5

Strand-specific Composition Bias in Bacterial Genomes

Feng-Biao Guo and Lu-Wen Ning
School of Life Science and Technology
University of Electronic Science and Technology of China
China

1. Introduction

In 1950, Chargaff experimentally found that nucleotides of G and C (or T and A) have the same abundance values when analyzing two DNA strands together (Chargaff, 1950). Three years later, Watson and Crick (1953) published the DNA double helix model and the base-pairing rule in the model explained such equivalent frequencies. This is called the first rule of Chargaff or parity rule 1. Surprisingly, Lin and Chargaff (1967) observed approximately equivalent frequencies of complementary nucleotides within each single DNA strand. This is called the second rule of Chargaff or parity rule 2. The rule 2 is theoretically explained as follows. When mutation and selection are symmetric with respect to the two strands of DNA, parity rule 1 holds the following six pairs of substitution rate to be equal, $r_{GC}=r_{CG}$, $r_{TA}=r_{AT}$, $r_{GA}=r_{CT}$, $r_{AG}=r_{TC}$, $r_{CA}=r_{GT}$, $r_{AC}=r_{TG}$, where, r_{GC} means the substitution rate of G to C in a specific strand and so on (Lobry, 1995). Having the six pairs of equal substitution rates, it is formally derived that complementary nucleotides within each strand have the same occurrence frequencies. Indeed, parity rule 2 only exists when there are not any strand biases of mutation or selection. Therefore, parity rule 2 is a natural derivation of parity rule 1 at the equilibrium state between two strands. And any deviation from parity rule 2 implies substitutional strand biases: the result of different mutations (and or repair) rates, different selective pressures, or both, between the two strands of DNA (Lobry and Sueoka, 2002). In the past two decades, these deviations from intra-strand equimolarities have been extensively studied in eukaryotes (Niu et al., 2003) and their organelles (Krishnan et al. 2004), viruses (Mrazek and Karlin, 1998), particularly in bacteria and archaea (Necsulea and Lobry, 2007). In bacteria, the observed deviations switch sign at the origin and terminus of replication. This chapter reviews the subject of strand-specific composition bias in bacterial genomes, varying strength of it in different species, the underlying mechanisms and the analyzing methods.

2. Strand-specific composition bias in bacterial genomes

2.1 Strand-specific substitution and composition biases

DNA replication is a semi-conservative process (Rocha, 2004). The two strands of the parental duplex are separated, and each serves as a template for the synthesis of a new partner strand. The parental duplex is replaced with two daughter duplexes, each of which

consists of one parental strand and one newly synthesized strand. Because of the duplex structure of the parental strands, one daughter strand would be synthesized in a 5' -> 3' direction and the other would have to be copied in a 3' -> 5' direction. However, DNA polymerases can only catalyze synthesis in the 5'-3' direction. Thus, the 5'->3' strand (known as the Leading strand), is continuously synthesized. For the 5'->3' strand (known as the lagging strand), the solution is addressed by adopting discontinuous synthesis. That is to say, lagging strand replication proceeds through the synthesis of relatively short polynucleotide segments (Okazaki fragments) that are then joined together to form a continuous strand (Rocha, 2004).

As mentioned above, the deviations from parity rule 2 observed in bacteria switch sign at the origin and terminus of replication. That is to say, the substitution bias occurs between the two replicating strands, namely leading and laggings strands. There are two major ways for studying asymmetric substitutions: observation of rate bias of substitutions between homologous sequences and direct detection of composition deviations from parity rule 2 (Frank and Lobry, 1999).

Wu and Maeda (1987) used the first method to test for asymmetric substitution in certain regions of chromosomal sequences from six primates. They obtained homologous sequences of the beta-globin complex for the six primates and then calculated the substitution matrix. After comparing the substitution rates of complementary nucleotides, they obtained the first observation of strand asymmetry. The sequence comparisons even allowed them to make predictions about the positions of replication origins. But in later studies, the examination of longer sequences (Bulmer, 1991) did not show the existence of strand asymmetry. Francino et al. (1996) used the same method to investigate asymmetric substitution in the bacterium *Escherichia coli*. They found no differences in substitution rates between leading and lagging strands. However, an excess of C -> T changes was observed on the coding strand when compared to the non-coding strand. Based on this result, they suggested that strand bias of substitution mainly resulted from transcription coupled mutation or repair. The two works were the early reports on asymmetry substitution between leading and lagging strands or between coding and non-coding strands. Rocha et al. (2006) evaluated substitution biases between leading and lagging strands in seven bacteria. Significant biases existed in all seven genomes. Among them, in *E. coli* the C -> T substitution is much higher in the leading strand than in the lagging strand. This result contradicts previous ones partly (Francino et al., 1996) and the contradiction may be caused by the very small size of gene samples used by Francino and colleagues. Recently, different substitution (C to T, A to G, and G to T) rates between coding strands and non-coding strands were also observed for 1630 human genes (Mugal et al., 2009).

The substitution bias could be reflected by the different occurrence frequencies of the four nucleotides between the two strands. The second method builds on the analysis of the DNA sequences for deviations from A=T and G=C. Such deviations in SV 40 were found to have a polarity switch at the origin of replication and thus were taken as evidence for asymmetric mutation in the replication process (Filipski, 1990). The strand nucleotide composition bias was then found in genomes of echinoderm and vertebrate mitochondria (Asakawa et al., 1991). Strand composition biases were observed in the genome of *Haemophilus influenzae* and in parts of the *E. coli* and *Bacillus subtilis* genomes by using the method of GC skew and AT skew (Lobry, 1996). In these genomes the leading strands are relatively enriched in G over C and T over A. However, the case is reversed for the lagging strands. McLean et al. (1998) examined GC skew and AT skew at the third codon position along genomic regions in

completely sequenced prokaryotes at that time. Among nine bacteria, eight have GC and AT skews that change sign at the origin of replication. The leading strand in DNA replication is G richer (over C) and T richer (over A) at codon position 3 in six eubacteria, but C and T richier in two *Mycoplasma* species. Tiller and Collins (2000) investigated the relative contributions of replication orientation, gene direction, and signal sequences to base composition asymmetries in 13 bacterial genomes by using qualitative graphical presentations and quantitative statistical analyses. The effect of replication orientation, i.e., the gene is locacted on the leading or lagging strand, was found to contribute a significant proportion of the GC and AT skews. This effect is independent of, and can have opposite signs to the effects of transcriptional or translational processes, such as selection for codon usage, expression levels. With the rapid growth in the number of sequenced genomes, more and more bacteria are described with strand composition bias. Here, *Chlamydia muridarum* (Guo and Yu, 2007) is taken as an example to illustrate strand-specific composition bias at the three codon positions of genes and results are shown in Table 1. As can be seen, the G versus C bias is statistically significantly (Paried t-test, $p<0.01$) different between the two repicating strands, whereas the T verus A bias is not (Paried t-test, $p>0.05$).

	Leading strand					
	a	c	g	t	g-c	t-a
1st codon position	0.26	0.18	0.33	0.23	0.16	-0.03
2nd codon position	0.30	0.21	0.17	0.32	-0.04	0.02
3rd codon position	0.28	0.12	0.21	0.38	0.09	0. 10
Overall	0.28	0.17	0.24	0.31	0.07	0.03
	Lagging strand					
1st codon position	0.28	0.22	0.27	0.23	0.05	0.05
2nd codon position	0.30	0.24	0.14	0.32	-0.10	0.10
3rd codon position	0.31	0.20	0.14	0.36	-0.06	0.06
Overall	0.30	0.22	0.18	0.30	-0.04	0.04

Table 1. Strand specific composition bias in the *C. muridarum* genome.

2.2 Methods used to elucidate the bias and to predict replication origins

For un-annotated bacterial genomes, information on the localization of the replication origin is not available. Therefore, it is unknown whether a gene is located on the leading or lagging strands and quantitative results as in Table 1 could not be obtained. In this circumstance, the strand composition biases, i.e. deviations from parity rule 2, are usually studied by graphical methods. GC-skew (and or AT-skew), cumulative GC-skew and Z curve are three such methods.

GC skews were first used to study mitochondrial strand asymmetry and then widely used to bacterial genomes (Lobry, 1996). The GC skew calculation is performed by the following equation:

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

where G and C denote the occurrences of the corresponding bases in a given sequence with given length. The skew values along a long sequence were studied often by using a sliding window. The window length is fixed and two adjacent windows may overlap partly in some cases. Take the chromosomal position as horizontal axis and the vertical axis denotes the skew value, a line chart could be drawn. In that way, a GC skew plot for *E. coli* K12 is obtained and shown in Figure 1. As can be seen from the figure, there is composition asymmetry along the chromosome. The skew switches signs at two sites and hence divides the genome into three parts. In fact, the two switching points correspond to experimentally determined replication termini and origins. So, the three regions are leading strand, lagging strand and leading strands, respectively. As can be seen, the leading strand has positive GC skew values and the region that is a lagging strand has negative GC skew values in the *E. coli* genome.

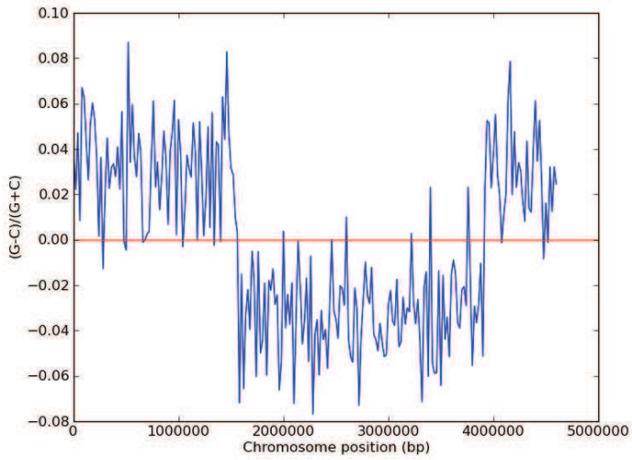

Fig. 1. GC skew for the *E. coli* K12 genome.

Although the window-based GC skew method is extensively used, the proper window size is hard to adjust. Such plots may not always be very illustrative due to many visible fluctuations for a small window size, while larger windows may hide precise coordinates of polarity switches. Therefore, an optimal window size does not exist in many cases. To address this point, a more convenient skew diagram was later proposed by Grigoriev (1998). He suggested to calculate directly the sum of (G–C)/(G+C) in adjacent windows from an arbitrary start to a given point in a sequence. Although this method is based on a sliding window, the diagram of cumulative GC skew tends to be smoother because it adopts the form of a sum. To avoid the dependence on the window size w and chromosome length c, Grigoriev (1998) suggested that the cumulative skew values are multiplied by w/c. A cumulative skew diagram for *E. coli* K12 is shown in Figure 2. As can be seen, there indeed are much less fluctuations than in Figure 1. It also shows that the switching points become peaks. The maximum skew value corresponds to the replication terminus and the minimum corresponds to the replication origin.

TA skew or cumulative TA skew could be calculated and plotted by replacing the symbol G by T and C by A in the equation (1). Similarly, keto-amino or purine-pyrimidine skew may be obtained by making appropriate replacements.

Both GC skew and cumulative GC skew are based on sliding windows. The Z curve is one method that thoroughly gets rid of sliding window. We briefly describe the Z curve method

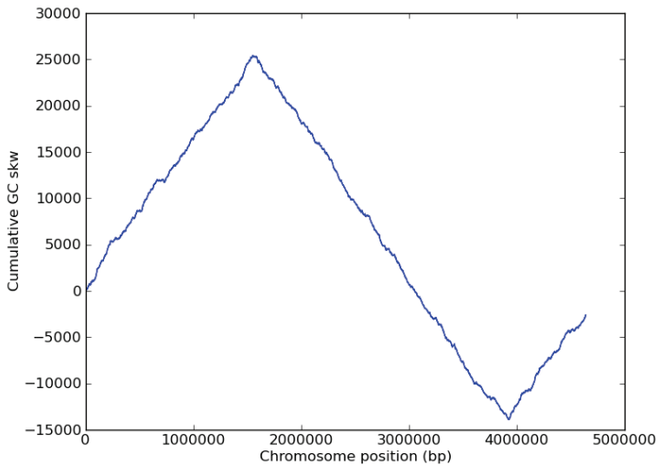

Fig. 2. Cumulative GC skew diagram for the *E. coli* K12 genome.

as follows. The Z curve is a three dimensional space curve constituting the unique representation of a given DNA sequence in the sense that for the curve or for the sequence each can be uniquely reconstructed from the other (Zhang and Zhang, 2003). Consider a DNA sequence read from the 5′-end to the 3′-end with N bases, inspecting the sequence one base at one time, and beginning with the first base. The number of inspecting steps could be denoted by n, i.e., n =1, 2,. . . , N. In the *n*th step, let us count the cumulative numbers of the bases A, C, G, and T, occurring in the subsequence from the first to the *n*th base and denote them by A_n, C_n, G_n, and T_n, respectively. The Z curve consists of a series of nodes P_n, where n =1, 2,. . . , N, whose coordinates are denoted by x_n, y_n, and z_n. It was shown that (Zhang and Zhang, 2003)

$$\begin{aligned} Xn &= (An + Gn) - (Cn + Tn) \equiv Rn - Yn \\ yn &= (An + Cn) - (Gn + Tn) \equiv Mn - Kn \\ zn &= (An + Tn) - (Cn + Gn) \equiv Wn - Sn \\ n &= 0, 1, 2, ..., N, \ xn, yn, zn \in [-N,N], \end{aligned} \tag{2}$$

where $A_0 = C_0 = G_0 = T_0 = 0$ and hence $x_0 = y_0 = z_0 = 0$. The symbols R, Y, M, K, W, and S represent the puRines, pYrimidines, aMino, Keto, Weak hydrogen bonds and Strong hydrogen bonds, respectively, according to the Recommendation 1984 by the NC-IUB (Cornish-Bowden, 1984). The connection of the nodes P_0 ($P_0 = 0$), P_1, P_2, ..., until P_N one by one sequentially by straight lines is called the Z curve for the DNA sequences inspected. The Z curve defined above is a 3-D space curve, having three independent components, i.e., x_n, y_n and z_n (Zhang and Zhang, 2003).

When being used for predicting replication origin or studying strand composition bias, only the x and y components of the 3-D Z curve are involved (Guo and Yu, 2007). According to equation (1), the x component curve denotes the plus of cumulative excess of G over C and A over T. Whereas, the y component curve represents the opposite number of the plus of cumulative excess of G over C and T over A. In short, the x component denotes the cumulative excess of purine over pyrimidine and the y component means the opposite number of cumulative excess of keto over amino. As an example, the x and y component curves for the *E. coli* K12 chromosome are shown in Figure 3. As can be seen, there are two

peaks in both of the two curves and they correspond to the replication terminus and the replication origin, respectively.

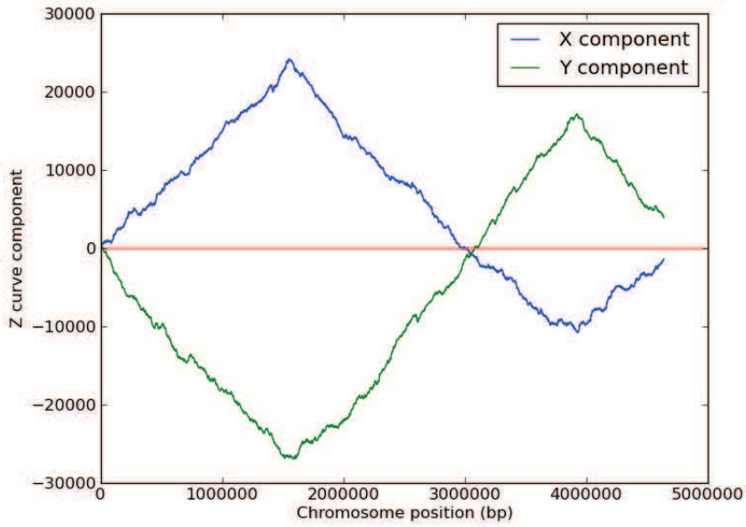

Fig. 3. X and Y component curves of the Z curve method for *E. coli* K12 genome.

According to analyses on bacterial genomes with experimentally replication origin, skew or Z curve plots for almost all of them inflect the sign or polarity at the sites of replication origins. This is the result of different nucleotide composition biases between the two replicating strands. Based on that fact, replication origins may be putatively predicted by using such methods in newly sequenced bacterial genomes. Indeed, during the annotation process for most of sequenced prokaryotes, replication origins were identified by using one, two or all three of these methods. Therefore, theoretically predicting replication origins is one of the practical applications from the universal phenomenon of strand composition bias in bacterial genomes.

2.3 Consistent direction and varying strengths of strand composition bias

Almost all of the literatures reporting significant strand composition bias revealed an excess of G over C in the leading strands in bacterial genomes. However, C over G excess in the leading strand is very rarely observed. Necsulea and Lobry (2007) performed a thorough analysis of base skew in 360 sequenced bacterial genomes. In this work, they investigated the direction or sign of bias between complementary nucleotides. Table 2 summarizes their results. Among 360 bacteria, only 33 chromosomes show no significant effect of replication. The absence of direct replication effects on base composition bias seems to be more frequent in certain bacterial families, such as *Cyanobacteria*, where 7 out of 17 chromosomes show no effect of replication on nucleotide skews, and Mollicutes (10 out of 16 chromosomes). Another noteworthy point is that only two out of 360 genomes have excess of C over G in the leading strands. Therefore, the direction (or sign) of G versus C bias is the same in nearly 100% of bacterial species. Comparatively, the bias of T versus A is not so consistent. As can be seen, about 14% (35/253) of chromosomes differ from the collective with statistically significant (randomisation test, $p < 0.05$) excess of T over A in the leading strands. Therefore,

the directions of G versus C and T versus A biases are generally consistent in most bacterial genomes.

	A>T	A=T	A<T
G>C	33	73	205
G=C	1	33	13
G<C	1	1	0

Table 2. Numbers of bacteria with various composition biases in the leading strand (adapted from Necsulea and Lobry, 2007)

However, the strength of specific composition biases varies from genome to genome in bacteria. Rocha (2004) once used one quantitative method to evaluate the strength of strand composition bias in 58 completely sequenced prokaryotes. The accuracy of the discrimination of the leading strand genes and proteins based on their nucleotide compositions is employed as the index measuring strand bias. If there are no composition biases between the two strands, the expected accuracy is about 50%. According to their results, *Streptomyces coelicolor* has the least bias and the classification accuracy is less than 60%. The accuracies of most genomes vary in the ranges between 60% and 90%. Most interestingly, three obligate intracellular parasites have the accuracy higher than 90%. That means they have very strong composition biases between the two replicating strands. Among them, *Borrelia burgdorferi*, has the highest accuracy of 95% when differentiating genes on the two strands based solely on the amino acid content and 97% using nucleotide composition.

Prior to Rocha, the different nucleotide compositions between genes on the two replicating strands of *B. burgdorferi* had been observed using graphical methods (Mcinerney, 1998). If the strand composition bias is strong enough, the individual nucleotide biases could propagate into higher-order biases in a correlated way, thereby changing the relative frequencies of codons and even amino acids of genes and encoded proteins in each of the replicating strands. Therefore, codon usage analysis could reflect the nucleotide composition in bacterial genomes with strong strand bias. In Mcinerney (1998), a correspondence analysis (COA) was made first for codon usages of all genes in *B. burgdorferi*. Then a scattering plot was drawn by using the two most important axes of COA. In the plot, points denoting about 567 genes were divided into two clusters. These two clusters appeared to be quite distinct, with very little overlap. And this meant that they had different nucleotide compositions or codon usages. On inspection, it was shown that these two groups defined the genes that were located on the leading or on the lagging strands. This was the first observation of separate codon usage associated with replication in bacterial genomes. In the past decade, another 10 bacterial genomes were also found to have extremely strong composition bias (Wei and Guo, 2010). In other words, genes on the two replicating strands were found to have separate base/codon usages in genomes of 11 bacteria including *B. burgdorferi*.

Among the 11 bacteria with extremely strong strand composition bias, the observations for three are from our group: *Chlamydia muridarum* (Guo and Yu, 2007), *Lawsonia intracellularis* (Guo and Yuan, 2009) and *Ehrlichia canis* (Wei and Guo, 2010). Here we take *Lawsonia intracellularis* as an example and briefly describe our work in the following. As an obligate intracellular bacterium, *Lawsonia intracellularis* could cause ileum inflammation in most animals, especially in pigs. The genome of *L. intracellularis* PHE/MN1-00 was determined in

2006. The complete genome sequence of *L. intracellularis* was downloaded from GeneBank. According to the annotation, the chromosome contains 1180 protein-coding genes. Most analyses were carried out using codonW. This software was used to determine the major source of variation of codon usage among the genes on the chromosome. Only those codons for which there is a synonymous alternative were used in the analysis. Hence, the three termination codons and the codons that encode Methionine and Tryptophan were excluded. Consequently, each gene is described by a vector of 59 variables (codons). COA plots all the genes analyzed in their 59-dimensional space and attempts to identify a series of new orthogonal axes accounting for the greatest variation among genes. The first principal axis is chosen to maximize the standard deviation of the derived variable and the second principal axis is the direction that maximizes the standard deviation among directions un-correlated with the first, and so forth.

Here, Figures 4 shows the positions of the genes along the first and second major axes produced by COA on codon counts. The closeness of any two genes on the plot reflects the similarities of their codon usages. As can be seen, the first axis individually could separate the genes into two clusters with little overlap. The following two facts indicate that the two groups correspond to genes on the leading and lagging strands of replication, respectively. (i) The first axis is found to strongly correlate with GC and AT skews. At the left end of the first axis, genes are characterized by richness in nucleotides G and T, whereas the case is opposite at the right end. (ii) The coordinates of individual genes along the first axis of COA are plotted against the chromosomal locations of the corresponding genes in Figure 5. Genes on the direct strand and those on the reverse complement strand are denoted by red and blue squares, respectively. It is found that genes on the left side of sequenced direct strand and genes on the right side of the reverse complement strand have lower coordinate values of the first axis, whereas, for the other genes, the opposite occurs. In fact, genes on the left side of direct strand and those on the right side of the reverse complement strand just correspond to genes on the leading strand, whereas the other ones correspond to the lagging strand. Therefore, it is reasonable to say that two clusters in Figure 4 correspond to genes on the leading strands and lagging strands, respectively. After marking genes located on the leading, lagging strands by different symbols in Figure 4, the speculation is confirmed.

A Chi-square test was then performed for comparing RSCU values of genes located on the two replicating strands and results are listed in Table 3. RSCU (Relative Synonymous Codon Usage) is defined in Equation 3. where x_{ij} is the occurence number of the *j*th codon for the *i*th amino acid, and n_i denotes the degree of codon degeneracy for the *i*th amino acid.

$$\mathrm{RSCU}\ ij = \frac{x_{ij}}{\frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}} \tag{3}$$

In the table 3, the symbl ++ indicates that the leading strand genes used the codon more frequently than the lagging strand genes, and the symbol -- indicates the lagging strand genes used the codon more frequently than the leading strand genes, whereas xx indicates that there is no significant difference in usage of the codon on either strand. In total, 49 among 59 codons are found to be significantly different between genes on the leading strand from those on the lagging strand. Among the 23 codons used more frequently in the leading strand, 19 are G-ending or T-ending and the exceptions are TTA, ACA, AGA and GCA.

Among the 26 codons used more frequently in the lagging strand, 16 are C-ending, 8 are A-ending and codons CTT and ACT constitute the outliers. Results of the chi-square test confirm that there is a bias towards G, T in the leading strand, and towards C, A in the lagging strand. Therefore, it could be concluded that in *L. intracellularis*, the leading and lagging strands of replication display an asymmetry in the compositions and this bias is the most important source of codon usage variation.

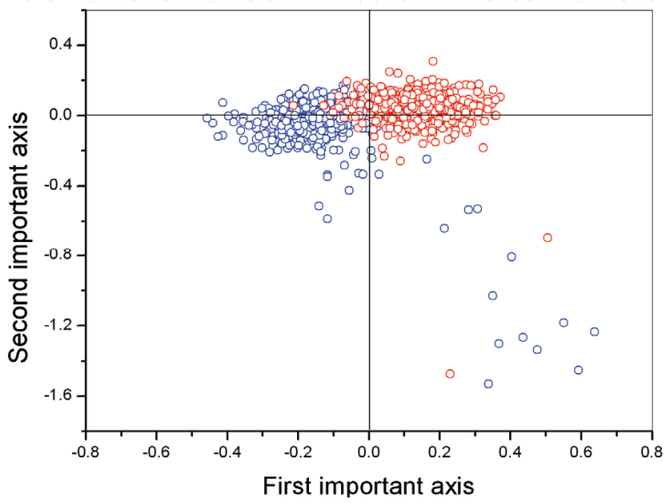

Fig. 4. Plot of the two most important axes after COA on codon counts for the 1180 genes on the *L. intracellularis* chromosome.

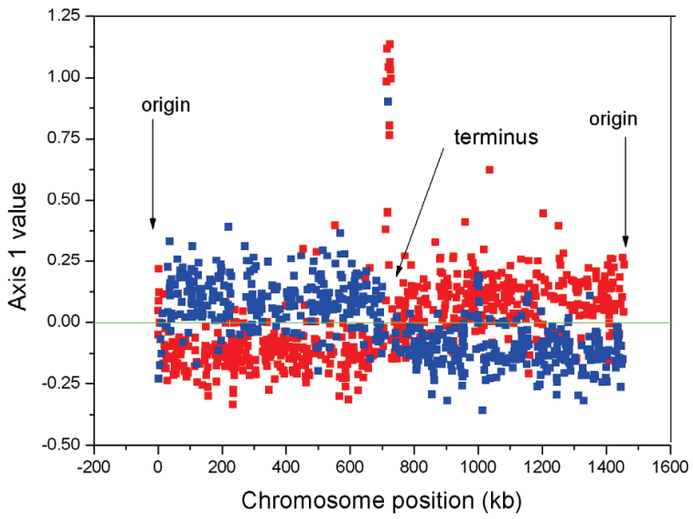

Fig. 5. Plot of axis 1 values of chromosomal genes against the corresponding chromosomal locations in the *L. intracellularis* genome.

AA	Codon	RSCU Leading	Significant	RSCU Lagging	AA	Codon	RSCU Leading	Significant	RSCU Lagging
Phe	UUU	1.81	++	1.64	Ser	UCU	2.21	XX	2.25
	UUC	0.19	--	0.36		UCC	0.23	--	0.39
Leu	UUA	2.53	++	1.86		UCA	1.47	--	1.60
	UUG	0.60	++	0.20		UCG	0.14	++	0.08
Tyr	UAU	1.74	++	1.60	Cys	UGU	1.77	++	1.59
	UAC	0.26	--	0.40		UGC	0.23	--	0.41
ter	UAA	0.00	XX	0.00	ter	UGA	0.00	XX	0.00
ter	UAG	0.00	XX	0.00	Trp	UGG	1.00	XX	1.00
Leu	CUU	1.98	--	2.44	Pro	CCU	2.08	XX	2.06
	CUC	0.17	--	0.46		CCC	0.18	--	0.30
	CUA	0.56	--	0.89		CCA	1.64	XX	1.59
	CUG	0.15	XX	0.14		CCG	0.10	XX	0.05
His	CAU	1.75	++	1.61	Arg	CGU	2.24	XX	2.18
	CAC	0.25	--	0.39		CGC	0.26	--	0.51
Gln	CAA	1.41	--	1.74		CGA	0.74	--	1.02
	CAG	0.59	++	0.26		CGG	0.25	XX	0.20
Ile	AUU	1.72	++	1.59	Thr	ACU	1.33	--	1.39
	AUC	0.26	--	0.41		ACC	0.24	--	0.36
	AUA	1.02	XX	1.01		ACA	2.24	++	2.14
Met	AUG	1.00	XX	1.00		ACG	0.19	++	0.12
Asn	AAU	1.66	++	1.49	Ser	AGU	1.58	++	1.24
	AAC	0.34	--	0.51		AGC	0.36	--	0.45
Lys	AAA	1.47	--	1.76	Arg	AGA	1.88	++	1.75
	AAG	0.53	++	0.24		AGG	0.63	++	0.34
Val	GUU	1.93	++	1.82	Ala	GCU	1.85	XX	1.84
	GUC	0.29	--	0.47		GCC	0.25	--	0.36
	GUA	1.40	--	1.51		GCA	1.78	++	1.73
	GUG	0.37	++	0.20		GCG	0.12	XX	0.08
Asp	GAU	1.73	++	1.60	Gly	GGU	1.67	++	1.41
	GAC	0.27	--	0.40		GGC	0.32	--	0.41
Glu	GAA	1.43	--	1.73		GGA	1.46	--	1.76
	GAG	0.57	++	0.27		GGG	0.55	++	0.42

Table 3. Chi-square results of RSCU of genes on the leading and lagging strands.

2.4 The underlying mechanism for the composition bias in bacterial genomes

As mentioned above, almost all the bacterial genomes have significant strand-specific composition biases. It is necessary and important to investigate the underlying mechanisms of such biases. Two published papers reviewed numerous explanations for the base composition bias in bacterial genomes (Frank and Lobry, 1999; Rocha, 2004). These hypotheses could be divided into two major categories (Necsulea and Lobry, 2007). The first hypothesis supposes that the replication mechanism is a direct cause of base composition asymmetry. The different mutation frequencies between the two replicating strands result in the nucleotide composition bias (Powdel et al., 2009). The second hypothesis states (Powdel et al., 2009) that the deviations from PR2 are associated with the strand asymmetry of the transcription mechanism, in combination with the gene distribution bias encountered in bacterial chromosomes (most protein-coding genes were located on the leading strands). This theory also falls back on mutation bias for detail interpretation. During transcription,

template strand and non-template strand have different mutation probabilities and subsequent repair. As for the main cause of these asymmetries, numerous authors have provided many solid evidence in favour of the mutationist view by demonstrating that the base skews are mainly expressed at the third codon positions of genes as well as in non-coding regions where selective pressure is minimal (Lobry, 1996). For either mutationist views, cytosine deamination of single-stranded DNA performs a vital role in the generation of strand composition bias. The deaminiation of cytosine leads to the formation of uracil. Because of the Watson-Crick base paring, cytosine is effectively protected against deamination in normal circumstances in vivo. However, the rate of cytosine deamination increases 140 times in the single-stranded DNA (Beletskii and Bhagwat, 1996). If the resulting uracil is not replaced with cytosine, C -> T mutations occur. During the process of replication, the leading strand is exposed more time in the single-stranded state than the lagging strand. Therefore, C to T mutations occur more frequently in the leading strand than in the lagging strand and then the excesses of G(C) relative to C(G) and T(A) relative to A(T) are produced in the leading(lagging) strand. During transcription, the coding strand is more exposed in the single-stranded state. Therefore, it has more G over C.

Extensive evidence has been proposed to support the replication mechanism as a direct cause of base composition asymmetries (Necsulea and Lobry, 2007). As mentioned above, the analyses of the codon usage patterns, through correspondence analysis or other statistical methods, showed that in some bacterial species genes located on the replicating strands can be distinguished by their synonymous codon choice (McInerney, 1998; Wei and Guo, 2010). Using the ANOVA method on GC and AT skews, with gene direction and replication orientation as the explanatory variables, Tillier and Collins showed that the nucleotide composition of a bacterial gene is significantly influenced by its position on the leading or the lagging strand for replication (Tillier and Collins, 2000). Lobry and Sueoka (2002) performed one thorough analysis on 43 prokaryotic chromosomes and confirmed that deviations from parity rule 2 differ significantly between leading and lagging strands. This is one of the convincing evidences. Worning et al. (2006) suggested that the sign of AT-skew is determined by the polymerase alpha subunit that replicates the leading strand. In bacteria such as *Firmicutes*, where both genes are present the AT-skew is positive on the leading strand, whereas it is negative in genomes that contain only *dnaE*. Qu et al. (2010) confirmed this conclusion based on a larger dataset.

The second hypothesis also has its supporting evidence. Francino et al. (1996) concluded that the substitution patterns were similar on the leading and lagging strands, but significantly different between the coding and non-coding strands, based on the observation of several genes in E. coli K12. Therefore, they suggested that a process linked to transcription rather than the mode of replication caused the nucleotide asymmetry. Note that a partly contradictory result was obtained by Rocha et al. (2006), at the whole genomic scale in the same species. According to them, the C to T substitution is much higher in leading strands than in lagging strands in *E. coli*. Nikolaou and Almirantis (2005) contributed an interesting work to the area, in favour of the latter type of mechanism. In order to produce a perfect gene orientation bias, they used the method of artificially rearranging the bacterial chromosome. In the case of *Nostoc* sp. the rearrangement generated a strong trend in base composition asymmetry. Thus, Nikolaou and Almirantis (2005) suggested that the gene orientation bias would be the main factor responsible for the existence of the nucleotide skews in this bacterium, and replication only had an indirect role on base asymmetry.

Based on the artificial genome rearrangement proposed by Nikolaou and Almirantis, Necsulea and Lobry (2007) developed one novel method to distinguish the replication and transcription effects on base composition asymmetry. Their results suggested that the effect of replication on the GC-skew is generally very strong. For numerous species, the AT-skew is caused by coding sequence-related mechanisms. Therefore, the cause of base composition bias in bacterial genomes would be the superposed effect of replication and transcription. The superposed effect of the two processes may be the sum or the difference. In other words, transcription-associated asymmetries can either increase or decrease replication-associated strand asymmetries, depending on the transcription direction and the position of the gene relative to the origin of replication (Necsulea and Lobry, 2007; Mugal et al., 2009). See also the chapter by Seligmann in this book.

2.5 Why there exists extremely strong strand composition bias in obligate intracellular parasites?

As mentioned above, 11 bacteria have been found to have extremely strong strand composition bias (Wei and Guo, 2010). The bias is strong enough to divide base and codon usages according to whether genes are located on the leading or lagging strands. Their names are *Borrelia burgdorferi*, *Treponema pallidum*, *Chlamydia trachomatis*, *Buchnera aphidicola*, *Blochmannia floridanus*, *Bartonella henselae*, *Bartonella quintana*, *Tropheryma whipplei*, *Chlamydia muridarum*, *Lawsonia intracellularis* and *Ehrlichia canis*, respectively. Among them, the replication associated codon usage seperation for the last three bacteria are reported by our group (Guo and Yu, 2007; Guo and Yuan, 2009; Wei and Guo, 2010). Investigating the common characters of 11 bacteria may be interesting and important.

As reported in many cases, the living environment and living styles may exert influence on the genomic G+C content and on codon usgaes of genes. Based on this consideration, we compare the living habitation of the 11 bacteria. Among them, 9 belong to oligate intracellular parasites and this means they live permanently in the cell of their host. Due to this safe habitation in the living cell, they would suffer less damage on DNA from ultraviolet radiation or other physical, chemical factors than freely living bacteria. After long-term evolution, some or most genes coding for DNA repair enzymes may be lost from these species. Due to the loss of such genes or enzymes, mutations generated during the replication process are not effectively corrected. The replicaiton associated mutation in obligate intracellular parasites would accumulate much more than in freely living bacteria. Such mutations might be a major cause for the strand composition bias in bacterial genomes. So, more mutations, more bias. The above deduction is our speculation. Its correctness should be validated by a large scale test in the future.

Secondly, chromosomes of the 11 bacteria are all shorter than 2000 kb. According to statistics on fully sequenced genomes, bacterial chromosomes vary from 160 kb to more than 10000 kb. However, all 11 species have small genome sizes although some of these bacteria are not endosymbionts. Hence, we supposed that small genome size is a necessary condition to generate strong enough strand-specific mutational bias. Perhaps in small bacterial genomes that have suffered reductive evolution, the repair mechanism of replication may be inefficient. Alternatively, in bacteria with larger chromosome, the mutation pressure is hard to prevail over translational selection.

Thirdly, all of the 11 bacteria have medium or low genomic G+C content. Among them, *B. aphidicola* has the lowest G+C content (26%), whereas *T. pallidum* has the highest G+C content (52%). Perhaps, the environment of high G+C contents is adverse to the generation

of strong strand mutation biases. Future experimental works are required to clarify the relationship between the asymmetric mechanism of replication and genomic GC content or genome size.

Fourthly, the strong mutation bias may be associated with the absence of certain genes involved in chromosome replication. As suggested by Klasson and Andersson (2006), the strong strand-specific mutational bias in endosymbiont genomes coincides with the absence of genes associated with replication restart. After a comparative analysis on 20 gamma-proteobacterial genomes, it was found that endosymbiont bacteria lacking *recA* and other genes coding for the replication restart pathway, such as *priA*, displayed the strongest strand bias. Following that study, here we investigate the absence of *mutH*, *priA*, *topA*, *dnaT*, *fis* and *recA*, which are all associated with replication initiation and the re-initiation pathway. Consequently, genes *mutH*, *dnaT* and *fis* are found to be absent in all 11 bacterial genomes with extreme strand asymmetry bias. Comparatively, all of the three genes exist in *E. coli* and other -proteobacteria, which have medium mutation biases. Klasson and Andersson (2006) suggested that cytosine deaminations accumulate during single-strand exposure at stalled replication forks and the extent of strand composition bias may depend on the time spent in repairing such lesions. Inefficient re-start mechanisms result in the replication fork to be arrested for longer time and hereby lead to higher DNA strand asymmetry. As a common character of the 11 genomes, we believe that genes associated with the replication restart pathway are very likely to be absent in the other genomes, found in the future, with strong strand mutational bias.

Finally, Figure 6 shows the y component curves of the Z curve defined in equation (2) for five representatives of the 11 bacteria with extremely strong strand composition bias. For comparison, the y component curve of the *E. coli* K12 chromosome is also shown. In *E. coli*, there also exists strand specific composition bias, however it is not strong enough to generate separate codon usages. As can be seen, all of the y component curves for the five bacteria are much smoother than for *E. coli*. The latter's y component curve has many prickles (or local fluctuations) along the chromosome. As shown in Grigoriev (1998), local fluctuations in the chromosome diagrams often correspond to sequence inversions or direct

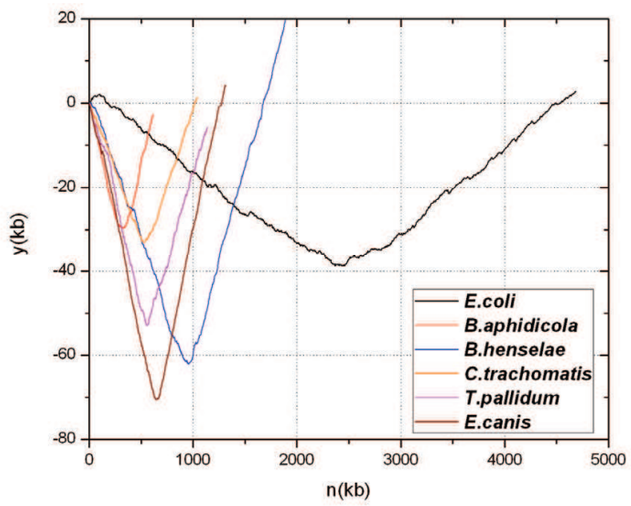

Fig. 6. Y component curves for 5 bacteria with extremely strong strand composition bias and that for the free-living bacterium *E.coli* K12.

translocations to another half of a chromosome, or integration of foreign DNA into the chromosome. Note that protection against mutations by secondary strcuture formation also explains such strand asymmetries (Krishnan et al., 2004). In other words, chromosome rearrangements often are exhibited as little prickles in the y component curves. Therefore, we could make the conclusion that the 11 bacterial chromosomes are highly stable and have very few rearrangements. According to Rocha (2004), lower rearrangement frequency are just the most likely reasons for the appearance of seperate codon usages in some obligate intracellular parasites. Our results confirmed this speculation.

3. Strand composition bias in eukaryotes, organelles, archaea and plasmids

Compared with bacterial genomes, studies on strand composition bias in eukaryotic genomes are limited. Most analyses of eukaryotic genomes did not show strand compositional asymmetry at chromosome scale (Grigoriev, 1998; Gierlik et al., 2000). It is probably a result of a relative excess of autonomously replicating sequences (ARS) and of random choice of these sequences in each replication cycle (Gierlik et al., 2000). However, the examination of three contigs from human genomes gave some evidence of strand compositional asymmetries. In addition, local asymmetries have been found in the last ARS from both ends of chromosomes of *Saccharomyces cerevisiae* (Grigoriev, 1998; Gierlik et al., 2000). In these regions, replicons may be longer. To circumvent the predicament of the lack of known replication origin, Niu et al. (2003) resorted to neighboring gene pairs that are located on different strands of nuclear DNA. Such gene pairs are most probably coded on two replicating strands. It was found that the relative frequencies of T versus A and of G versus C are significantly skewed in most studied eukaryotes when examining the introns and the fourfold degenerate sites of codons in the genes of each pair. After using quadrant diagrams to distinguish the effects of replication and transcription, the study demonstrated that there are different causes in studied genomes although the composition bias existed in most of them. For example, both transcription-associated mutation bias and replication-associated mutation bias may play an important role in causing strand asymmetry in *S. cerevisiae*. In *Schizosaccharomyces pombe*, transcription asymmetry is more likely to be the major cause of the DNA strand bias. In *A. thaliana*, transcription-associated asymmetric A to T or A to C to T substitution may be the genuine cause of the bias (Niu et al., 2003).

As for human genomes, Francino and Ochman (2000) failed to detect the asymmetry of some replicons by the phylogenetic comparisons. Analysis of the whole set of human genes revealed that most of them presented TA and GC skews (Touchon et al., 2003). The two kinds of biases are correlated to each other and they are specific to gene sequences, exhibiting sharp transitions between transcribed and non-transcribed regions. At the same time, Green et al. (2003) also described a qualitatively different transcription-associated strand asymmetry in humans. In their study, human orthologous sequences were generated by aligning with eight other mammals. The authors saw pronounced asymmetric transition substitutions in the transcribed regions of human chromosome 7. The transitions of A to G were 58% more frequent than T to C and G to A transitions were 18% more frequent than C to T. With 'maximal segment' analysis, they showed that the strand asymmetry was associated specifically with transcribed regions. Two years later, Touchon et al. (2005) analyzed intergenic and transcribed regions flanking experimentally identified human replication origins and the corresponding mouse and dog homologous regions. They demonstrated that there existed compositional strand asymmetries associated with replication. By using wavelet

transformations of skew profiles, the authors revealed the existence of 1000 putative replication origins associated with randomly distributed termination sites in human genome (Touchon et al., 2005). Around these putative origins, the skew profile displayed a characteristic jagged pattern which was also observed in mouse and dog genomes. By analyzing the nucleotide composition of intergenic sequences larger than 50 kb by cumulative skew diagrams, Hou et al. (2006) found replication-associated strand asymmetry in vertebrates including humans. Therefore, they proposed that transcription-associated strand asymmetries masked the replication-associated ones in the human genome. Huvet et al. (2007) found with multi-scale analysis that the base skew profile presented characteristic patterns consisting of successions of N-shaped structures in more than one-quarter of the human genome. These N domains are bordered by putative replication origins. Wang et al. (2008) illustrated that transcription-associated strand compositional asymmetries and replication-associated ones coexist in most vertebrate (including human) large genes although in most cases the former conceals the latter. The three most frequent types of asymmetric substitution, C to T, A to G, and G to T, were examined in the human genome (Mugal et al., 2009). All three rates were found to be on average higher on the coding strands than on the transcribed. Such finding points to the simultaneous action of rate increasing effects on the coding strands, such as increased adenine and cytosine deamination, and transcription-coupled repair as a rate-reducing effect on the transcribed strands. Furthermore, the author showed that the rate asymmetries of genes are to some extent also produced by the process of replication, depending on the distance to the next ORI and the relative direction of transcription and replication (Mugal et al., 2009). With the help of the very recently published work by Chen et al. (2011), we conclude that strand composition asymmetry (bias) is the superposed effect of replication and transcription asymmetries in the human genome. Among them, transcription associated mutation and or repair bias exert effects on transcribed regions. However, replication induced mutation and repair biases act on the whole chromosome. This is quite similar to bacterial genomes.

As for eukaryotic organelles, there are quite a few reports of strand bias. For example, Seligmann and colleagues observed strand asymmetric gradients in various mitochondria and investigated in the past five years how properties of replication origins affect the gradients (Seligmann, 2010; Seligmann and Krishnan, 2006; Seligmann et al., 2006a, 2006b).

Regarding archaea, a few have shown significant strand composition skews, which are associated with replication. Among them, some are determined or predicted to contain a single replication origin, while others have multiple origins of replication, similar to eukaryotes. According to Necsulea and Lobry (2007), 18 out of 29 archaeal chromosomes showed significant effects of replication on nucleotide skews

Usually, it is believed that bacterial plasmids replicate using a different mechanism than that of the chromosome of their host cell. In 2000, cumulative skew diagrams showed that plasmid and chromosome of *B. burgdorferi* adopted a similar bi-directional replication (Picardeau et al., 2000). Recently, our group performed skew analysis on the largest plasmid of *L. Intracellularis*. As shown in Figure 7, the cumuliatve GC-skew diagram shows two peaks, at two points, around 27 kb and 115 kb in the replicon. And this suggests that the plasmid replicates bi-directionally from an internal origin as the chromosome does. Leading strands and lagging strands are hence determined based on the putative origin and terminus. Result of COA shows that genes on the two replicating strands have distinct codon usages. Note that similar results were observed in genes on the chromosome. Based on these two facts, we suppose that common asymmetric replication would be involved in

the chromosome and the largest plasmid of *L. intracellularis*. Not only both replicate bi-directionally from an internal origin, but also they have biased mutation/repair rates between the two replicating strands. Recently, Arakawa et al. (2009) performed thorough analyses on skew profiles of hundreds of plasmids. Their results suggested the existence of rolling-circle replication in plasmids. Correlation of skew strength between plasmids and their corresponding host chromosomes, which was observed by the authors on 302 host chromosomes and 606 plasmids, suggested that within the same strain, these replicons had reproduced using the same replication machinery.

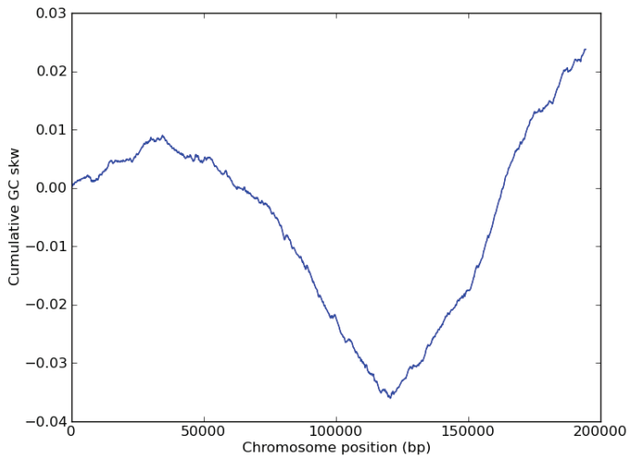

Fig. 7. Cumulative GC skew for the largest plasmid of *L. Intracellularis*.

4. Conclusion and future research

Strand composition bias has been found in various genomes for 20 years. The cause of base composition bias in bacterial genomes is supposed to be the superposed effect of replication and transcription asymmetries in mutation biases. In some species, the former mechanism is mainly responsible for the bias, while in some others the latter constitutes the major force driving the bias. In others, the two mechanisms have equally important effects. Transcription-associated asymmetries can either increase or decrease replication-associated strand asymmetries, depending on the transcription direction and the position of the gene relative to the origin of replication. Theoretically predicting replication origins is one of the practical applications of the universal phenomenon of strand composition bias in bacterial genomes. Future work should focus on the following aspects: (1) Investigation of the common characters and mechanisms of the biases between prokaryotic and eukaryotic genomes; (2) The cause for the varying strength of composition bias in different bacterial genomes; (3) More works should be performed on strand composition bias in eukaryotes other than *Homo* sapiens and in archaeal genomes.

5. Acknowledgment

The present study was supported by the National Natural Science Foundation of China (grant 60801058 and 31071109), and the Fundamental Research Fund for the Central Universities of China (grant ZYGX2009J082).

6. References

Arakawa K., Suzuki H. & Tomita M. (2009). Quantitative analysis of replication-related mutation and selection pressures in bacterial chromosomes and plasmids using generalised GC skew index. BMC Genomics, Vol. 10, (December 2009), 10: 640, ISSN 1471-2164

Asakawa S., Kumazawa Y., Araki T., Himeno H., Miura K. & Watanabe K. (1991). Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. Journal of Molecular Evolution, Vol. 32, No. 6, (June 1991), pp. 511-520, ISSN 0022-2844

Beletskii A. & Bhagwat A.S. (1996). Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in Escherichia coli. Proceedings of the National Academy of Sciences of the United States of America, Vol. 93, No. 24, (November 1996), pp. 13919-13924, ISSN 0027-8424

Bulmer M. (1991). Strand symmetry of mutation rates in the beta-globin region. Journal of Molecular Evolution, Vol. 33, No. 4, (October 1991), pp. 305-310, ISSN 0022-2844

Chargaff E. (1950). Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experientia, Vol. 6, No. 6, (June 1950), pp. 201-209, ISSN 0014-4754

Chen C.L., Duquenne L., Audit B., Guilbaud G., Rappailles A., Baker A., Huvet M., d'Aubenton-Carafa Y., Hyrien O., Arneodo A., & Thermes C. (2011). Replication-associated mutational asymmetry in the human genome. Molecular Biology Evolution, [Epub ahead of print], (March 2011), ISSN 0737-4038

Cornish-Bowden A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: Recommendation 1984. Nucleic Acids Research, Vol. 13, Issue 9, (May 1985), pp. 3021-3030, ISSN 0305-1048

Filipski J. (1990). Evolution of DNA sequence, contributions of mutational bias and selection to the origin of chromosomal compartments. Advances in Mutagenesis Research 2, Publisher Springer, (February 1991), pp. 1–54, ISBN-10 3540524281

Frank A.C. & Lobry J.R. (1999). Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. Gene, Vol. 238, Issue 1, (September 1999), pp. 65-77, ISSN 0378-1119

Francino M.P., Chao L., Riley M.A., & Ochman H. (1996). Asymmetries generated by transcription-coupled repair in enterobacterial genes. Science, Vol. 272, No. 5258, (April 1996), pp. 107-109, ISSN 0036-8075

Francino M.P. & Ochman H. (2000). Strand symmetry around the beta globin origin of replication in primates. Molecular Biology Evolution, Vol. 17, Issue 3, (March 2000), pp. 416–422, ISSN 0737-4038

Gierlik A., Kowalczuk M., Mackiewicz P., Dudek M.R., & Cebrat S. (2000). Is there replication-associated mutational pressure in the Saccharomyces cerevisiae genome? Journal of Theoretical Biology, Vol. 202, Issue 4, (February 2000), pp. 305–314, ISSN 0022-5193

Green P., Ewing B., Miller W., Thomas P.J., NISC Comparative Sequencing Program & Green E.D. (2003). Transcription-associated mutational asymmetry in mammalian evolution. Nature Genetics, Vol. 33, No. 4, (April 2003), pp. 514-517, ISSN 1061-4036

Grigoriev A. (1998). Analyzing genomes with cumulative skew diagrams. Nucleic Acids Research, Vol. 26, Issue 10, (April 1998), pp. 2286-2290, ISSN 0305-1048s

Guo F.B. & Yu X.J. (2007). Separate base usages of genes located on the leading and lagging strands in Chlamydia muridarum revealed by the Z curve method. BMC Genomics, Vol. 8, (October 2007), 8: 366, ISSN 1471-2164

Guo F.B. & Yuan J.B. (2009). Codon usages of genes on chromosome, and surprisingly, genes in plasmid are primarily affected by strand-specific mutational biases in Lawsonia intracellularis. DNA Research, Vol. 16, Issue 2, (January 2009), pp. 91-104, ISSN 1340-2838.

Hou W.R., Wang H.F. & Niu D.K. (2006). Replication-associated strand asymmetries in vertebrate genomes and implications for replicon size, DNA replication origin, and termination. Biochemical Biophysical Research Communications, Vol. 344, Issue 4, (June 2006), pp. 1258-1262, ISSN 0006-291X

Huvet M., Nicolay S., Touchon M., Audit B., d'Aubenton-Carafa Y., Arneodo A. & Thermes C. (2007). Human gene organization driven by the coordination of replication and transcription. Genome Research, Vol. 17, Issue 9, (September 2007), pp. 1278-1285, ISSN 1465-6906

Klasson L. & Andersson S.G. (2006). Strong asymmetric mutation bias in endosymbiont genomes coincide with loss of genes for replication restart pathways. Molecular Biology Evolution, Vol. 23, Issue 5, (February 2006), pp. 1031-1039, ISSN 0737-4038

Krishnan, N.M, Seligmann, H., Raina, S.Z., Pollock, D.D. (2004) Detecting gradients of asymmetry in site-specific substitutions in mitochondrial genomes. DNA Cell Biol. 23, 707–714.

Lin H.J. & Chargaff E. (1967). On the denaturation of deoxyribonucleic acid: II. Effects of concentration. Biochimica Biophysica Acta, Vol. 145, Issue 2, (September 1967), pp. 398-409, ISSN 0006-3002

Lobry, J.R. (1995). Properties of a general model of DNA evolution under no-strand-bias conditions. Journal of Molecular Evolution, Vol. 40, No. 3, (December 1994), pp. 326–30, ISSN 0022-2844

Lobry J.R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. Molecular Biology Evolution, Vol. 13, No. 5, (May 1996), pp. 660–665, ISSN 0737-4038

Lobry J.R. & Sueoka N. (2002). Asymmetric directional mutation pressures in bacteria. Genome Biology, Vol. 3, Issue 10, (September 2002), research0058.1-research0058.14, ISSN 1465-6906

Mcinerney J.O. (1998). Replicational and transcriptional selection on codon usage in Borrelia burgdoreferi. Proceedings of the National Academy of Sciences of the United States of America, Vol 95, No. 18, (June 1998), pp. 10698-10703, ISSN 0027-8424

McLean J.M., Wolfe K.H. & Devine K.M. (1998). Base composition skews, replication orientation and gene orientation in 12 prokaryote genomes. Journal of Molecular Evolution, Vol 47, No. 6, (June 1998), pp. 691–696, ISSN 0022-2844

Mrazek, J., Karlin, S. (1998) Strand compositional asymmetry in bacterial and large viral genomes. Proc. Nat. Acad. Sci. USA 95, 3720-3725.

Mugal C.F., von Grünberg H.H. & Peifer M. (2009). Transcription-induced mutational strand bias and its effect on substitution rates in human genes. Molecular Biology Evolution, Vol .26, Issue 1, (October 2008), pp. 131-142, ISSN 0737-4038

Necsulea A. & Lobry J.R. (2007). A new method for assessing the effect of replication on DNA base composition asymmetry. Molecular Biology Evolution, Vol. 24, Issue 10, (July 2007), pp. 2169-79, ISSN 0737-4038

Nikolaou C. & Almirantis Y. (2005). A study on the correlation of nucleotide skews and the positioning of the origin of replication: different modes of replication in bacterial species. Nucleic Acids Research, Vol. 33, Issue 21, (November 2005), pp. 6816-6822, ISSN 0305-1048

Niu D.K., Lin K. & Zhang D.Y. (2003). Strand compositional asymmetries of nuclear DNA in eukaryotes. Journal of Molecular Evolutian, Vol. 57, No. 3 (April 2003), pp. 325-334, ISSN 0022-2836

Qu H., Wu H., Zhang T., Zhang Z., Hu S., & Yu J. (2010), Nucleotide compositional asymmetry between the leading and lagging strands of eubacterial genomes. Research in Microbiology, Vol. 161, Issue 10, (December 2010), pp. 838-846, ISSN 0923-2508

Picardeau M., Lobry J.R. & Hinnebusch B.J. (2000), Analyzing DNA strand compositional asymmetry to identify candidate replication origins of Borrelia burgdorferi linear and circular plasmids. Genome Research, Vol 10, (July 2000), pp. 1594-604, ISSN 1088-9051

Powdel B.R., Satapathy S.S., Kumar A., Jha P.K., Buragohain A.K., Borah M. & Ray S.K. (2009). A study in entire chromosomes of violations of the intra-strand parity of complementary nucleotides (Chargaff's second parity rule). DNA Research, Vol. 16, (October 2009), pp. 325-343, ISSN 1340-2838

Rocha E.P.C. (2004). The replication-related organization of bacterial genomes. Microbiology, Vol. 150, (2004), pp. 1609-1627, ISSN 1350-0872

Seligmann, H. (2010) The ambush hypothesis at the whole-organism level: Off frame, 'hidden' stops in vertebrate mitochondrial genes increase developmental stability. Comp. Biol. Chem. 34, 80-85.

Seligmann, H., Krishnan, N.M. (2006) Mitochondrial replication origin stability and propensity of adjacent tRNA genes to form putative replication origins increase developmental stability in lizards. J. Exp. Zool. 306B, 433-439.

Seligmann, H., Krishnan, N.M., Rao, B.J. (2006a) Mitochondrial tRNA sequences as unusual replication origins: Pathogenic implications for Homo sapiens. J. Theor. Biol. 243, 375-385.

Seligmann, H., Krishnan, N.M., Rao, B.J. (2006b) Possible multiple origins of replication in Primate mitochondria: alternative role of tRNA sequences. J. Theor. Biol. 241, 321-332.

Tillier E.R. & Collins R.A. (2000). The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. Journal of Molecular Evolution, Vol 50, No. 3, (March 2000), pp. 249–257, ISSN 0022-2836

Touchon M., Nicolay S., Arneodo A., d'Aubenton-Carafa Y. & Thermes C. (2003). Transcription-coupled TA and GC strand asymmetries in the human genome. FEBS Letters, Vol. 555, Issue 3, (December 2003), pp. 579-582, ISSN 0014-5793

Touchon M., Nicolay S., Audit B., Brodie of Brodie E.B, d'Aubenton-Carafa Y. Arneodo A., & Thermes C. (2005). Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. Proceedings of the National

Academy of Sciences of the United States of America, Vol 102, No. 28, (July 2005), pp. 9836-9841, ISSN 0027-8424

Wang H.F., Hou W.R. & Niu D.K. (2008). Strand compositional asymmetries in vertebrate large gens. Molecular Biology Reports, Vol 35, (April 2007), pp. 163-169, ISSN 1573-4978

Watson J.D. & Crick F.H.C. (1953). Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. Nature, Vol 171, Issue 4356, (April 1953), pp. 737-738, ISSN 0028-0836

Wei W. & Guo F.B. (2010). Strong strand composition bias in the genome of Ehrlichia canis revealed by multiple methods. The Open Microbiology Journal, Vol 4, (October 2010), pp. 98-102, ISSN 1874-2858

Worning P., Jensen L.J., Hallin P.F., Staerfeldt H.H. & Ussery D.W. (2006). Origin of replication in circular prokaryotic chromosomes. Environmental Microbiology, Vol. 8, Issue 2, (February 2006), pp. 353–361, ISSN 1462-2912

Wu C.I. & Maeda N. (1987). Inequality in mutation-rates of the two strands of DNA. Nature, Vol 327, Issue 6118, (May 1987), pp. 169-170, ISSN 0028-0836

Zhang R. & Zhang C.T. (2003). Multiple replication origins of the archaeon Halobacterium species NRC-1. Biochemical and Biophysical Research Communications, Vol 302, Issue 4, (March 2003), pp. 728-734, ISSN 0006-291X

DNA Replication-Current Advances

Edited by Dr Herve Seligmann

ISBN 978-953-307-593-8
Hard cover, 694 pages
Publisher InTech
Published online 01, August, 2011
Published in print edition August, 2011

The study of DNA advanced human knowledge in a way comparable to the major theories in physics, surpassed only by discoveries such as fire or the number zero. However, it also created conceptual shortcuts, beliefs and misunderstandings that obscure the natural phenomena, hindering its better understanding. The deep conviction that no human knowledge is perfect, but only perfectible, should function as a fair safeguard against scientific dogmatism and enable open discussion. With this aim, this book will offer to its readers 30 chapters on current trends in the field of DNA replication. As several contributions in this book show, the study of DNA will continue for a while to be a leading front of scientific activities.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Feng-Biao Guo and Lu-Wen Ning (2011). Strand-specific Composition Bias in Bacterial Genomes, DNA Replication-Current Advances, Dr Herve Seligmann (Ed.), ISBN: 978-953-307-593-8, InTech, Available from: http://www.intechopen.com/books/dna-replication-current-advances/strand-specific-composition-bias-in-bacterial-genomes

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen