GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Review

Selectional versus mutational mechanism underlying genomic features of bacterial strand asymmetry: a case study in *Clostridium acetobutylicum*

H.-L. Zhao, Z.-K. Xia, Z.-G. Hua and W. Wei

Center of Bioinformatics and Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, China

Corresponding author: W. Wei
E-mail: unheard.wei@gmail.com

Genet. Mol. Res. 14 (1): 1911-1925 (2015)
Received January 2, 2014
Accepted March 25, 2014
Published March 20, 2015
DOI http://dx.doi.org/10.4238/2015.March.20.1

ABSTRACT. Strand biases are widespread in bacterial genomes. In this review, we discuss 5 types of bias, including gene orientation, the number of open reading frames, nucleotide composition, substitution rate, and gene length, between leading and lagging strands during replication. For each type of strand bias, related studies were summarized and *Clostridium acetobutylicum* ATCC 824 was used as a representative example to illustrate bias. Our results in *C. acetobutylicum* indicate that there is little asymmetry between 2 replication strands on open reading frame number and gene length, whereas the other 3 features presented significant strand bias. The underlying mechanisms of mutation and/or selection are discussed. It is hoped that this review will improve the understanding of the extent and reasons for various types of strand bias in bacterial genomes.

Key words: *Clostridium acetobutylicum*; Gene orientation bias; Nucleotide composition bias; Strand bias; Substitution rate bias

INTRODUCTION

Bacterial chromosome replication typically starts at a defined origin from which 2 replication forks propagate in opposite directions. The process is semi-conservative; 2 strands of the parental duplex separate at the replication fork and serve as templates for the synthesis of a new cognate strand by DNA polymerases. The parental duplex is replaced by 2 daughter duplexes, each of which consists of 1 parental strand and 1 newly synthesized strand. The DNA double helix is anti-parallel, with nucleotides added only to the 3' end of the growing chain; DNA polymerases can only catalyze synthesis in the 5'-3' direction. Thus, the 5'-3' strand (leading strand) is continuously synthesized in the same direction as the movement of replication fork. However, the lagging strand replicates through the synthesis of relatively smaller chains segments (known as Okazaki fragments), which are then joined together to form an integrated strand. Replication continues until a termination signal is reached or the 2 replication forks meet. Synthesis of DNA in different directions leads to various asymmetric genomic characteristics between the 2 replicating strands. Various studies have been conducted to examine the mechanism of this process (Rocha, 2002, 2004; Guo, 2012).

Clostridium acetobutylicum is most frequently found in soil, although it can also survive in a number of different environments. Because this organism is capable of breaking down sugar, it is referred to as saccharolytic; it also produces a number of commercially useful products including acetone, ethanol, and butanol (Nolling et al., 2001). Thus, this bacterium is industrially important. The genome sequence and corresponding annotation file of *C. acetobutylicum* ATCC 824 were extracted from GenBank (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/); the origin and terminus of replication were annotated using the Doric database (http://tubic.tju.edu.cn/doric/). We used this information to identify the orientation of all genes in order to determine whether the gene is located on the leading or lagging strand.

RESULTS AND DISCUSSION

Gene orientation bias

Gene orientation bias occurs when there is an unequal distribution of genes between the leading and lagging strands. We discuss this asymmetry considering multiple aspects in the following sections.

Whole genome

Previous studies

Accompanied by increased analysis of available complete genomic sequences, researchers have shown that the extent of bias in gene orientation varies widely among species. The first systematic survey of gene strand bias revealed that genomes contain 55-80% of genes in the leading strand. In *Escherichia coli*, the frequency of genes present on the leading strand is 54%, while this value is 74% in *Bacillus subtilis* (McLean et al., 1998).

Example

In *C. acetobutylicum* ATCC 824, 2900 genes are located on the leading strand, where-

as 771 genes are present on the lagging strand.

Highly expressed genes

Previous studies

As early as 1977, Nomura and Morgan (1977) analyzed the *E. coli* chromosome and found that genes coding for ribosomal proteins and rRNAs were generally localized on the leading strand. In 1987, Sharp and Li (1987) found that the highly expressed genes are preferentially distributed on the leading strand. According to McLean et al. (1998), a peculiarity that appeared to be shared by most bacterial species was that orientation bias was greater for highly expressed genes. The following year, Karlin (1999) also concluded that in most cases, highly expressed genes are overrepresented on the leading strand.

In contrast, Rocha (2002) found that expression level was not a determinant of gene strand bias (Rocha, 2002); rather, essentiality was considered to be more important, as observed in *B. subtilis* and *E. coli* (Rocha and Danchin, 2003). Furthermore, Hu et al. (2007) discovered that highly expressed genes accounted for less than 50% of the leading strand in 30 of 211 prokaryotic genomes examined. They also found that highly expressed genes do not always preferentially reside on the leading strand.

Example

We defined highly expressed genes based on experimental data (gene expression atlas) as well as theoretical information (codon adaptation index, CAI).

For the gene expression atlas, we selected an experimental sample of a wild-type strain that was grown under favorable environmental conditions. The expression level of the gene was reflected by the value of gene expression atlas, which was obtained conveniently from the array express archive database (http://www.ebi.ac.uk/arrayexpress/). The top 5% expressed genes were regarded as highly expressed. We counted these genes and found that 159 and 23 highly expressed genes were distributed on the leading and lagging strands, respectively.

The CAI is a measure of codon usage, which uses a reference set of highly expressed genes, typically ribosomal protein genes, from a species to assess the relative merits of each codon. A score for each gene was calculated based on the frequency of use of all codons in that gene (Sharp and Li, 1987). Wu et al. (2005) analyzed the correlation between CAI values and experimental expression levels and showed that CAI can predict highly expressed genes. CAI values vary from 0-1.0. A higher CAI value indicates that an objective gene has a similar codon usage pattern to the reference genes. Thus, we also used CAI to measure gene expression levels. CAI for a gene was calculated using the following formula (Sharp and Li, 1987):

$$CAI = \left(\prod_{k=1}^{L} w_k\right)^{\frac{1}{L}}$$

where w_k is the frequency of the use of a particular codon compared to the frequency of the optimal codon for that amino acid and L is the number of codons in the gene.

In our study, the collection of ribosomal protein genes was chosen as the reference set. We calculated the CAI value of each gene based on the equation above, and then tested the sig-

nificance of our results for genes on the leading and lagging strands ($P < 2.2 \times 10^{-16}$). We also found the mean CAI of leading strand genes (0.67) was higher than that of lagging strand genes (0.64). The results based on theoretical CAI were very consistent with experimental values.

Essential genes

Previous studies

As described above, it is widely known that gene strand bias results from the preference for highly expressed genes in the leading strand (McLean et al., 1998). However Tao et al. (1999) analyzed genes expressed during the exponential growth phase of *E. coli* and found that essential genes, most of which are expressed at lower levels, are highly biased. Nevertheless, nonessential, highly expressed genes were equally distributed between the 2 strands (Tao et al., 1999). Rocha and Danchin (2003) also found that essentiality rather than expressiveness is the basis of gene strand bias based on their analysis of gene distributions in *B. subtilis* and *E. coli*. These results were also observed in Firmicutes and g-Proteobacteria; essential genes were found to be more biased than non-essential genes (Rocha and Danchin, 2003), suggesting that essentiality is the primary determinant of the chromosome structure. In a recent study by Lin et al. (2010), it was found that in 10 bacteria, essential genes were primarily located on the leading strand compared to the lagging strand. In addition, their results suggested that a particular Clusters of Orthologous Groups (COG) functional category plays a key role in shaping the gene strand bias in bacterial genomes.

Example

We predicted essential genes using the Geptop webserver, which first provided an online platform for detecting essential genes (http://cefg.uestc.edu.cn/geptop/) (Wei et al., 2013). We submitted the entire proteome for a bacterial species in FASTA format. This webserver can calculate the essentiality score for each gene. The default cutoff for the essentiality score is 0.15; thus, a gene whose value is more than 0.15 is predicted to be essential. According to our analysis, 297 of 3671 genes were predicted to be essential. A total of 269 genes were located on the leading strand, while only 28 genes were on the lagging strand. The mean essential score of leading strand genes was 0.044, which is greater than that of lagging strand genes (0.024); the 2 sets of essential scores between leading and lagging strands were significant ($P = 1.4 \times 10^{-12}$).

Functional categories

Above, we presented the strand bias of genes sorted according to their expression and essentiality; in this section, we examine the relationship between bias and gene function.

Previous studies

As described above, strand bias for essential genes only emerge for particular COGs (Lin et al., 2010). The information storage and process (J, K, and L), and subcategories D (cell cycle control), M (cell wall biogenesis), O (posttranslational modification), C (energy production and conversion), G (carbohydrate transport and metabolism), E (amino acid transport and

metabolism), and F (nucleotide transport and metabolism) were preferentially located on the leading strand, whereas other COG functional subcategories showed no statistically significant strand bias.

The analysis by Lin et al. (2010) provided valuable insight into the functional preference of genes to leading and lagging strands, but essential genes accounted for only a small portion (~10% in *E. coli* and *B. subtilis*) of the whole bacterial genome (Kobayashi, et al., 2003; Kato and Hashimoto, 2007). Mao et al. (2012) carried out a larger computational analysis examining a larger number of genes and organisms to confirm the generality of gene strand bias. They used Gene Ontology (GO) to define functional categories of genes across 725 bacterial genomes. Their results demonstrated that genes in different functional categories differ in their tendency to be on the leading strand. The variable distribution of genes on the 2 strands were hypothesized to result from 2 balancing forces: the first generally keeps the genome as compact as possible to remain energetically efficient when replicating and maintaining the genome, while the second drives genes of certain functional categories to leading strands to make the bacteria more efficient when responding to environmental changes. Therefore, we hypothesize that the percentage of genes belonging to different functional categories that are asymmetric between the 2 strands is subject to selection pressure.

Example

We counted the occurrence of each type of COG. Gene classification information regarding COGs can be obtained from the annotation file of a species, and all types of COGs can be acquired from the COGs database (http://www.ncbi.nlm.nih.gov/COG/). Our data are displayed in Table 1. As shown in the table, different COGs prefer different strands; extensive testing of the data ($P = 1.4 \times 10^{-3}$) also supported asymmetry between the 2 strands.

Table 1. Numbers of various type of COGs between leading and lagging strands.

COG	Lead	Lag	COG	Lead	Lag
C	90	23	I	43	13
CHR	2	1	IQ	7	0
D	36	5	J	148	11
E	116	41	K	130	62
EF	3	0	KE	1	1
EG	1	1	KG	7	0
EH	9	1	KL	3	0
EM	0	2	KT	6	4
EJ	1	0	L	97	15
EP	14	3	LK	2	0
ER	4	1	LKJ	3	1
F	60	8	LR	0	2
FE	2	0	M	123	34
FGR	2	0	MG	6	1
FJ	1	0	MJ	2	0
FP	0	1	N	23	1
FR	0	1	NT	6	0
G	141	43	O	59	14
GC	3	0	P	61	24
GE	1	1	PH	1	2
GER	3	0	Q	8	2
GT	2	0	R	228	56
H	86	19	S	118	38
HC	2	1	T	66	10
HI	1	0	TK	29	3
HR	2	1	TQ	2	0

Underlying mechanism of gene orientation bias

In general, most genes are preferentially distributed on the leading strand. This bias may result from the collisions between DNA polymerase and RNA polymerase. Chromosomal replication often occurs during highly active transcription when both polymerases are bound to the same template, and thus collisions between the 2 polymerases are inevitable (Rocha, 2004). In addition, both polymerases progress in the same 5'-3' direction, and transcribed genes on the leading strand lead to co-oriented collisions, while genes on the lagging strand bring about head-on collisions. Thus, there are different consequences of collision events between the 2 polymerases, leading to an asymmetric distribution of genes between the 2 strands. Co-oriented collisions may significantly slow replication, but the transcript may be completed. However, the head-on collisions may retard replication and result in the production of aborted transcripts. These may be translated into truncated non-functional peptides, which is likely to be deleterious for cellular activity (Rocha, 2004). Consequently, the collision model is thought to force a higher gene density onto the leading strand (Rocha, 2008), indicating that selective pressure contributes to gene strand bias.

Open reading frame (ORF) number bias

There have been some studies examining the bias of ORFs between leading and lagging strands. Moreover, to further verify that gene orientation bias is the result of selection rather than mutation, we used ORF number as a research object. If gene orientation bias is caused by mutation, the ORF number should exhibit similar strand bias, whereas if the appearance of ORFs is approximate between leading and lagging strands, selection would be confirmed to play a key role in gene orientation bias.

The ORF is a portion of a reading frame containing no stop codons. When a new gene is identified and its DNA sequence is determined, its corresponding protein sequence is unclear. The DNA sequence can be translated or read in 6 possible reading frames (3 for each strand, in line with 3 different start positions for the 1st codon). Identifying an ORF provides the first evidence that a new sequence of DNA is part or all of a gene encoding a particular protein. The idea of a coding sequence (CDS) is similar to the ORF. Overall, ORF is only a potential coding sequence that is typically predicted based on the DNA sequence and could be not transcribed, whereas CDS is the segment of the DNA that is translated to encode proteins (http://www.biostars.org/p/47022/).

Example

We defined a sequence of DNA beginning with the start codon ATG and ending with any of the 3 termination codons (TAA, TAG, TGA) as an ORF. We confirmed the ORF with different minimum lengths ranging from 90-300 base pairs (bp) in 30-bp steps, and then distinguished the positions of these ORFs similarly to the orientation of genes. Details regarding the ORFs with various lengths between leading and lagging strands are shown in Table 2. The data revealed no clear difference (P = 0.40) in the number of ORFs.

Table 2. Numbers of ORF with different minimum length between two strands.

Strand	90 bp	120 bp	150 bp	180 bp	210 bp	240 bp	270 bp	300 bp
Lead	6313	4263	3257	2677	2326	2052	1902	1780
Lag	6441	4309	3246	2675	2305	2049	1887	1785

Nucleotide composition bias

Chargaff parity rule 2 (PR2) states that the intrastrand nucleotide composition should be such that A = T and C = G when there is not any strand bias such as mutation or selection. However, rapid development of the genomic era has revealed that an increasing number of bacteria show bias of PR2 between replication-associated leading and lagging strands (Arakawa et al., 2009).

Single nucleotide composition

Previous studies

PR2 bias was first identified in the genomes of echinoderm and vertebrate mitochondria (Asakawa et al., 1991). Subsequently, Lobry (1996) analyzed the genomes of *E. coli*, *B. subtilis*, and *Haemophilus influenzae*, and found their nucleotide composition to be asymmetric. In these genomes, the leading strands are relatively enriched in G compared to C (GC skew) and T compared to A (AT skew) (Arakawa et al., 2009). Necşulea and Lobry (2007) sequenced additional bacterial genomes to analyze nucleotide composition bias. Their results indicated that 311 of 360 genomes contained excess G over C in the leading strands. A recent study supported the ubiquity of strand composition asymmetry in prokaryotic genomes (Xia, 2012). GC skew was proposed to predict replication origin in bacterial genomes (Lobry, 1996) and has been widely used. This value can be calculated using the equation: (G - C)/(G + C), where G and C denote the occurrences of the corresponding nucleotides in a given sequence (Rocha, 2004).

Example

We determined the absolute frequency per base (A, T, C, G) in each codon position (1st, 2nd, 3rd) for every gene, and then processed these data using principal component analysis (PCA). We used the first 2 principal components to draw a scatter plot and found that the single nucleotide composition of genes located on leading and lagging strands were separated with only slight overlap (Figure 1) and that there was a clear distinction between the 2 strands. The difference rate was 97% using a 60% confidence interval.

Segregated codon usage

Previous studies

In some bacteria, sufficiently strong single nucleotide composition bias can to lead to separate codon usage of genes. In 1998, the codon usages of all genes in *Borrelia burgdorferi*

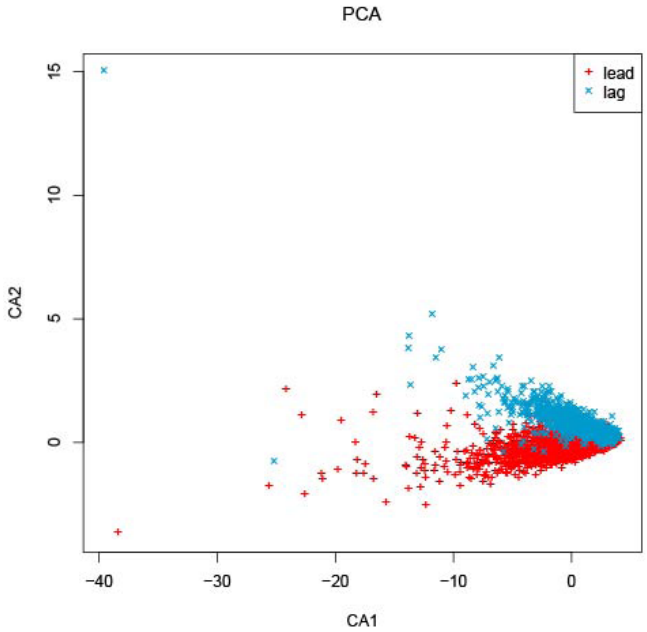

Figure 1. PCA for single nucleotide composition. First (CA1) and second (CA2) principal component scores of PCA for 12 single nucleotide variables determined for 3671 *Clostridium acetobutylicum* genes are shown. Each element represents a gene, red '+' symbols correspond to values of leading genes, and blue 'x' symbols are those of lagging genes. The difference rate between the 2 sets of data was 97% using a 60% confidence interval.

were studied by correspondence analysis (CA). The results suggested that the 2 strands were quite distinct at the codon level (McInerney, 1998). This was the first observation of separate codon usage associated with replication in bacterial genomes. Subsequently, several researchers showed the same results in different species (Lafay, et al., 1999; Romero, et al., 2000; Das, et al., 2005). These analyses were based on the relative synonymous codon usage-CA method (Perrière and Thioulouse, 2002), which can only reflect relative synonymous codon usage. Wei and Guo (2010) adopted the Z curve, PR2-plot, and relative synonymous codon usage-CA to fully analyze the *Ehrlichia canis* genome. Their results revealed clear divided codon usage of genes on the 2 replicating strands. Eleven intracellular bacteria, among which 7 belong to obligate and 4 belong to facultative species, have been found to contain separate codon usages based on whether the genes are located on leading or lagging strands (Das et al., 2005; Guo and Yu, 2007; Guo and Yuan, 2009; Dutta and Paul, 2012). With respect to the extremely strong nucleotide composition bias in obligate intracellular parasites, Guo and Ning (2011) speculated that these species live in the cells of their hosts, and thus their habitats are relatively safe and some genes coding for DNA repair enzymes may be lost during long-term evolution. Such mutations generated during replication may accumulate and be an important cause of composition bias (Guo and Ning, 2011). Moreover, some researchers found that the genomes of obligate intracellular parasites include 2 common characteristics: small genome size and low genomic G+C content (Rocha and Danchin, 2002). For the first characteristic, Guo and Ning (2011) suggested that in small bacterial genomes that have suffered reductive evolution, repair mechanisms for replication may be inefficient. However, in larger bacterial

genomes, it is difficult for mutation pressure to exceed translational selection. For the second characteristic, Rocha and Danchin (2002) considered that the cost of energy metabolism is lower in AT-rich genomes, so that competition for metabolic resources with hosts may lead to the higher frequency of AT in intracellular bacteria.

Example

We examined the codon usage for all genes using within-group correspondence analysis (WCA). WCA adjusts the value for each codon based on the average value of all codons encoding the same amino acid (Suzuki et al., 2008).

We used this method to analyze a data matrix of codons, which involve all genes (rows) and 59 codons (columns). The results showed that the codon usage of genes was more biased between the 2 strands, with their divergence reaching 98% using a 60% confidence interval. The results are illustrated in Figure 2 and imply a completely different codon usage pattern.

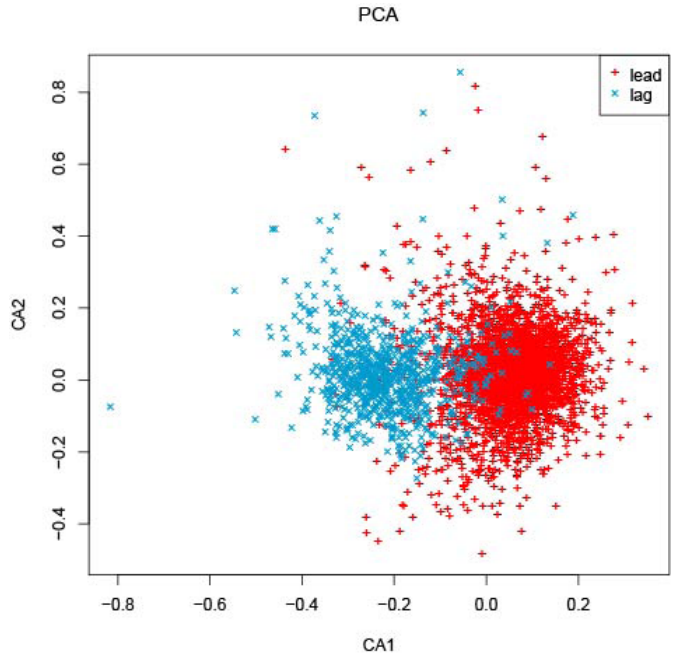

Figure 2. WCA of codon usage. First (CA1) and second (CA2) principal component scores of WCA for 59 codon variables determined for 3671 *Clostridium acetobutylicum* genes are shown. Each element represents a gene, red '+' symbols correspond to the values of leading genes, and blue 'x' symbols are those of lagging genes. The difference rate between the 2 sets of data was 98% using a 60% confidence interval.

Underlying mechanisms of nucleotide composition bias

Because most bacterial genomes have been found to contain significant nucleotide composition bias, it is necessary to determine the underlying mechanisms of such asymmetry. There are 2 alternative hypotheses (Necşulea and Lobry, 2007). The first hypothesis is related

to replication. Replication is a dissymmetric process, and the 2 strands are synthetized by separate polymerases, which allows for variation in the error rate between the 2 strands (Kunkel, 1992; Stillman, 1994). In addition, the structurally asymmetrical replication fork gives rise to different forms of damage to the template strand during replication (Trinh and Sinden, 1991). The second hypothesis is associated with transcription. Francino et al. (1996) analyzed genes of *E. coli* and found that the substitution patterns were similar between the leading and lagging strands, whereas these patterns were different between coding and non-coding strands. Francino and Ochman (1997) suggested that composition bias results when transcription overexposes the non-transcribed strand to DNA damage while targeting repair enzymes to the transcribed strand. Thus, the asymmetric transcription process can bias the occurrence of mutations between the transcription strands (Francino and Ochman, 1997), and most protein-coding genes are located on the leading strands as described above. In addition, some studies found that gene orientation bias was positively related to nucleotide composition bias (Hu et al., 2007; Wu et al., 2012). Thus, the asymmetric transcription mechanism may also bring about composition bias when combined with gene orientation bias.

Regardless of whether nucleotide composition bias is related to replication or transcription, mutation is active during the process. In particular, cytosine, which is the most unstable of the 4 bases of nucleic acids, easily deaminates to uracil. If a resulting uracil is not replaced with cytosine, a C to T mutation results. Cytosine deaminates at a rate of 3-7 x 10^{-13}/s in double-stranded DNA (Frederico, et al., 1990). However, this rate increases by 140 times in single-stranded DNA relative to double-stranded DNA (Beletskii and Bhagwat, 1996). Consequently, mutations resulting from the chemical instability of bases in single-stranded DNA are responsible for nucleotide composition asymmetry.

Substitution rate bias

The impetus of genetic evolution is the nucleotide substitution, which may result in changes to the genetic code and hereditary information (Nei and Kumar, 2000), and the nucleotide substitution rate varies among genes. There are 2 outcomes to substitutions: synonymous substitution does not change the amino acid sequence of a protein or nonsynonymous substitution modifies the amino acid sequence.

Previous studies

Any significant deviation from the intrastrand A = T or C = G relationships implies that there is asymmetry in the substitution patterns between the leading and lagging strands (Lobry, 1996). Many bacteria show this deviation; therefore, we referred to studies regarding asymmetric substitution. Wu and Maeda (1987) investigated the rate bias of substitution between homologous sequences of the beta-globin complex in 6 primates. Their comparison to the substitution rates of complementary nucleotides was the first observation of strand asymmetry. Francino et al. (1996) also studied asymmetric substitution of several genes in *E. coli*. They found no differences in substitution rates between the leading and lagging strands, but significant deviation between coding and non-coding strands. A somewhat contradictory result was found by Rocha et al. (2006). They evaluated substitution bias between the 2 strands on a genome-wide scale in 7 bacteria and found that clear bias existed in all of the studied genomes.

Example

Evolutionary analysis requires a pair of strains, and we selected *Clostridium beijerinckii* NCIMB 8052. Although a different strain, this species belongs to *C. acetobutylicum*. To calculate evolutionary rates, related files of the 2 genomes were retrieved from GenBank (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/).

The orthologous gene pairs between the 2 genomes were identified based on the reciprocal best hit using the Blast program. Protein sequences encoded by the orthologous gene pairs were aligned using ClustalW with default parameters (Thompson et al., 1994) and were then back-translated into nucleotide sequences. The number of nonsynonymous substitution sites was computed following Yang's (2007) definition using the PAML package. We calculated the *Ka* of orthologous genes using these 3 procedures. The mean *Ka* of leading strand genes was 0.34 and that of the lagging strand genes was 0.37. Moreover, the 2 sets of *Ka* values analyzed using the *t*-test were markedly different ($P = 7.9 \times 10^{-3}$). Based on this result, leading strand genes may be more conserved.

Underlying mechanisms of substitution rate bias

Different substitution rates are observed in different studies. In 1996, Lobry studied 3 prokaryote genomes and found that mutational bias was responsible for asymmetric substitution patterns in the 2 DNA strands. In the absence of any selection bias between the leading and lagging strands, the disparity of replication error was considered to lead to differences in substitution patterns. A similar hypothesis was proposed by Szczepanik et al. (2001) who considered that different rates of nucleotide substitution accumulation on leading and lagging strands implicate qualitative and quantitative differences in the accumulation of mutations in protein coding sequences on different DNA strands. Marín and Xia (2008) presented a substitution model showing that an increased rate of C to T mutation will lead to positive GC skew in 1 strand but negative GC skew in the other. In addition, a recent study of *B. subtilis* showed that the rate of point mutations in core genes on the lagging strand was higher than that on the leading strand, with this difference occurring primarily for nonsynonymous mutations (Paul et al., 2013).

However, Francino and Ochman (1997) proposed that differential selective constraints control much of the variation in substitution rates. At sites under little or no selective constraints, mutations were relatively neutral. Francion and Ochman (1997) confirmed that genes evolve at different rates depending on the strength of selective pressure to maintain their functions. According to Furusawa (2012), living organisms can represent the heredity phase and evolution phase. The fundamental reason for the precise heredity can be attributed to a leading strand of high fidelity and evolution in the lagging strand, which shows low fidelity. Thus, the evolutionary rates of nucleotide substitution may be determined mainly by mutation rates, selection effects, or both.

Gene length bias

Previous studies

Gene length is also an important indicator relevant to other genomic characteristics.

Previous studies have indicated that a strong negative correlation exists between codon usage bias and protein length in some eukaryotes (Duret and Mouchiroud, 1999), whereas a significant positive correlation was observed in *E. coli* genes (Moriyama and Powell, 1998). It was hypothesized that the different relationships between codon usage bias and gene length observed in prokaryotes and eukaryotes may result from different types of selection (Moriyama and Powell, 1998). Ribeiro et al. (2012) found that in *E. coli*, the nucleotide length of a gene affected expression dynamics. Moreover, the length of essential genes was found to be smaller than that of non-essential genes. In addition, the distribution of gene positions is not random; most genes are located on the leading strand. In accordance with a recent analysis, there is a significant positive correlation between increased mutation rates and gene length on the lagging compared with the leading strand (Paul et al., 2013). Together, these results indicate that it is necessary to compare gene length distribution between the two replication strands.

Example

As described above, gene length may affect various cellular activities. There have been no studies examining gene length between the 2 replication strands; thus, we surveyed strand bias by introducing gene lengths as candidate indicators. Our calculations of gene length in target species demonstrated that the mean length of genes between the leading (927 bp) and lagging strands (904 bp) were nearly equal. In addition, the 2 groups of length data were analyzed using the Student *t*-test, but no significant difference was observed (P = 0.39).

CONCLUSIONS

Various strand biases were first identified in our study; we introduced previous studies and performed associative analysis as well as examined each type of bias.

Decades of research have revealed the occurrence of bias and its correlations. Thus, we show the bias that has been identified in previous studies and their underlying mechanisms in Table 3. 1) Regardless of whether the gene strand bias originates in the polymerase collision model, as a response to environmental change, or serves to maintain genome function, the selection forces acting on different genes located appropriate strand. 2) The cause of nucleotide composition bias in bacterial genomes is thought to be the superposed effect of replication and transcription asymmetries in base mutations. 3) Substitution rate bias may be due to the combined action of mutation and selection.

Table 3. Mechanisms of various strand bias.

Mechanism	Gene orientation	ORF	Nucleotide composition	Substitution rate	Gene length
Mutation		-	√	√	-
Selection	√			√	

The results of analysis of *C. acetobutylicum* ATCC 824 are summarized in Table 4. Among the 5 types of strand biases, gene orientation, nucleotide composition, and *Ka* showed significant asymmetries between the leading and lagging strands, which are in agreement with previous studies. The impact of ORF number and gene length on strand bias was studied for

the first time; neither showed deviation between the 2 strands. The approximate numbers of ORFs demonstrate that selection rather than mutation contributes to gene orientation bias. In addition, gene length appears to be balanced regarding replication strands.

Various strand biases are superficial phenomena; however, there are complex mechanisms behind bias, and additional studies are needed to determine the mechanism of leading and lagging strand bias.

Table 4. Various strand composition bias in *Clostridium acetobutylicum* ATCC 824.

Strands	Gene orientation						ORF	Nucleotide composition	*Ka*	Gene length
	Total No.	Highly expressed		Essential		Functional categories				
		No.	CAI	No.	Score					
Leading	2900	159	0.67	269	0.04	Table 1	Table 2	Figures 1 and 2	0.34	927 bp
Lagging	771	23	0.64	28	0.02				0.37	904 bp
t-test (P)	-	-	$<2.2 \times 10^{-16}$	-	1.4×10^{-12}	1.4×10^{-3}	0.40	-	7.9×10^{-3}	0.39

ACKNOWLEDGMENTS

Research supported by the National Natural Science Foundation of China (grant #31071109), the Fundamental Research Funds for the Central Universities of China (#ZYGX2013J101) and the Program for New Century Excellent Talents in University (#NCET-11-0059). The authors are indebted to Dr. Feng-biao Guo for valuable support, inspiring discussion and helps in revising the manuscript.

REFERENCES

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

Asakawa S, Kumazawa Y, Araki T, Himeno H, et al. (1991). Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. *J. Mol. Evol*. 32: 511-520.

Beletskii A and Bhagwat AS (1996). Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. *Proc. Natl. Acad. Sci. U. S. A*. 93: 13919-13924.

Das S, Paul S, Chatterjee S and Dutta C (2005). Codon and amino acid usage in two major human pathogens of genus *Bartonella* - optimization between replicational-transcriptional selection, translational control and cost minimization. *DNA Res*. 12: 91-102.

Duret L and Mouchiroud D (1999). Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A*. 96: 4482-4487.

Dutta C and Paul S (2012). Microbial lifestyle and genome signatures. *Curr. Genomics* 13: 153-162.

Francino MP and Ochman H (1997). Strand asymmetries in DNA evolution. *Trends Genet*. 13: 240-245.

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

Frederico LA, Kunkel TA and Shaw BR (1990). A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry* 29: 2532-2537.

Furusawa M (2012). Implications of fidelity difference between the leading and the lagging strand of DNA for the acceleration of evolution. *Front. Oncol*. 2: 144.

Guo FB (2012). Replicating strand asymmetry in bacterial and eukaryotic genomes. *Curr. Genomics* 13: 2-3.

Guo FB and Yu XJ (2007). Separate base usages of genes located on the leading and lagging strands in *Chlamydia muridarum* revealed by the Z curve method. *BMC Genomics* 8: 366.

Guo FB and Yuan JB (2009). Codon usages of genes on chromosome, and surprisingly, genes in plasmid are primarily affected by strand-specific mutational biases in *Lawsonia intracellularis*. *DNA Res*. 16: 91-104.

Guo FB and Ning LW (2011). Strand-Specific Composition Bias in Bacterial Genomes. In: DNA Replication-Current Advances (Seligmann H, ed.). InTech Press, Croatia. Chapter 5.

Hu J, Zhao X and Yu J (2007). Replication-associated purine asymmetry may contribute to strand-biased gene distribution. *Genomics* 90: 186-194.

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

Kato J and Hashimoto M (2007). Construction of consecutive deletions of the *Escherichia coli* chromosome. *Mol. Syst. Biol*. 3: 132.

Kobayashi K, Ehrlich SD, Albertini A, Amati G, et al. (2003). Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. U. S. A.* 100: 4678-4683.

Kunkel TA (1992). Biological asymmetries and the fidelity of eukaryotic DNA replication. *Bioessays* 14: 303-308.

Lafay B, Lloyd AT, McLean MJ, Devine KM, et al. (1999). Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Res*. 27: 1642-1649.

Lin Y, Gao F and Zhang CT (2010). Functionality of essential genes drives gene strand-bias in bacterial genomes. *Biochem. Biophys. Res. Commun*. 396: 472-476.

Lobry JR (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol*. 13: 660-665.

Mao X, Zhang H, Yin Y and Xu Y (2012). The percentage of bacterial genes on leading versus lagging strands is influenced by multiple balancing forces. *Nucleic Acids Res*. 40: 8210-8218.

Marín A and Xia X (2008). GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: new substitution models incorporating strand bias. *J. Theor. Biol*. 253: 508-513.

McInerney JO (1998). Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proc. Natl. Acad. Sci. U. S. A.* 95: 10698-10703.

McLean MJ, Wolfe KH and Devine KM (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J. Mol. Evol*. 47: 691-696.

Moriyama EN and Powell JR (1998). Gene length and codon usage bias in *Drosophila melanogaster*, *Saccharomyces cerevisiae* and *Escherichia coli*. *Nucleic Acids Res*. 26: 3188-3193.

Necşulea A and Lobry JR (2007). A new method for assessing the effect of replication on DNA base composition asymmetry. *Mol. Biol. Evol*. 24: 2169-2179.

Nei M and Kumar S (2000). Molecular Evolution and Phylogenetics: Oxford University Press, New York.

Nolling J, Breton G, Omelchenko MV, Makarova KS, et al. (2001). Genome sequence and comparative analysis of the solvent-producing bacterium *Clostridium acetobutylicum*. *J. Bacteriol*. 183: 4823-4838.

Nomura M and Morgan EA (1977). Genetics of bacterial ribosomes. *Annu. Rev. Genet*. 11: 297-347.

Paul S, Million-Weaver S, Chattopadhyay S, Sokurenko E, et al. (2013). Accelerated gene evolution through replication-transcription conflicts. *Nature* 495: 512-515.

Perrière G and Thioulouse J (2002). Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res*. 30: 4548-4555.

Ribeiro AS, Häkkinen A and Lloyd-Price J (2012). Effects of gene length on the dynamics of gene expression. *Comput. Biol. Chem*. 41: 1-9.

Rocha EP (2002). Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol*. 10: 393-395.

Rocha EP (2004). The replication-related organization of bacterial genomes. *Microbiology* 150: 1609-1627.

Rocha EP (2008). The organization of the bacterial genome. *Annu. Rev. Genet*. 42: 211-233.

Rocha EP and Danchin A (2002). Base composition bias might result from competition for metabolic resources. *Trends Genet*. 18: 291-294.

Rocha EP and Danchin A (2003). Essentiality, not expressiveness, drives gene-strand bias in bacteria. *Nat. Genet*. 34: 377-378.

Rocha EP, Touchon M and Feil EJ (2006). Similar compositional biases are caused by very different mutational effects. *Genome Res*. 16: 1537-1547.

Romero H, Zavala A and Musto H (2000). Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. *Nucleic Acids Res*. 28: 2084-2090.

Sharp PM and Li WH (1987). The codon Adaptation Index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res*. 15: 1281-1295.

Stillman B (1994). Smart machines at the DNA replication fork. *Cell* 78: 725-728.

Suzuki H, Brown CJ, Forney LJ and Top EM (2008). Comparison of correspondence analysis methods for synonymous codon usage in bacteria. *DNA Res*. 15: 357-365.

Szczepanik D, Mackiewicz P, Kowalczuk M, Gierlik A, et al. (2001). Evolution rates of genes on leading and lagging DNA strands. *J. Mol. Evol*. 52: 426-433.

Tao H, Bausch C, Richmond C, Blattner FR, et al. (1999). Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. *J. Bacteriol*. 181: 6425-6440.

Thompson JD, Higgins DG and Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 22: 4673-4680.

Trinh TQ and Sinden RR (1991). Preferential DNA secondary structure mutagenesis in the lagging strand of replication in *E. coli*. *Nature* 352: 544-547.

Wei W and Guo FB (2010). Strong strand composition bias in the genome of *Ehrlichia canis* revealed by multiple methods. *Open Microbiol. J*. 4: 98-102.

Wei W, Ning LW, Ye YN and Guo FB (2013). Geptop: a gene essentiality prediction tool for sequenced bacterial genomes based on orthology and phylogeny. *PLoS One* 8: e72343.

Wu CI and Maeda N (1987). Inequality in mutation rates of the two strands of DNA. *Nature* 327: 169-170.

Wu G, Culley DE and Zhang W (2005). Predicted highly expressed genes in the genomes of *Streptomyces coelicolor* and *Streptomyces avermitilis* and the implications for their metabolism. *Microbiology* 151: 2175-2187.

Wu H, Qu H, Wan N, Zhang Z, et al. (2012). Strand-biased gene distribution in bacteria is related to both horizontal gene transfer and strand-biased nucleotide composition. *Genomics Proteomics Bioinformatics* 10: 186-196.

Xia X (2012). DNA replication and strand asymmetry in prokaryotic and mitochondrial genomes. *Curr. Genomics* 13: 16-27.

Yang Z (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol*. 24: 1586-1591.