at the third position. With respect to the GC skew at the first two positions of codons, the first positions prefer G to C, and the second positions C to G, in all these bacteria as has been noted previously (Andersson and Kurland 1990). The GC skew at third positions varies in sign (see below).

The ANOVA shows statistical significance with gene direction in most cases. This gene direction effect could be due to either translational selection or transcription-level mutational biases. A mutational mechanism involving the deamination of C (leading to a C → T mutation) on the coding strand would lead to a positive GC skew and a negative AT skew with respect to coding orientation. *B. burgdorferi, E. coli, R. prowazekii, T. pallidum,* and *B. subtilis* have a positive GC skew at the third position, which is consistent with a mutational bias. *H. pylori, M. tuberculosis, M. genital, C. pneumoniae,* and *M. pneumoniae* have a negative GC skew at the third position, however, indicating that this deamination mechanism cannot be responsible for the skew in the latter group of bacteria. The different direction of the skew may be due to a different mutational bias or different translational selection in different organisms.

Replication Orientation

There are many examples of significant correlations of base composition biases with replication orientation (Table 1, boldface values). In general the AT skews are negative, smaller in magnitude than the GC skews, and, in many cases, not statistically significant. These results are consistent with those from the linear discriminant analysis performed on the base composition of these genomes (Rocha et al. 1999a). *B. burgdorferi* has very large AT and GC skews at the third codon position, correlating with replication. The effect of replication on base composition biases has been independently inferred by McInerney (1998) from an analysis of codon usage patterns in *B. burgdorferi* which revealed a preference for T or G at third positions of codons on the leading strand. *H. pylori, H. influenzae, M. pneumoniae,* and *M. genitalium* show no significant AT skew correlated with replication. *B. subtilis* has a small but significantly positive AT skew at the first and second codon positions and in the non-CDS regions. AT skews in the other bacteria are usually small and negatively correlated with replication orientation.

Almost all the bacteria show a positive correlation between GC skew and replication orientation at all three codon positions and in non-CDS regions (i.e., $\beta > 0$ in Table 1). Only in the *Mycoplasma* are the GC skews nonsignificant. *M. genitalium* is exceptional in that it does not show any significant GC skew with replication orientation, although there is a significant skew with gene direction. This bacterium has the fewest genes and a high degree of correlation between coding orientation and replication orientation, such that no bias can be attributed solely to replication orientation. The coefficients for the GC skews in the non-*Mycoplasma* bacteria show that they are all significantly positively correlated with replication orientation even when they are significantly negatively correlated with gene direction. This occurs with *H. pylori* and *M. tuberculosis* at the third position, for example, and in all the bacterial genomes at the second position. The results thus show that even when there is an overall strong preference for C on the coding strand of genes, a replication bias for G on the leading strand can still be significant.

H. influenzae is the only bacterium that does not have any AT or GC skew significantly attributable to gene direction, while it has significant GC skews correlated with replication orientation. *H. influenzae* shows that replication orientation can have an even larger effect than selection for codon usage patterns. *H. influenzae* thus illustrates the importance of considering replication orientation when analyzing codon usage patterns.

Non-CDS

We also calculated the skew in non-protein-coding regions of the genome to determine if base composition asymmetries were apparent in the regions that are not subject to amino acid coding constraints. The non-CDS sequences contain rRNA, tRNA, promoters, attenuators, and other functional as well as nonfunctional DNA. Additionally, even if most of these sequences are transcribed, the transcription direction is not necessarily known. Given these caveats and assuming that no unidentified protein-coding genes or pseudogenes are present in these regions, we can nevertheless assert that any bias observed in these regions cannot be attributed to codon usage. We performed one-way ANOVAs with respect to the orientation of replication *(r)* for these sequences. We found large, positive correlations between GC skews and replication orientation and nonsignificant or small negative correlations of AT skews with replication orientation (GCS and ATS in Table 1). *B. subtilis* presents a significantly positive AT skew in non-CDS regions, possibly indicating some level of selection in these regions that is not found in the other organisms. Usually, the AT and GC skews in non-CDS regions are similar in size and sign to the skews seen with replication orientation at the third positions of codons, yet the non-CDS skew is not attributable to codon usage.

Signal Sequences

It is possible that the GC and AT skews could be caused by an excess of chi or other signal sequences on the leading strand (Lobry 1996a; Mrázek and Karlin 1998; Salzberg et al. 1998). In *E. coli,* the chi signal sequence GCTGGTGG is used in recombination, and of approximately 1000 such sites in the *E. coli* genome, more than three-quarters are found on the leading strand (Blattner et al. 1997). We performed an ANOVA of the *E. coli* genome where all chi sequences were ignored for the calculation of GC and AT skews. The results of these ANOVAs were not significantly different from the results shown in Table 1, where chi sequences were included (data not shown). The biases in the orientation of solely the chi sequences therefore do not explain the GC and AT bias in *E. coli.*

Another potential for bias between the leading and the lagging strands could come from biases in unknown signal sequences. Skews in oligonucleotides of short length have been found in prokaryotic genomes (Salzberg et al. 1998). We removed all skewed octamer sequences from the genome sequences of *E. coli* and from *H. pylori.* This led to an exclusion of about 10% of all nucleotides from the analysis in *E. coli* (including the chi sequences) and 6% for *H. pylori.* The ANOVAs (shown for *E. coli* in Table 2) reveal that although asymmetries were reduced, they were not eliminated. There is still the potential that many smaller signal sequences are causing the asymmetries, but skewed octamers are not the sole source of the GC and AT skews.

Relationship Between Expression Levels and Base Composition Skew

Although there are significant correlations between base composition skew and replication orientation, this does not necessarily imply that the skews are caused by replication events. In fact, a possible explanation for the

Table 2. Effects of skewed octamers and CAI values on GC and AT skews[a]

Sequence	Coefficient	GC1		GC2		GC3		GCS		AT1		AT2		AT3		ATS	
E. coli	α	18.2	*	−10.7	*	3.9	*			23.4	*	−2.6	*	−17.0	*		
	β	**1.7**	*	**1.0**	*	**4.7**	*	**4.2**	*	**−0.5**		**0.2**		**−1.3**	*	**−1.0**	*
E. coli—no skewed octamers	α	17.0	*	−11.1	*	0.4				23.1	*	−1.0	*	−16.7	*		
	β	**0.7**	*	**0.0**		**2.8**	*	**2.8**	*	**−0.3**		**0.2**		**−1.3**	*	**−1.4**	*
E. coli—0.3 < CAI < 0.5	α	16.8	*	−10.7	*	6.6	*			21.2	*	−4.1	*	−16.1	*		
	β	**1.9**	*	**1.3**	*	**5.4**	*	**N/A**		**−1.0**	*	**0.1**		**−1.0**	*	**N/A**	

[a] The coefficients of correlation with transcription orientation and replication orientation are given for the GC and AT skews at the three codon positions and in non-CDS regions (see Table 1, footnote a). Results using the complete *E. coli* sequence are compared to results when all skewed octamers are removed from the sequence or when the analysis is restricted to genes with intermediate CAI values. An asterisk indicates values significant at the 0.01 level.

correlations is that these are a consequence of selection for gene regulation purposes. Correlations between gene regulation and gene direction could potentially lead to apparent correlations between base composition and replication orientation. Indeed, highly transcribed genes such as ribosomal proteins are preferentially located on the leading strand (Sharp and Li 1987) and there can be a preference for G nucleotides at the third position (Pan et al. 1998; this study). If genes that are not highly expressed are located on the lagging strand and have an excess of C, this would lead to a bias for G on the leading strand and for C on the lagging strand. Such biases would not have a mutational origin but would reflect selection for coding orientation for gene regulation purposes. This effect may lead to an overestimate of the contribution of replication in causing the observed bias (and an underestimate of that of transcription).

However, the data do not support the idea that differences in transcription levels are a major cause of apparently replication-linked bias. Significant GC skews are observed in organisms displaying small correlations between gene direction and replication (e.g., *E. coli, C. pneumoniae,* and *C. trachomatis*). Noncoding regions also show significant skews (as discussed above). Additionally, we examined the relationship between CAI values and replication orientation in *E. coli, B. subtilis, M. tuberculosis,* and *H. influenzae.* The CAI value for a gene is a measure of its use of optimal codons (i.e., the codons used in the most highly expressed genes) and therefore is thought to reflect the level of expression of that gene (Sharp and Li 1987). We found that the distribution of the CAI Index is similar on both strands (leading and lagging), except that there is an excess of genes associated with the highest CAI values that are located on the leading strand (data not shown). This excess of the most highly expressed genes on the leading strand is not sufficiently large to produce a significant correlation between replication orientation and CAI values, as we found no significant difference between the CAI values for genes on the leading and those for genes on the lagging strands. ANOVAs of the CAI index for each of these four bacteria were nonsignificant with respect to replication orientation. (The *p* values were 0.93 for *E. coli,* 0.74 for *H. influenzae,* 0.83 for *M. tuberculosis,* and 0.99 for *B. subtilis.*

ANOVAs with gene direction and replication orientation were performed as before, with the exclusion of those genes with the lowest and highest CAI values (less than 0.3 and greater than 0.5). These did not give significantly different results than when these genes were included (Table 2), as the skews were still significantly correlated with replication orientation. Genes with the highest and lowest CAI do not account for the correlation of base composition skews with replication orientation and therefore these skews are not completely explained by the selection of highly expressed genes on the leading strand.

Conclusion

Unlike previous graphical analyses (Fig. 1) (Cebrat et al., 1999; Mackiewicz et al. 1999) and linear discriminant analysis methods (Rocha et al. 1999a; Perriere et al. 1996; for discussion see Karlin 1999; Rocha et al., 1999b), the ANOVA allows the quantification of the relative effects of replication orientation and gene direction to the base composition asymmetries and of their statistical confidence. A large portion of the GC and AT skews can be attributed to translational selection, such as a great preference for G at the first position of codons and T at the third position. When combined with an excess of genes on the leading strand, base composition skews will appear to correlate with replication orientation. The ANOVA can separate the effects of replication and gene direction and we find that a portion of the GC and AT skews is significantly attributable to replication orientation, independent of any transcriptional or translational bias. The correlation of the skews with replication orientation cannot be explained by coding orientation biases, expression-level biases, or biases in chi sequences or in other potential signal sequences and, therefore, are indicative of a bias between the leading and the lagging strands.

The most likely explanation for base composition differences between the leading and the lagging strands is a mutational difference between these two strands. These mutational differences could occur in many ways. The templates for the leading and lagging strands may be differentially damaged. Damage on the DNA templates could also be differentially repaired. There could also be a different rate of misincorporation of nucleotides in the leading and lagging strands and/or of differential repair of mutations on the two strands.

The lack of asymmetry observed in several of the prokaryotes considered (*Synechocystis* and the archaebacteria) may be due to the presence of multiple origins of replication, different methods of replication (i.e., rolling circle vs bidirectional), better protection of the DNA from damage, different polymerase actions, or better mutation repair mechanisms.

Bias mutation pressures have been shown to affect sequence evolution (for a review see Li 1997). Whatever the cause of the base compositional biases, for which we find a significant correlation with replication orientation, it will have consequences for sequence analysis. For example, biases in base composition can cause sequences to appear similar when they are not phylogenetically closely related or functionally important, and conversely, the inversion of a gene onto the opposite strand may lead to a very fast sequence divergence. Phylogenetic and functional analysis methods may need to be modified to consider the significant effect of replication orientation on both nucleotide and amino acid composition in bacterial genomes.

Acknowledgments. We thank J.R. Lobry and another reviewer for their helpful comments on the manuscript. R.A.C. is a fellow of the Canadian Institute for Advanced Research (CIAR). This work was funded by the National Sciences and Engineering Council (NSERC) grant to R.A.C.

References

Andersson SG, Kurland CG (1990). Codon preferences in free-living organisms. Microbiol Rev 54:98–210

Andersson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. Nature 396:133–140

Beletskii A, Bhagwat A (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

Blattner FR, Plunkett G, Bloch CA, et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277:1453–1474

Bult CJ, White O, Olsen GJ, et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschil.* Science 273:1058–1073

Cebrat S, Dudek MR, Gierlik A, Kowalcczuk M, Mackiewicz P (1999) Effect of replication on the third base of codons. Physica A 265: 78–84

Deckert G, Warren PV, Gaasterland T, Young WG, Lenox AL, Graham DE, Overbeek R, Snead MA, Keller M, Aujay M (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus.* Nature 392:353–358

Fleischmann RD, Adams MD, White O, et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512

Francino MP, Ochman H (1997) Strand asymmetries in DNA evolution. Trends Genet 13:240–245

Francino MP, Chao L, Riley MA, Ochman H (1996) Asymmetries Generated by transcription-coupled repair in enterobacterial genes. Science 272:107–109

Fraser CM, Gocayne JD, White O, et al. (1995) The minimal gene complement of *Mycoplasma genitalium.* Science 270:397–403

Fraser CM, Casjens S, Huang WM, et al. (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi.* Nature 390:580–586

Fraser CM, Norris SJ, Weinstock GM, et al. (1998) Complete genome sequence of Treponema pallidum, the syphilis spirochete. Science 281:375–388

Freeman JM, Plasterer TN, Smith TF, Mohr SC (1998) Patterns of genome organization in bacteria. Science 279:1827a

Grigoriev A (1998) Analysing genomes with cumulative skew diagrams. Nucleic Acids Res 26:2286–2290

Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li BC, Hermann R (1996) Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae.* Nucleic Acids Res 24:4420–4449

Kalman S, Mitchell W, Marathe R, Lammel C, Fan J, Olinger L, Grimwood J, Davis RW, Stephens RS (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis.* NATURE Genetics 21:385–389

Kaneko T, Sato S, Kotani H, et al. (1996) Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res 3:185–209

Karlin S. (1999) Bacterial DNA strand asymmetry. Trends Microbiol 7:305–308

Kawarabayasi Y, Sawada M, Horikawa H, et al. (1998) Complete sequence and gene organization of the genome of a hyperthermophilic archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res 5:55–76

Klenk HP, Clayton RA, Tomb JF, et al. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus.* Nature 390:364–370

Kunst F, Ogasawara N, Moszer I, et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis.* Nature 390:249–256

Li, W-H. (1997) Molecular evolution. Sinauer Associates, Sunderland, MA

Lobry JR (1996a) Asymmetric substitution patterns in the two DNA strand of bacteria. Mol Biol Evol 13:660–665

Lobry JR (1996b) A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. Biochimie 78: 323–326

Lobry JR (1996c) Origin of replication of *Mycoplasma genitalium.* Science 272:745–746

Mackiewicz P, Gielik A, Kowalczuk M, Dudek M, Cebrat S (1999) How does replication-associated mutational pressure influence amino acid composition of proteins. Genome Res 9:409–416

Marczynski G, Shapiro L (1993) Bacterial chromosome origins of replication. Curr Opin Genet Dev 3:775–782

Marsh R, Worcel A (1977) A DNA fragment containing the origin of replication of the *Escherichia coli* chromosome. Proc Natl Acad Sci USA 74:2720–2724

Mclnemey JO (1998) Replication and transcriptional selection on codon usage in *Borrelia burgdorfori.* Proc Natl Acad Sci USA 95:10698–10703

McLean M, Wolfe KH, Devine KM (1998) Base composition skews, replication orientation and gene orientation in 12 prokaryotic genomes. J Mol Evol 47:691–696

Mrázek J, Karlin S (1998) Strand compositional asymmetry in bacterial and large viral genomes. Proc Natl Acad Sci USA 95:3720–3725

Ogasawara N, Yoshikawa H (1992) Genes and their organization in the replication origin region of the bacterial chromosome. Mol Microbiol 6:629–634

Ogasawara N, Mizumoto S, Yoshikawa H (1984) Replication origin of the *Bacillus subtilis* chromosome determined by hybridization of the first-replicating DNA with cloned fragments from the replication region of the chromosome. Gene 30:173–182

Pan A, Dutta C, Das J (1998) Codon usage in highly expressed genes of *Haemophillus influenzae* and *Mycobacterium tuberculosis:* Translational selection versus mutational bias. Gene 215:405–413

Perriere G, Lobry JR, Thioulouse J (1996) Correspondence discriminant analysis: A multivariate method for comparing classes of protein and nucleic acid sequences. Comput Appl Biosci 12:519–524

Picardeau M, Lobry JR, Hinnebusch BJ (1999) Physical mapping of an origin of bidirectional replication at the centre of the Borrelia burgdorferi linear chromosome. Mol Microbiol 32:437–445

Rocha EPC, Danchin, Vlari A (1999a) Universal replication biases in bacteria. Mol Microbiol 32:11–16

Rocha EPC, Danchin, Viari A (1999b) Bacterial DNA strand compositional asymmetry: Response. Trends Microbiol 7:308

Salzberg SL, Salzberg AJ, Kerlavage AR, Tomb J-F (1998) Skewed oligomers and origins of replication. Gene 217:57–67

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

Shields DC, Sharp PM (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acids Res 15:8023–8040

Smith DR, Doucette-Stamm LA, Deloughery C, et al. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: Functional analysis and comparative genomics. J Bacteriol 179:7135–7155

Stephens RS, Kalman S, Lammel C, et al. (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis.* Science 282:754–759

Tomb JF, White O, Kerlavage AR, et al. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori.* Nature 388: 539–547