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

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

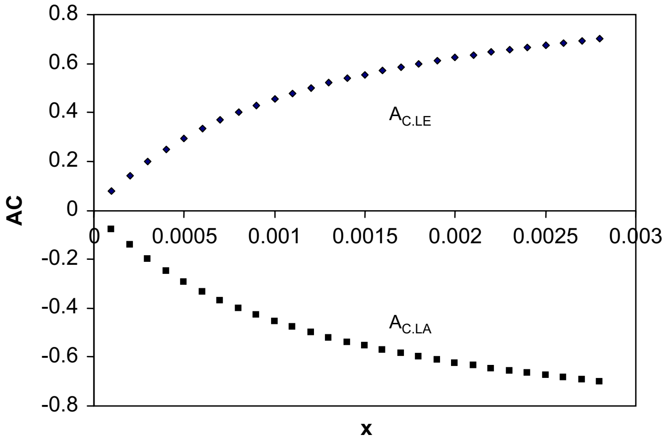

Fig. 1. Expected change of $A_{C.LE}$ and $A_{C.LA}$ for genomes with different values of x (the part of C→T mutations due to deamination). Computed with $\alpha = 0.0002$ and $\beta = 0.0001$.