Journal of Theoretical Biology 253 (2008) 508–513

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: New substitution models incorporating strand bias

Antonio Marín[a], Xuhua Xia[b,*]

[a] *Departamento de Genética, Universidad de Sevilla, Avenida Reina Mercedes 6, E-41012 Sevilla, Spain*

[b] *Department of Biology and Center for Advanced Research in Environmental Genomics, University of Ottawa, 30 Marie Curie, P.O. Box 450, Station A, Ottawa, Ontario, Canada K1N 6N5*

ARTICLE INFO

Article history:
Received 5 November 2007
Received in revised form
29 February 2008
Accepted 4 April 2008
Available online 11 April 2008

Keywords:
Strand bias
DNA replication
GC skew
Genomic evolution

ABSTRACT

The DNA strands in most prokaryotic genomes experience strand-biased spontaneous mutation, especially C→T mutations produced by deamination that occur preferentially in the leading strand. This has often been invoked to account for the asymmetry in nucleotide composition, typically measured by GC skew, between the leading and the lagging strand. Casting such strand asymmetry in the framework of a nucleotide substitution model is important for understanding genomic evolution and phylogenetic reconstruction. We present a substitution model showing that the increased C→T mutation will lead to positive GC skew in one strand but negative GC skew in the other, with greater C→T mutation pressure associated with greater differences in GC skew between the leading and the lagging strand. However, the model based on mutation bias alone does not predict any positive correlation in GC skew between the leading and lagging strands. We computed GC skew for coding sequences collinear with the leading and lagging strands across 339 prokaryotic genomes and found a strong and positive correlation in GC skew between the two strands. We show that the observed positive correlation can be satisfactorily explained by an improved substitution model with one additional parameter incorporating a general trend of C avoidance.

1. Introduction

Many studies have documented strand asymmetry in eubacterial genomes associated with their single-origin mode of genome replication (Frank and Lobry, 1999; Karlin, 1999; Lobry, 1996; Lobry and Sueoka, 2002; Rocha et al., 1999). In general, there is an excess of (*G*+*T*) in the leading strand and an excess of (*A*+*C*) in the lagging strand in many prokaryotic genomes examined (Francino and Ochman, 1997; Freeman et al., 1998; Grigoriev, 1998; McLean et al., 1998; Perriere et al., 1996), with the bias generally attributed to strand-biased deamination (Frank and Lobry, 1999; Frederico et al., 1990; Lindahl, 1993; Lobry and Sueoka, 2002; Sancar and Sancar, 1988). The strand compositional asymmetry is strong enough to identify the location of the bacterial origin of replication whose flanking sequences change direction in GC skew (Frank and Lobry, 2000; Green et al., 2003; Lobry, 1996; Worning et al., 2006; Zhang and Li, 2003; Zhang and Zhang, 2003). GC skew correlates with the distribution of inverted repeats (Achaz et al., 2003) and essential genes (Rocha and Danchin, 2003) and associates with amino acid composition (Mackiewicz et al., 1999).

Because of the difficulty in identifying strand affiliation of individual genes during its evolutionary history, it is difficult to study the effect of strand bias based on conventional methods using observed substitution patterns. Instead, within-genome indices have been developed to characterize strand bias (Lobry, 1996; Morton and Morton, 2007). One simple index to measure the strand bias in nucleotide composition in a genome is the GC skew (Lobry, 1996) which now exists in two versions differing only in sign, one being (*C*−*G*)/(*C*+*G*) (Fujimori et al., 2005; Lobry, 1996) and the other being (*G*−*C*)/(*G*+*C*) (Blattner et al., 1997; Chambaud et al., 2001; Contursi et al., 2004; Grigoriev, 1998), where *C* and *G* designate the number of nucleotides cytosine and guanine, respectively. To avoid confusion, we explicitly define

$$A_C = \frac{G-C}{G+C} \tag{1}$$

We further designate $A_{C.LE}$ and $A_{C.LA}$ as A_C for leading and lagging strands, respectively. In general, $A_{C.LE} > A_{C.LA}$ (Lobry, 1996; Lobry and Sueoka, 2002), and a number of contributing factors involving specific types of mutation and selection have been proposed (Frank and Lobry, 1999) and quantitatively assessed (Morton and Morton, 2007).

[*] Corresponding author. Tel.: +1 613 562 5800x6886; fax: +1 613 562 5486.
E-mail addresses: anmarin@us.es (A. Marín), xxia@uottawa.ca (X. Xia).

0022-5193/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2008.04.004

Casting the strand bias in the framework of a nucleotide substitution model is important for our understanding of genomic evolution and phylogenetic reconstruction in prokaryotic genomes because none of the existing substitution models for phylogenetic reconstruction has taken the strand-biased substitution into consideration. We present substitution models showing that an increased C→T mutation pressure on the leading strand will lead to a positive A_C in the leading strand and a negative A_C in the lagging strand. Greater C→T mutation pressure is associated with greater differences in A_C between the leading and the lagging strands. We further demonstrate that empirical results of $A_{C.LE}$ and $A_{C.LA}$ are inconsistent with the substitution model invoking the strand-biased C→T mutation only and require an alternative substitution model incorporating a tendency toward C avoidance/shortage in coding sequences.

We define the vector of the four nucleotide frequencies, $P(t)$, and the transition probability matrices for the leading and the lagging strand (designated by M_{LE} and M_{LA}, respectively), as

$$P(t) = [P_A(t)\, P_G(t)\, P_C(t)\, P_T(t)] \tag{2}$$

$$M_{LE} = \left[\begin{array}{ccccc} & A & G & C & T \\ A & \bullet & a_1 & a_2 & a_3 \\ G & a_1 & \bullet & a_4 & a_5 \\ C & a_2 & a_4 & \bullet & a_6 + x \\ T & a_3 & a_5 & a_6 & \bullet \end{array}\right] \tag{3}$$

$$M_{LA} = \left[\begin{array}{ccccc} & A & G & C & T \\ A & \bullet & a_1 & a_2 & a_3 \\ G & a_1 + x & \bullet & a_4 & a_5 \\ C & a_2 & a_4 & \bullet & a_6 \\ T & a_3 & a_5 & a_6 & \bullet \end{array}\right] \tag{4}$$

where a_i values are the transition probabilities and the diagonal elements of M_{LE} and M_{LA} are subjected to the constraint of each row sum equal to 1. The symbol x in matrix M_{LE} and M_{LA} indicates the increased probability of C→T transitions in the leading strand and the consequently increased probability of G→A transitions on the lagging strand. It is positive, can vary across genomes, and may be substantially larger than a_1 or a_6 as indicated in previous studies on bacterial genomes (Lobry, 1996; Lobry and Sueoka, 2002; McInerney, 1998), vertebrate mitochondrial genomes (Reyes et al., 1998; Tanaka and Ozawa, 1994; Xia, 2005; Xia et al., 2006) and viral genomes (Xia and Yuen, 2005). If $x = 0$, then the two transition probability matrices in Eqs. (3) and (4) are symmetrical, and the equilibrium nucleotide frequencies will be all equal to 1/4. In the terminology of Morton and Morton (2007), the parameter x represents the replication-dependent effect that differ between the leading and lagging strands.

The dynamic behavior of the Markov chain specified in Eqs. (3) and (4) follows the equation below:

$$P(t+1) = P(t)M \tag{5}$$

To obtain equilibrium frequencies of $P(t)$, which is conventionally designated as π_i (where $i = 1, 2, 3, 4$ corresponding to the four nucleotides), we solve Eq. (5) by setting $P(t{+}1) = P(t)$ and imposing the constraint of $\Sigma P(t) = 1$. This yields the equilibrium frequencies for the leading and lagging strands:

$$\begin{aligned} \pi_{A.LE} &= \frac{(a_1a_3 + a_1a_5 + a_3a_4 + a_3a_5)x + C_1}{C_2x + 4C_1} \\ \pi_{G.LE} &= \frac{(a_1a_3 + a_1a_5 + a_2a_5 + a_3a_5)x + C_1}{C_2x + 4C_1} \\ \pi_{C.LE} &= \frac{C_1}{C_2x + 4C_1} \\ \pi_{T.LE} &= \frac{a_1a_3 + a_1a_5 + a_2a_5 + a_3a_5 + a_1a_2 + a_1a_4 + a_2a_4 + a_3a_4)x + C_1}{C_2x + 4C_1} \\ C_1 &= a_1a_2a_3 + a_1a_2a_5 + a_1a_2a_6 + a_1a_3a_4 + a_1a_3a_6 + a_1a_4a_5 + a_1a_4a_6 + a_1a_5a_6 \\ &\quad + a_2a_3a_4 + a_2a_3a_5 + a_2a_4a_5 + a_2a_4a_6 + a_2a_5a_6 \\ &\quad + a_3a_4a_5 + a_3a_4a_6 + a_3a_5a_6 \\ C_2 &= 3(a_1a_3 + a_1a_5 + a_3a_5) + 2(a_2a_5 + a_3a_4) + a_1a_2 + a_1a_4 + a_2a_4 \end{aligned} \tag{6}$$

$$\begin{aligned} \pi_{C.LA} &= \frac{(a_2a_3 + a_2a_5 + a_2a_6 + a_3a_6)x + C_1}{C_0x + 4C_1} \\ \pi_{T.LA} &= \frac{(a_1a_3 + a_1a_5 + a_3a_4 + a_3a_5)x + C_1}{C_0x + 4C_1} \\ \pi_{A.LA} &= \frac{(a_2a_3 + a_3a_6 + a_2a_5 + a_2a_6 + a_5a_4 + a_4a_3 + a_4a_6 + a_5a_6)x + C_1}{C_0x + 4C_1} \\ \pi_{G.LA} &= \frac{C_1}{C_0x + 4C_1} \\ C_0 &= 3(a_3a_6 + a_2a_3 + a_2a_6) + 2(a_4a_3 + a_5a_2) + a_4a_6 + a_5a_6 + a_5a_4 \end{aligned} \tag{7}$$

where C_1 and C_2 are specified in Eq. (6). Note that $\Sigma\pi_i = 1$. From the equilibrium frequencies above, we can obtain A_C for leading and lagging strands

$$\begin{aligned} A_{C.LE} &= \frac{Z_1x}{Z_2 + Z_1x} \\ Z_1 &= a_2a_5 + a_1a_3 + a_1a_5 + a_3a_5 \\ Z_2 &= 2(a_1a_2a_6 + a_4a_3a_2 + a_4a_3a_6 + a_4a_2a_5 + a_1a_6a_5 \\ &\quad + a_1a_3a_6 + a_1a_4a_3 + a_1a_2a_5 \\ &\quad + a_1a_4a_5 + a_1a_6a_4 + a_5a_4a_3 + a_1a_3a_2 + a_3a_5a_6 \\ &\quad + a_3a_5a_2 + a_4a_2a_6 + a_2a_5a_6) \\ A_{C.LA} &= -\frac{Y_1x}{Y_2 + Y_1x} \\ Y_1 &= a_3a_2 + a_2a_5 + a_2a_6 + a_3a_6 \\ Y_2 &= 2a_2a_3a_4 + 2a_2a_1a_6 + 2a_6a_3a_4 + 2a_3a_6a_1 + 2a_5a_6a_1 \\ &\quad + 2a_3a_6a_5 + 2a_4a_1a_6 + 2a_4a_3a_1 \\ &\quad + 2a_4a_5a_1 + 2a_4a_3a_5 + 2a_2a_4a_6 + 2a_2a_3a_1 \\ &\quad + 2a_2a_5a_1 + 2a_2a_5a_6 + 2a_2a_3a_5 + 2a_2a_4a_5 \end{aligned} \tag{8}$$

If we assume that $a_1 = a_6 = \alpha$ and $a_2 = a_3 = a_4 = a_5 = \beta$ in Eqs. (3) and (4), then Eq. (8) is reduced to

$$\begin{aligned} A_{C.LE} &= \frac{x}{4\alpha + 4\beta + x} \\ A_{C.LA} &= -A_{C.LE} \end{aligned} \tag{9}$$

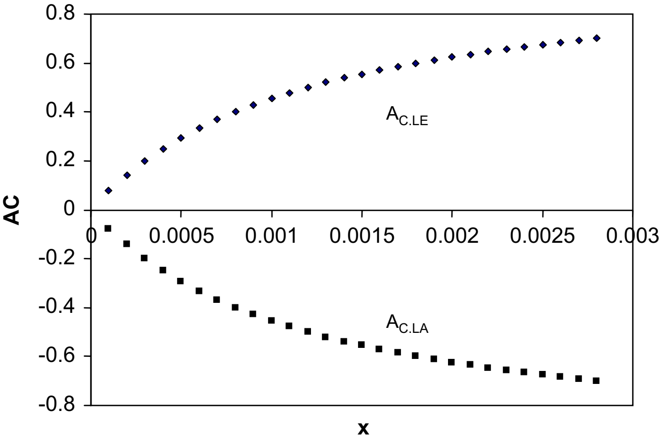

Fig. 1. Expected change of $A_{C.LE}$ and $A_{C.LA}$ for genomes with different values of x (the part of C→T mutations due to deamination). Computed with $\alpha = 0.0002$ and $\beta = 0.0001$.

Eq. (9) shows that strand bias expressed in the form of A_C will disappear when x approaches 0, but $A_{C.LE}$ will increase with x, approaching 1 when x approaches infinity. Similarly, $A_{C.LA}$ will approach -1, as x approaches infinity (Fig. 1). This can explain the previous empirical documentation of $A_{C.LE}>A_{C.LA}$ (Lobry, 1996; Lobry and Sueoka, 2002). Moreover, $A_{C.LE}$ and $A_{C.LA}$ are expected to be distributed equally above and below zero given Eq. (9), but may be asymmetrical given the more general Eq. (8). In any case, the two are expected to be negatively correlated. This prediction can be easily tested with empirical data.

2. Empirical test of the mutation-based model

The data set we used consists of 339 bacterial chromosomes (corresponding to 136 genera and 226 bacterial species). The predicted origin (ori) and terminus (ter) of replication were taken from the 'Origin of Replication' table of the Genome Atlas Database, Center for Biological Sequence Analysis, BioCentrum-DTU, Technical University of Denmark, Denmark, http://www.cbs.dtu.dk/services/GenomeAtlas/show-kingdom.php? kingdom = Bacteria (Hallin and Ussery, 2004; Worning et al., 2006). This allows us to classify each CDS according to whether it is collinear with the leading or the lagging strand.

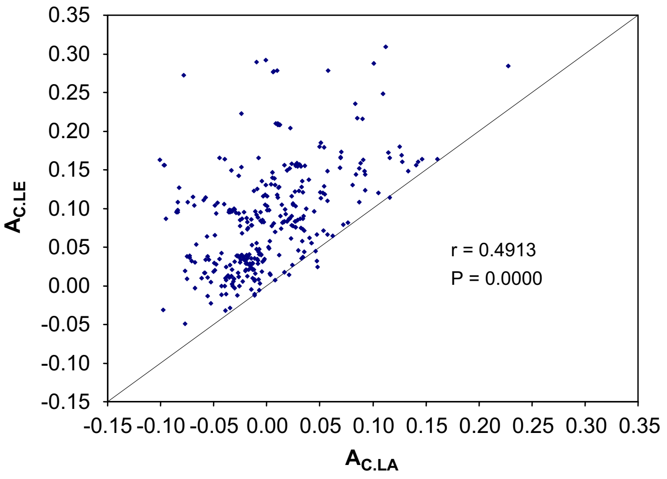

Fig. 2. Positive correlation between $A_{C.LE}$ and $A_{C.LA}$ across genomes, with $r = 0.4913$, $n = 339$ and $p<0.000001$. The straight line indicates where $A_{C.LE} = A_{C.LA}$.

The bacterial genome .ffn files containing the FASTA formatted entries of all CDSs in each chromosome were downloaded from the NCBI ftp server, ftp://ftp.ncbi.nih.gov/genomes/Bacteria/. $A_{C.LE}$ and $A_{C.LA}$ for each species (genome) were computed by pooling all CDSs in the leading strand and the lagging strand, respectively. Thus, we are not studying differences in A_C among genes within the leading or within the lagging strand. Also, our use of CDSs eliminates the complication of transcription-coupled bias (Francino et al., 1996) because the transcription-coupled bias is expected to be the same for protein-coding genes, regardless which strand the gene is located.

Our results (Fig. 2) are interesting in two ways. First, $A_{C.LE}$ is generally larger than $A_{C.LA}$, which is expected from the mutation-only substitution model presented above. Second, a strong positive correlation exists between $A_{C.LE}$ and $A_{C.LA}$, which is not expected from the mutation-only substitution model above which predicts a negative correlation. Thus, our substitution model based on biased C→T mutation on the leading strand and biased G→A mutation on the lagging strand is insufficient to explain the empirical observation.

A plot of $A_{C.LE}$ and $A_{C.LA}$ versus genomic A_C (Fig. 3) shows that both $A_{C.LE}$ and $A_{C.LA}$ are highly significantly and positively correlated with genomic A_C. Furthermore, the greater the genomic A_C, the greater the difference is between $A_{C.LE}$ and $A_{C.LA}$.

3. Replication-independent effect: *C* avoidance

A positive correlation between $A_{C.LE}$ and $A_{C.LA}$ indicates shared factors operating on both strands. At present, only a deficiency in

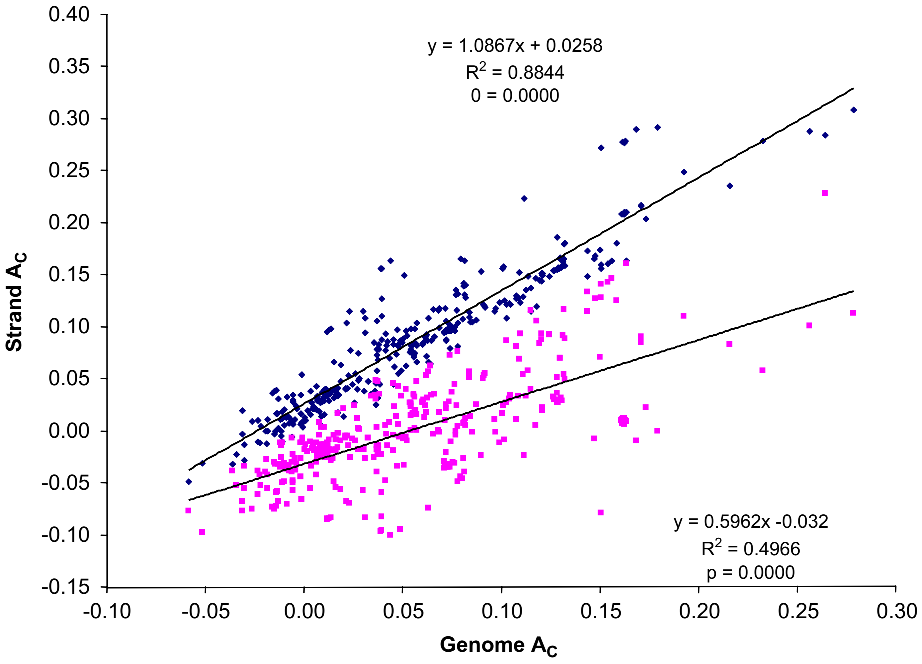

Fig. 3. $A_{C.LE}$ and $A_{C.LA}$ are both positively correlated with genomic A_C, with $A_{C.LE}$ (blue dots) increasing with genomic A_C faster (having a greater slope) than $A_{C.LA}$ (pink dots).

C usage has been documented almost universally in a variety of organisms (Rocha and Danchin, 2002; Xia et al., 2006). Two factors may have contributed to *C* deficiency. First, a relatively low intracellular cytosine availability in both eukaryotes and prokaryotes has long been reported. For example, in the exponentially proliferating chick embryo fibroblasts in culture, the concentration of ATP, CTP, GTP and UTP, in the unit of $(\text{mol} \times 10^{-12}$ per 10^6 cells), is 1890, 53, 190, and 130, respectively, in 2 h culture, and 2390, 73, 220, and 180, respectively, in 12 h culture (Colby and Edlin, 1970). The protozoan parasite, *Trypanosoma brucei*, exemplifies C-limitation in mammalian blood. The parasite has lost its ability to synthesize ATP, GTP and TTP, but maintains its *de novo* synthesis pathway for CTP. Inhibiting its CTP synthetase effectively eradicates the parasite population in the host (Hofer et al., 2001). This suggests that little CTP can be salvaged from the host. C-limitation appears to be a general feature in bacterial species, and a biochemical explanation has been offered to explain the general C-limitation in bacterial species (Rocha and Danchin, 2002). Given the low C availability, an organism with its RNA containing few C's may have a selective advantage over an organism with its RNA containing many C's.

The other factor that has been postulated to contribute to *C* deficiency is the inherent high mutation rate of *C*. Spontaneous deamination *C* to *U* or methylated *C* to *T* dominates chemical processes leading to the decay of DNA and RNA (Frederico et al., 1990, 1993; Lindahl, 1993; Sancar and Sancar, 1988). The single-stranded mRNA is particularly prone to mutations caused by the spontaneous deamination (Francino and Ochman, 2001) because the deamination rate is about 100 times higher in single-stranded nucleotide sequences than in double-stranded ones (Frederico et al., 1990). Thus, an organism with increased *C* usage may suffer from reduced reliability of its RNA products and consequently have a reduced fitness. These two factors might have contributed to long genes having reduced *C* usage (Omont and Kepes, 2004; Xia et al., 2006).

4. Substitution model incorporating *C* avoidance

We accommodate the hypothesized *C* avoidance by adding a *y* parameter to the transition probability matrices in Eqs. (3) and (4) to obtain

$$M_{LE} = \left[\begin{array}{ccccc} & A & G & C & T \\ A & \bullet & a_1 & a_2y & a_3 \\ G & a_1 & \bullet & a_4y & a_5 \\ C & a_2 & a_4 & \bullet & a_6+x \\ T & a_3 & a_5 & a_6y & \bullet \end{array}\right] \tag{10}$$

$$M_{LA} = \left[\begin{array}{ccccc} & A & G & C & T \\ A & \bullet & a_1 & a_2y & a_3 \\ G & a_1+x & \bullet & a_4y & a_5 \\ C & a_2 & a_4 & \bullet & a_6 \\ T & a_3 & a_5 & a_6y & \bullet \end{array}\right] \tag{11}$$

where $0 \leqslant y \leqslant 1$. If $y = 1$, then there is no *C* avoidance and the two *M* matrices in Eqs. (10) and (11) are reduced to those in Eqs. (3) and (4). In the terminology of Morton and Morton (2007), the parameter *y* represents the replication-independent effect that is shared by both the leading and lagging strands.

Now again solving for π_i, we have

$$\pi_{A.LE} = \frac{a_2a_4a_6y^2 + C_3y + a_3a_4xy + C_4x + C_5}{a_2a_4a_6y^3 + a_2a_4xy^2 + C_6y^2 + C_7xy + C_8y + 3C_4x + 3C_5}$$

$$\pi_{G.LE} = \frac{a_2a_4a_6y^2 + C_3y + a_2a_5xy + C_4x + C_5}{a_2a_4a_6y^3 + a_2a_4xy^2 + C_6y^2 + C_7xy + C_8y + 3C_4x + 3C_5}$$

$$\pi_{C.LE} = \frac{a_2a_4a_6y^3 + C_3y^2 + C_5y}{a_2a_4a_6y^3 + a_2a_4xy^2 + C_6y^2 + C_7xy + C_8y + 3C_4x + 3C_5}$$

$$\pi_{T.LE} = \frac{a_2a_4xy^2 + a_2a_4a_6y^2 + C_3y + (a_1a_2 + a_1a_4 + a_2a_5 + a_3a_4)xy + C_4x + C_5}{a_2a_4a_6y^3 + a_2a_4xy^2 + C_6y^2 + C_7xy + C_8y + 3C_4x + 3C_5}$$

$$\begin{aligned} C_3 &= a_1a_2a_6 + a_1a_4a_6 + a_2a_3a_4 + a_2a_4a_5 + a_2a_5a_6 + a_3a_4a_6 \\ C_4 &= a_1a_3 + a_1a_5 + a_3a_5 \\ C_5 &= a_1a_2a_3 + a_1a_2a_5 + a_1a_3a_4 + a_1a_3a_6 + a_1a_4a_5 + a_1a_5a_6 \\ &\quad + a_2a_3a_5 + a_3a_4a_5 + a_3a_5a_6 \\ C_6 &= a_1a_2a_6 + a_2a_5a_4 + 3a_2a_4a_6 + a_2a_5a_6 + a_2a_3a_4 + a_1a_4a_6 + a_3a_4a_6 \\ C_7 &= 2a_3a_4 + a_1a_2 + a_1a_4 + 2a_2a_5 \\ C_8 &= 3C_3 + C_5 \end{aligned} \tag{12}$$

$$\pi_{A.LA} = \frac{a_2a_4a_6y^2 + (a_2a_6 + a_4a_6)xy + C_3y + (C_{11} + a_5a_2 + a_5a_6 + a_5a_4)x + C_5}{a_2a_4a_6y^3 + a_2a_6xy^2 + C_6y^2 + C_9xy + C_8y + C_{10}x + 3C_5}$$

$a4\ y\ x\ a6 + (a5\ a6 + a5\ a4 + a5\ a2)x$

$$\pi_{G.LA} = \frac{a_2a_4a_6y^2 + C_3y + C_5}{a_2a_4a_6y^3 + a_2a_6xy^2 + C_6y^2 + C_9xy + C_8y + C_{10}x + 3C_5}$$

$$\pi_{C.LA} = \frac{a_2a_4a_6y^3 + C_3y^2 + a_2a_6xy^2 + (a_2a_3 + a_2a_5 + a_3a_6)xy + C_5y}{a_2a_4a_6y^3 + a_2a_6xy^2 + C_6y^2 + C_9xy + C_8y + C_{10}x + 3C_5}$$

$$\pi_{T.LA} = \frac{a_2a_4a_6y^2 + a_2a_6xy + C_{11}x + C_3y + C_5}{a_2a_4a_6y^3 + a_2a_6xy^2 + C_6y^2 + C_9xy + C_8y + C_{10}x + 3C_5}$$

$$\begin{aligned} C_9 &= a_2a_3 + a_2a_5 + 2a_2a_6 + a_3a_6 + a_4a_6 \\ C_{10} &= a_5a_2 + 2a_3a_6 + a_5a_6 + a_5a_4 + 2a_3a_2 + 2a_3a_4 \\ C_{11} &= a_2a_3 + a_3a_4 + a_3a_6 \end{aligned} \tag{13}$$

where C_3–C_8 are defined in Eq. (12).

These equilibrium frequencies lead to the $A_{C.LE}$ and $A_{C.LA}$ for the leading and the lagging strands, respectively:

$$A_{C.LE} = \frac{-a_2a_4a_6y^3 + (a_2a_4a_6 - C_3)y^2 + (C_3 - C_5)y + a_2a_5xy + C_4x + C_5}{a_2a_4a_6y^3 + (a_2a_4a_6 + C_3)y^2 + (C_3 + C_5)y + a_2a_5xy + C_4x + C_5}$$

$$A_{C.LA} = \frac{-a_2a_4a_6y^3 + (a_2a_4a_6 - C_3)y^2 - a_2a_6xy^2 - C_{12}xy + (C_3 - C_5)y + C_5}{a_2a_4a_6y^3 + (a2a_4a_6 + C_3)y^2 + a_2a_6xy^2 + C_{12}xy + (C_3 + C_5)y + C_5}$$

$$C_{12} = a_2a_3 + a_2a_5 + a_3a_6 \tag{14}$$

where C_3–C_{11} are defined in Eqs. (12) and (13).

To see how the incorporation of *C* avoidance would change the correlation patter between $A_{C.LE}$ and $A_{C.LA}$, we may simplify the situation by assuming $a_1 = a_6 = \alpha$ and $a_2 = a_3 = a_4 = a_5 = \beta$. This reduces Eqs. (13) and (14) to

$$\begin{aligned} \pi_{A.LE} &= \pi_{G.LE} = \frac{1}{y+3} \\ \pi_{T.LE} &= \frac{\alpha y + xy + \alpha + x + 2\beta}{4\alpha y + yx + 2\beta y + \alpha y^2 + 3\alpha + 3x + 6\beta} \\ \pi_{C.LE} &= \frac{y(\alpha + 2\beta + \alpha y)}{4\alpha y + yx + 2\beta y + \alpha y^2 + 3\alpha + 3x + 6\beta} \end{aligned} \tag{15}$$

$$\begin{aligned} \pi_{T.LA} &= \pi_{C.LA} = \frac{1}{y+3} \\ \pi_{A.LA} &= \frac{2\alpha + \beta y + 2x + \beta}{6\alpha + 3x + 3\beta + 2\alpha y + xy + 4\beta y + \beta y^2} \\ \pi_{G.LA} &= \frac{2\alpha + \beta + \beta y}{6\alpha + 3x + 3\beta + 2\alpha y + xy + 4\beta y + \beta y^2} \end{aligned} \tag{16}$$

$$\mathrm{A}_{C.LE} = -\frac{\alpha y^2 + 2\beta y - \alpha - x - 2\beta}{\alpha y^2 + 2\alpha y + 2\beta y + \alpha + x + 2\beta} \tag{17}$$

$$A_{C.LA} = -\frac{\beta y^2 + 2\alpha y + xy - 2\alpha - \beta}{\beta y^2 + 2\beta y + 2\alpha y + xy + 2\alpha + \beta} \tag{18}$$

$A_{C.LE}$ and $A_{C.LA}$ both increases with decreasing *y*, i.e., with increasing *C* avoidance (Fig. 4). Three predictions can be derived from this new model concerning $A_{C.LE}$ and $A_{C.LA}$. First, the stronger the strand-biased mutation, the greater the difference is between $A_{C.LE}$ and $A_{C.LA}$. Second, with *C* avoidance in both strands, $A_{C.LE}$ and $A_{C.LA}$ become positively correlated for genomes experiencing

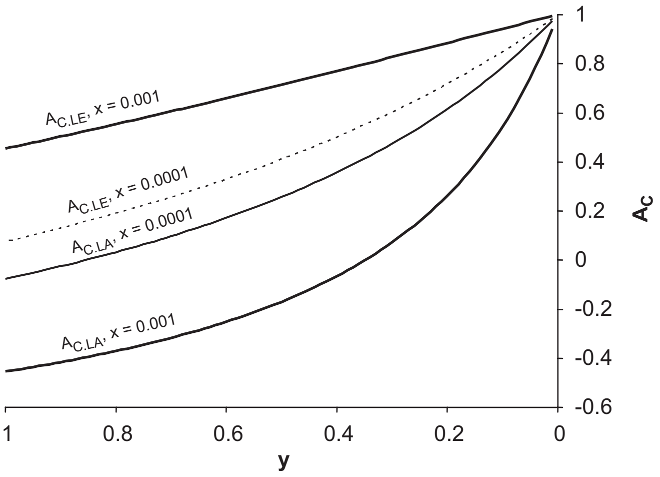

Fig. 4. As *C* avoidance increases (i.e., as *y* become smaller), both $A_{C.LE}$ and $A_{C.LA}$ will increase, leading to a positive correlate between the two. When $y = 1$, i.e., no *C* avoidance selection, $A_{C.LE}$ and $A_{C.LA}$ are symmetrically distributed above and below the zero line. Computed with $\alpha = 0.0002$, $\beta = 0.0001$ and two *x* values, 0.001 and 0.0001, respectively.

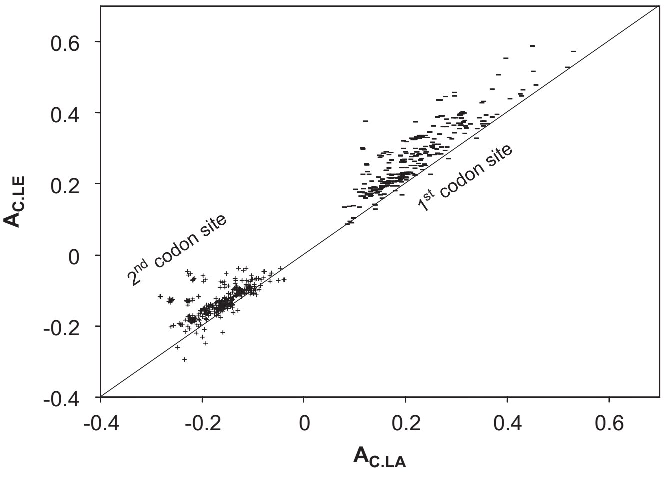

Fig. 5. Relationship between $A_{C.LE}$ and $A_{C.LA}$ for the first and second codon sites. The straight line has intercept = 0 and slope of 1, indicating where $A_{C.LE} = A_{C.LA}$.

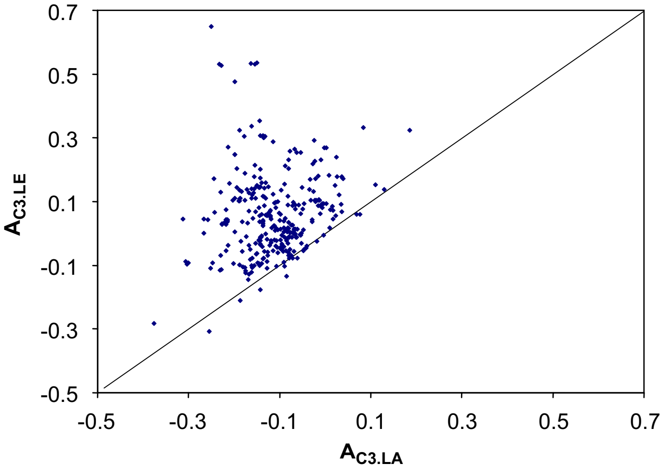

Fig. 6. Relationship between $A_{C.LE}$ and $A_{C.LA}$ at the third codon site.

different degrees of *C* avoidance. Third, when *C* avoidance is weak (when *y* is close to 1), the positive correlation between $A_{C.LE}$ and $A_{C.LA}$ are approximately linear. The empirical result (Fig. 3) is consistent with this new model incorporating *C* avoidance.

5. Different pattern among the three codon sites

The formulation of *C* avoidance, in addition to strand-biased mutation, suggests differences among the three codon sites. The second codon site should be strongly constrained by amino acid usage and should be most resistant to mutation and *C* avoidance . In fact, alanine (coded by GCN) and serine (coded by UCN) are the most frequently used amino acids in most of the prokaryotic genomes. This implies that *C* will be overused at the second codon site (resulting in negative A_C). In contrast, the third codon site should be most mutable and most susceptible to *C* avoidance.

The plot of $A_{C.LE}$ and $A_{C.LA}$ for the first and second codon sites (Fig. 5) reveals several interesting patterns. First, $A_{C.LE}$ is consistently greater than $A_{C.LA}$ for both the first and the second codon position. Thus, the strand-biased mutation is visible even in these two highly conserved codon sites. Second, $A_{C.LE}$ and $A_{C.LA}$ for the second codon site are constrained within a relatively narrow range and are both negative. This is consistent with the fact that the second codon site is strongly constrained by amino acid usage, especially by alanine and serine codons with a middle *C*. Second, $A_{C.LE}$ and $A_{C.LA}$ for first codon site are both positive, with their variances significantly greater than those for second codon site by the variance ratio test ($F = 2.7118$, $DF1 = DF2 = 338$, $p = 0.0000$ for $A_{C.LA}$ and $F = 4.6508$, $DF1 = DF2 = 338$, $p = 0.0000$ for $A_{C.LE}$). Third, the points for second codon sites are closer to the straight line than those for the first codon site, indicating that the effect of strand-biased mutation is more visible in the first codon site than that in the second codon site. To confirm this, we performed a paired-sample *t*-test between D_1 ($= A_{C1.LE}–A_{C1.LA}$) and D_2 ($= A_{C2.LE}–A_{C2.LA}$), where the subscripts 1 and 2 indicate codon site. The means are 0.0575 and 0.0339, respectively, for the first and second codon sites, and differ significantly with $t = 17.6$, $DF = 338$, $p = 0.0000$.

$A_{C.LE}$ is much greater than $A_{C.LA}$ for the third codon site relative to the difference for the first and second codon site (Fig. 6), with D_3 ($= A_{C3.LE}–A_{C3.LA}$) significantly greater than both D_1 and D_2. This is expected because the third codon site should be the most susceptible to the strand-biased mutation pressure. $A_{C3.LE}$ and $A_{C3.LA}$ are also positively correlated, but the correlation is much weaker than that between $A_{C1.LE}–A_{C1.LA}$ or between $A_{C2.LE}–A_{C2.LA}$, with $r = 0.1332$ and $p = 0.0166$.

How well is the strand-biased substitution model applicable to mitochondrial genomes is unknown because mitochondrial DNA (mtDNA) replication does not involve leading or lagging strand (Bogenhagen and Clayton, 2003; Clayton, 1982, 2000; Shadel and Clayton, 1997). During mtDNA replication, the L-strand is first used as a template to replicate the daughter H-strand, while the parental H-strand was left single-stranded for an extended period because the complete replication of mtDNA takes nearly 2 h (Clayton, 1982, 2000; Shadel and Clayton, 1997). Spontaneous deamination of both A and C (Lindahl, 1993; Sancar and Sancar, 1988) occurs frequently in human mitochondrial DNA (Tanaka and Ozawa, 1994). Deamination of A leads to hypoxanthine that forms stronger base pair with C than with T, generating an A.T→G.C mutation. Deamination of C leads to U, generating C.G→U.A mutations. Among these two types of spontaneous deamination, the C→U mutation occurs more frequently than the A→G mutation (Lindahl, 1993). In particular, the C→U mutation mediated by the spontaneous deamination occurs in single-stranded DNA more than 100 times as frequent as double-stranded DNA (Frederico et al., 1990). Note that these C→U mutants will immediately be used as a template to replicate the daughter L-strand, leading to a G→A mutation in the L-strand after one round of DNA duplication. Therefore, the H-strand, left single-stranded for an extended period during DNA replication, tend to accumulate A→G and C→U mutations and become rich in G and T, while the L-strand will become rich in A and C. The strand bias exhibits strong effect on codon usage bias and tRNA evolution (Xia, 2005). A similar strand-biased model may also be needed in modeling mitochondrial genomic evolution.

If genes switch between the leading and the lagging strands during evolution, then different strand-specific substitution models should be used along different branches in a phylogenetic analysis. How this should be implemented is still under investigation.

In summary, our comparative genomic analysis strongly suggests that both mutation and selection have acted to shape the strand asymmetry in bacterial genomes. This proposed substitution model that incorporates both strand-biased mutation and selection against C usage appears to be sufficient for the observed pattern and is expected to improve molecular phylogenetic reconstruction involving prokaryotic genomes.

Acknowledgments

This study is supported by the CAS/SAFEA International Partnership Program for Creative Research Teams and by NSERC's Discovery Grant and Strategic Research Grant to XX. A.M. acknowledges funding by the Spanish Ministry of Science and Technology (BFU2007-60791) and the Regional Government of Andalusia (CVI-162). We thank S. Aris-Brosou, C. Brunk, J. R. Lobry, E. Rocha, and H. Xiong and for comments and suggestions. Two anonymous referees helped clarify several ambiguities and provided relevant references.

References

Achaz, G., Coissac, E., Netter, P., Rocha, E.P., 2003. Associations between inverted repeats and the structural evolution of bacterial genomes. Genetics 164, 1279–1289.

Blattner, F.R., Plunkett 3rd., G.C., Bloch, A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., Shao, Y., 1997. The complete genome sequence of *Escherichia coli* K-12. Science 277, 1453–1474.

Bogenhagen, D.F., Clayton, D.A., 2003. The mitochondrial DNA replication bubble has not burst. Trends Biochem. Sci. 28, 357–360.

Chambaud, I., Heilig, R., Ferris, S., Barbe, V., Samson, D., Galisson, F., Moszer, I., Dybvig, K., Wroblewski, H., Viari, A., Rocha, E.P., Blanchard, A., 2001. The complete genome sequence of the murine respiratory pathogen *Mycoplasma pulmonis*. Nucleic Acids Res. 29, 2145–2153.

Clayton, D.A., 1982. Replication of animal mitochondrial DNA. Cell 28, 693–705.

Clayton, D.A., 2000. Transcription and replication of mitochondrial DNA. Hum. Reprod. 15, 11–17.

Colby, C., Edlin, G., 1970. Nucleotide pool levels in growing, inhibited, and transformed chick fibroblast cells. Biochemistry 9, 917.

Contursi, P., Pisani, F.M., Grigoriev, A., Cannio, R., Bartolucci, S., Rossi, M., 2004. Identification and autonomous replication capability of a chromosomal replication origin from the archaeon Sulfolobus solfataricus. Extremophiles 8, 385–391.

Francino, M.P., Ochman, H., 1997. Strand asymmetries in DNA evolution. Trends Genet. 13, 240–245.

Francino, M.P., Ochman, H., 2001. Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. Mol. Biol. Evol. 18, 1147–1150.

Francino, M.P., Chao, L., Riley, M.A., Ochman, H., 1996. Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science 272, 107–109.

Frank, A.C., Lobry, J.R., 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene 238, 65–77.

Frank, A.C., Lobry, J.R., 2000. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. Bioinformatics 16, 560–561.

Frederico, L.A., Kunkel, T.A., Shaw, B.R., 1990. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. Biochemistry 29, 2532–2537.

Frederico, L.A., Kunkel, T.A., Shaw, B.R., 1993. Cytosine deamination in mismatched base pairs. Biochemistry 32, 6523–6530.

Freeman, J.M., Plasterer, T.N., Smith, T.F., Mohr, S.C., 1998. Patterns of genome organization in bacteria. Science 279, 1827a.

Fujimori, S., Washio, T., Tomita, M., 2005. GC-compositional strand bias around transcription start sites in plants and fungi. BMC Genomics 6, 26.

Green, P., Ewing, B., Miller, W., Thomas, P.J., Green, E.D., 2003. Transcription-associated mutational asymmetry in mammalian evolution. Nat. Genet. 33, 514–517.

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

Hallin, P.F., Ussery, D.W., 2004. CBS genome atlas database: a dynamic storage for bioinformatic results and sequence data. Bioinformatics 20, 3682–3686.

Hofer, A., Steverding, D., Chabes, A., Brun, R., Thelander, L., 2001. *Trypanosoma brucei* CTP synthetase: a target for the treatment of African sleeping sickness. Proc. Natl. Acad. Sci. USA 98, 6412–6416.

Karlin, S., 1999. Bacterial DNA strand compositional asymmetry. Trends Microbiol. 7, 305–308.

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

Lobry, J.R., 1996. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13, 660–665.

Lobry, J.R., Sueoka, N., 2002. Asymmetric directional mutation pressures in bacteria. Genome Biol. 3 (research), 58.1–58.14.

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

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

McLean, M.J., Wolfe, K.H., Devine, K.M., 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J. Mol. Evol. 47, 691–696.

Morton, R.A., Morton, B.R., 2007. Separating the effects of mutation and selection in producing DNA skew in bacterial chromosomes. BMC Genomics 8, 369.

Omont, N., Kepes, F., 2004. Transcription/replication collisions cause bacterial transcription units to be longer on the leading strand of replication. Bioinformatics 20, 2719–2725.

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

Reyes, A., Gissi, C., Pesole, G., Saccone, C., 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol. Biol. Evol. 15, 957–966.

Rocha, E.P., Danchin, A., 2002. Base composition bias might result from competition for metabolic resources. Trends Genet. 18, 291–294.

Rocha, E.P., Danchin, A., 2003. Gene essentiality determines chromosome organisation in bacteria. Nucleic Acids Res. 31, 6570–6577.

Rocha, E.P., Danchin, A., Viari, A., 1999. Universal replication biases in bacteria. Mol. Microbiol. 32, 11–16.

Sancar, A., Sancar, G.B., 1988. DNA repair enzymes. Annu. Rev. Biochem. 57, 29–67.

Shadel, G.S., Clayton, D.A., 1997. Mitochondrial DNA maintenance in vertebrates. Annu. Rev. Biochem. 66, 409–435.

Tanaka, M., Ozawa, T., 1994. Strand asymmetry in human mitochondrial DNA mutations. Genomics 22, 327–335.

Worning, P., Jensen, L.J., Hallin, P.F., Staerfeldt, H.H., Ussery, D.W., 2006. Origin of replication in circular prokaryotic chromosomes. Environ. Microbiol. 8, 353–361.

Xia, X., 2005. Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. Gene 345, 13–20.

Xia, X., Yuen, K.Y., 2005. Differential selection and mutation between dsDNA and ssDNA phages shape the evolution of their genomic AT percentage. BMC Genet. 6, 20.

Xia, X., Wang, H.C., Xie, Z., Carullo, M., Huang, H., Hickey, D.A., 2006. Cytosine usage modulates the correlation between CDS length and CG content in prokaryotic genomes. Mol. Biol. Evol. 23, 1450–1454.

Zhang, J., Li, K., 2003. Single-base discrimination mediated by proofreading 3′ phosphorothioate-modified primers. Mol. Biotechnol. 25, 223–228.

Zhang, R., Zhang, C.T., 2003. Multiple replication origins of the archaeon *Halobacterium* species NRC-1. Biochem. Biophys. Res. Commun. 302, 728–734.