Measures of Compositional Strand Bias Related to Replication Machinery and its Applications

Kazuharu Arakawa* and Masaru Tomita

Institute for Advanced Biosciences, Keio University, Fujisawa 252-8520, Japan

Abstract: The compositional asymmetry of complementary bases in nucleotide sequences implies the existence of a mutational or selectional bias in the two strands of the DNA duplex, which is commonly shaped by strand-specific mechanisms in transcription or replication. Such strand bias in genomes, frequently visualized by GC skew graphs, is used for the computational prediction of transcription start sites and replication origins, as well as for comparative evolutionary genomics studies. The use of measures of compositional strand bias in order to quantify the degree of strand asymmetry is crucial, as it is the basis for determining the applicability of compositional analysis and comparing the strength of the mutational bias in different biological machineries in various species. Here, we review the measures of strand bias that have been proposed to date, including the ΔGC skew, the B_1 index, the predictability score of linear discriminant analysis for gene orientation, the signal-to-noise ratio of the oligonucleotide bias, and the GC skew index. These measures have been predominantly designed for and applied to the analysis of replication-related mutational processes in prokaryotes, but we also give research examples in eukaryotes.

Received on: July 23, 2011 - Revised on: September 10, 2011 - Accepted on: September 20, 2011

Keywords: Nucleotide composition bias, bacterial replication, GC skew, replication-related mutations.

INTRODUCTION

Genomic nucleotide sequences exhibit extraordinary variability in their guanine + cytosine (G + C) compositions, ranging from as low as 16.6% in *Carsonella ruddii* [1] and up to 74.9% in *Anaeromyxobacter dehalogenans* [2]. While the cause for this plasticity is still debated, proposed reasons include energetic [3, 4], genetic [5], biochemical [6], and environmental factors [7, 8] (see [2] for a recent review). Almost all of the genomes sequenced to date show symmetry in the composition of complementary bases: A ≈ T and G ≈ C. This intriguing equilibrium of the complementary bases within a single strand of the DNA duplex is commonly referred to as the Chargaff's second parity rule (PR2) [9]. Chargaff's first parity rule of complementary bases in the duplex DNA molecule [10] defines the base pairing in the Watson-Crick model [11], and Chargaff and coworkers later extended this rule to a single strand of the duplex based on empirical observations [12]. Using hundreds of complete genomes, extensive computational analysis recently showed that PR2 applies to almost all double-stranded DNA genomes that are sufficiently long [13]. Sueoka showed that PR2 is also expected to theoretically hold under no-strand-bias conditions, where the mutation rates are similar between the two strands [14].

PR2, however, is violated when the mutation/selection rates in the two strands of the DNA molecule are not at equilibrium. This would occur, for example, in single-stranded viral genomes and in organellar genomes, such as mitochondria, that harbor a unidirectional mode of replication, where the mutation rates are asymmetric in the two strands [13, 15]. In fact, such deviation from PR2 is almost universal in local regions of genomic sequences. Szybalski and coworkers first noted an excess of purines over pyrimidines in the coding sequences of bacteriophage in 1960s [16], and this transcription-coupled compositional asymmetry (presumably due to coding requirements and transcription-coupled repair and mutagenesis) has generally been confirmed in a variety of species [17-19]. In eubacteria with circular chromosome, replication progresses bidirectionally along the genome from a finite origin until the replication forks meet at a terminus [20]. This theta-type of replication results in two replicating arms of equal lengths, or the replichores, with opposite polarity in nucleotide composition due to asymmetric modes of leading and lagging strand synthesis [21]. While the genome as a whole maintains compositional parity because the opposite polarity of the replichores cancel each other out, each replichore shows a highly asymmetric composition of complementary bases, especially characterized by an excess of G over C in the leading strand [21, 22]. Chromosome replication therefore exerts a strand-specific mutation/selection pressure and results in a number of biases in genomic features in addition to the asymmetric compositional bias [23, 24]. First, coding sequences are preferentially located in the leading strand in most genomes [25], and this gene strand bias is as high as 90% in *Acetohalobium arabaticum* (but there are exceptions [26, 27]). Moreover, base as well as codon composition of genes are also biased depending on the strand [28-32]. Second, a large number of complementary oligonucleotides are asymmetrically distributed in the leading and lagging strands [33]. These oligonucleotides include highly skewed octamers

*Address correspondence to this author at the Institute for Advanced Biosciences, Keio University, Endo 5322, Fujisawa Kanagawa 252-8520, Japan; Tel/Fax: +81-466-47-5099; E-mail: gaou@sfc.keio.ac.jp

1389-2029/12 $58.00+.00

that are overrepresented in *Escherichia coli*, such as the homologous recombination hotspot named the Chi site, which is recognized by the RecBCD recombinase [34-36], and the FtsK-orienting polar sequences (KOPS) that direct FtsK translocase to the *dif* site, where XerCD recombinase binds in order for chromosome dimer resolution *via* homologous recombination [37-39]. Chi-analogues are found in multiple phylums [40], and KOPS-like sequences have been computationally predicted in hundreds of bacteria [41]. Third, strand preference correlates with gene essentiality [42] and lengths [43, 44], and replication further affects gene positioning and the organization of chromosomal domains [45, 46]. While the relationship between the orientation of replication and transcription is still controversial, a similar replication-related strand bias in terms of nucleotide and oligonucleotide composition and gene orientation has also been observed in eukaryotes, including mammals [47-50].

A deviation from PR2 in the nucleotide sequence therefore implies the existence of a mutational or selectional bias in the DNA duplex, and consequently, a biological mechanism that generates the strand bias. Therefore, analysis of the shift-points of strand bias can be used to identify transcription start sites in eukaryotes [51-53] and for the prediction of replication origins and termini in prokaryotes [54-56] and eukaryotes [50]. Strand bias is commonly calculated as the relative excess of a base among the complementary bases [22, 57] using the formula

$$XYskew = (X - Y)/(X + Y) \quad (1)$$

where X and Y represent the complementary bases. While the most frequently used combination of *X* and *Y* are C and G (GC skew), or T and A (AT skew), other combinations of bases are also used, such as purines (G and A) and pyrimidines (C and T), as well as keto bases (G and T) and amino bases (AC) [58]. The set of GC skew and AT skew is alternatively called the Chargaff difference, where the variables in the above formula are often replaced by the IUPAC codes for degenerate nucleotides; i.e., S for G + C, and W for A + T [59]. *X* and *Y* could even be genes on the Watson and Crick strands of DNA that could be used to calculate the gene orientation skew. The use of Euclidean distance from a theoretical parity point is also demonstrated as a combination of the AT and GC skews, expressed as [15]:

$$d(PR2) = \sqrt{ATskew^2 + GCskew^2} \quad (2)$$

Local regions with strand bias are detected by plotting this nucleotide skew along the sequence using sliding windows. Nucleotide skew is close to zero when there is no strand bias; however, when a sequence window contains an excess of Gs over Cs, the GC skew becomes a negative value based on the above formula, (C-G)/(C+G). Because the circular chromosome of eubacteria is subdivided into two replichores of opposite polarity that correspond to the leading and lagging strands, respectively, the GC skew graph of this genome shifts its sign at the junction of the two replichores. These shift-points in the GC skew graph correspond to the replication origin and terminus in eubacteria, and it is the basis for skew-based prediction of these sites (see [60, 61] for lists of software for origin prediction in bacteria). The uses of two-dimensional [62, 63] and three-dimensional [64, 65] trajectories have also been explored in order to accommodate multiple sets of bases and to overcome the need for windowed calculations. Cumulative plots of nucleotide skews are also frequently used in the analysis of skew shift points, which become the maxima and minima in these graphs [66]. Cumulative skew diagrams also correspond to the projection of random walk graphs for a single pair of complementary bases.

Numerous methods have been proposed for the analysis of compositional asymmetries, including multivariate statistics [31, 67] and signal processing methods based on Fourier [68] and wavelet transformations [50, 69] and wavelet-based multifractal analysis [70] (see [61] for comprehensive review). The key to these analyses is the quantitative measurement of the strength of strand bias. The prediction of replication origins by the analysis of compositional asymmetry is used in almost all sequencing projects of circular, bacterial genomes in order to name the first base of the sequences submitted to a public database based on the position of the replication origin. However, skew-based prediction is only applicable when a GC skew is clearly visible in the genome, which is not necessarily true in all prokaryotes [33, 71]. Quantitative indices of the degree of replication-related strand bias within these genomes become useful in such cases to provide the prerequisite validity of the application of skew-based analyses. Moreover, such measures of strand bias are indispensable for comparative studies of strand-specific mutational and selectional biases exerted by various mechanisms and for evolutionary studies of genetic changes in multiple genomes. To this end, we review several measures of replication-related strand bias and their applications in studying compositional asymmetry in genomic sequences.

MEASURES OF STRAND BIAS

ΔGC Skew and B_I Index

The GC skew is simply the measure of the bias towards a base within a pair of complementary bases; therefore, it consists of the sum of all mutational effects and does not delimit those of replication from transcription. Therefore, the simplest approach to extract only the replication-related strand bias is to take the difference of the GC skew values in the leading and lagging strands, so that other biases that are independent from the replication bias will cancel out. This ΔGC skew was first proposed in one of the earliest studies that used GC skew and is defined as follows [22]:

$$\Delta GCskew = GCskew_{leading} - GCskew_{lagging} \quad (3)$$

ΔGC skew can be used with other nucleotide pairs, including AT, keto, and amino bases, to calculate the skew. However, the above formula cannot exclude the effects of biased gene distribution in the leading and lagging strands and may contain strand bias effects from coding requirement or transcription. Followed by the work by Mrazek and Karlin that differentiated the effects of replication and transcription [26, 27], Rocha and coworkers therefore proposed the normalized ΔGC skew of genes for this purpose [72]:

$$\Delta GCskew_{gene} = \frac{\sum_{i \in genes_{leading}} GCskew_i}{N_{leading}} - \frac{\sum_{i \in genes_{lagging}} GCskew_i}{N_{lagging}} \quad (4)$$

They later further simplified the formula by considering only the third codon position of four-fold degenerate codons (*q*) [73] such that

$$\Delta GCskew_q = GCskew_{q,leading} - GCskew_{q,lagging} \quad (5)$$

While the ΔGC skew is stronger than the ΔAT skew in almost all bacteria, ΔAT skews are nonetheless observed in a number of species, and is sometimes more represented than ΔGC skew especially in A+T rich genomes [72]. Therefore, an index that takes account of both the GC and AT pairs is proposed by Lobry and Sueoka, as follows [74]

$$B_I = \sqrt{\left(x_{leading} - x_{lagging}\right)^2 + \left(y_{leading} - y_{lagging}\right)^2}$$
$$x = \frac{G}{G+C} \quad (6)$$
$$y = \frac{T}{T+A}$$

where *A, T, G,* and *C* are the frequencies of the corresponding nucleotides in the third codon positions. Consequently, the overall contribution of transcription- and translation-related bias can be formulated by the averages of *x* and *y* (x_c and y_c) of the leading and lagging strands:

$$B_{II} = \sqrt{\left(x_c - 0.5\right)^2 + \left(y_c - 0.5\right)^2} \quad (7)$$

As with the calculation of the ΔGC skew, Rocha and coworkers later modified the B_I index to only consider four-fold degenerate codons [73]. The maximal value of GC skew is 1 (when only C is found in a sequence), and therefore, the maximum ΔGC skew is 2 and that of B_I is $\sqrt{2}$. When there is no strand bias, both indices become 0.

The Predictability Score of Linear Discriminant Analysis of Gene Orientation

Nucleotide compositional skews often coincide with gene orientation skews [25, 67]. Genes are almost always more likely to be found on the leading strand, and there are presumably mutational or selectional pressures that affect the choice of nucleotides, codons, and amino acids. Linear discriminant analysis (LDA) is one of the simplest multivariate statistics approaches to compare these effects, particularly as there are only two states to be considered with regard to gene orientation, that is, whether a gene is found on the leading or the lagging strand. Here, the linear discriminant function is first obtained from the explanatory variables of interest [31] with

$$F(x) = a_0 + \sum_{i=1}^{n} \alpha_i x_i \quad (8)$$

where α_i is the calculated contribution coefficients for *i*th variable having composition x_i such that $F(x) > 0$ describes one group and another group is described by $F(x) < 0$. The number of variables, *n,* is therefore 4 for nucleotides, 12 for nucleotides in each codon position, 61 for codons, and 20 for amino acids. The predictabilities of the obtained discriminant functions can then be compared to assess the relative effects of compositional biases. Moreover, by calculating the predictability score along the genome one gene at a time and by dividing genes into two groups (left or right) relative to the observation position, the replication origin or terminus can be predicted to occur at the position where the predictability is highest. Nucleotides and codons generally have high prediction accuracy; however, the amino acid composition also shows a weaker but still clear predictability [31]. When there is no compositional bias, the prediction accuracy of LDA becomes 50%, implying a completely random distribution of the two classes (leading or lagging). The presence of stronger bias brings this number up to 100% (completely biased).

The Signal-to-Noise Ratio of Oligonucleotide Bias

The cause for the skew of oligonucleotides is multifactorial, including the Markovian probability based on the nucleotide composition [34, 36], the specific selection of biological signal motifs, such as Chi and KOPS [21, 67], and the codon usage, especially in light of biased gene distribution [73]. Nevertheless, strand bias generally coexists with nucleotide composition and the distribution of complementary oligonucleotides [33]. Worning and coworkers therefore presented the use of the oligonucleotide strand bias for the prediction of the replication origin by calculating the weighted double Kullback-Leibler distance, D_i, of each oligonucleotide, *I,* whose occurrence count is N_i as

$$D_i = (N_{i,leading} - N_{i,lagging}) \log_2 \left(\frac{N_{i,leading} + r}{N_{i,lagging} + r} \right) \quad (9)$$

where *r* (=5) is a control number to accommodate the low frequency of oligonucleotides. Here, $N_{i,leading}$ and $N_{i,lagging}$ are complementary oligonucleotides in a single-stranded sequence, and therefore, this calculation cannot be applied to palindromic (self-complementary) sequences such as 5'–GATC–3', which are identical on both strands. This D_i is calculated for all of the applicable oligonucleotides up to octamers, and the sum of all D_i at a position, *p,* is considered the overall bias:

$$S_p = \sum_{i \in oligo} D_i \quad (10)$$

S_p is calculated by moving the position, *p,* along the genome, and the regions upstream and downstream of *p* are considered as hypothetical origins. S_p is repeatedly calculated for window sizes of 50%, 55%, 60%, 65%, and 70% of the genome, and the median S_p among these windows is chosen, and is normalized to have the same scale as S_p calculated using the window size of 60% of the genome. Similar to the prediction accuracy of LDA, the S_p should be maximal at the real replication origin. Worning and coworkers took the signal-to-noise ratio of S_p further as the unique measure of strand bias for a given genome by using the maximal S_{max} and minimal S_{min}:

$$S/N = \frac{S_{\max}}{S_{\min}} \quad (11)$$

Unlike other measures of strand bias, the S/N of oligomer skew does not have a theoretical maximum value for completely biased sequences (such as 2 for ΔGC skew, 1 for

B_I, 100% for prediction accuracy of LDA, or 1 for GC skew index). When there is no strand bias, the S/N is at its minimum and is equal to 1.

The GC Skew Index

The calculation of the ΔGC skew or B_I requires the knowledge of the positions of origin and terminus that are typically predicted computationally using the skew shift points. Therefore, the application of these indices for skew analyses could become circular. Moreover, the skew shift points in the GC skew, or the maxima and minima in the cumulative GC skew graphs may not represent the true origin and terminus of replication in weakly biased sequences where genomic inversions and the horizontal transfer of biased sequences can introduce pseudo-shift points. In eubacteria, the replication-related strand asymmetry results in two replichores of nearly equal lengths (note that replichores are not exactly symmetrical in many bacteria, especially in the phylum Firmicutes) but with opposite polarity. GC skew graphs in these species therefore resemble the graph of a discrete sine curve, a graph composed of Y = -1 for $t_0 \sim (t_1 - t_0)/2$ and Y = 1 for $(t_1 - t_0)/2 \sim t_1$, and this "shape" of the GC skew graph can be assessed by observing the strength of the 1 Hz signal of its Fourier transformation [68]. Fourier transformation mathematically decomposes a given signal into a set of constituent frequencies, and the most simple frequency component of 1 Hz corresponds to a sine curve spanning all across the given signal duration. Therefore, pattern recognition of the shape of shifting GC skew graph can be mathematically considered as finding the 1 Hz spectral component of a given signal. The GC Skew Index (GCSI) thus combines the strand compositional difference that is calculated like the ΔGC skew ($dist_{norm}$) with the conformity of the GC skew graph to a discrete sine wave (*SA*) using a Fast Fourier Transformation (FFT) by taking the geometric mean of the two values [75, 76] such that

$$GCSI = \sqrt{k_1 SA \times k_2 dist_{norm}} \quad (12)$$

where k_1 and k_2 are normalization constants that have been empirically obtained to be 1/6000 and 1/600, respectively, in order to set the range of GCSI to 0 (no bias) or to approximately 1 (high bias). The *SA* is the spectral amplitude of the 1 Hz signal obtained by FFT, $F(k)$, of a signal of length, N, $f(n)$, where $n = 0, 1, \ldots, N-1$, at the frequency, k, is calculated as follows

$$F(k) = \sum_{n=0}^{N-1} f(n) e^{-i 2\pi k n / N} \quad (13)$$

where $i = \sqrt{-1}$. The power spectrum, $PS(k)$ of $F(k)$, is then given by

$$PS(k) = |F(k)|^2, k = 0,1,2,...,N-1 \quad (14)$$

at each frequency, k. The 1 Hz signal is therefore $PS(1)$. The *SA* is calculated by normalizing this signal strength as

$$SA = k_4 (k_3 PS(1))^{\alpha} \quad (15)$$

where k_3=600,000, k_4=40, and α=0.4, as calculated by regression analysis. The $dist_{norm}$ is calculated from the ΔGC skew with a windowed calculation using all bases (i.e., not limited to the third codon position) and using the regions between the maxima and the minima of the cumulative GC skew graphs, as follows

$$dist_{norm} = \Delta GCskew_{all} \times 4096 / W \quad (16)$$

where W is the number of windows used in the analysis for normalization.

Although windowed calculation of GC skew is required to produce initial signal for FFT, the GCSI does not depend on the choice of window size, as long as the number of windows is a power of 2. A window number of 2048 is recommended for eubacteria because each window contains regions longer than approximately 1 Kbp in these genomes (approximately 2 Mbp ~ 4 Mbp in size), which eliminates the local strand bias effects exerted by coding sequences that can be an average of 1 Kbp long. In small genomes, such as plasmids, a window number greater than 32 is recommended. GCSI also provides *P* values for the significance of its value based on *z*-testing with randomized iterations. A GCSI > 0.05 usually implies the existence of a strand bias in genomes with bidirectional replication machinery.

IMPLICATIONS OF THE DIFFERENT MEASURES OF STRAND BIAS

The five measures of replication-related strand asymmetry described above are summarized in Table **1**, and these indices can be categorized into two groups based on the requirement for knowledge of the coding regions and the positions of the replication origin and terminus. ΔGC skew, B_I, and LDA prediction accuracy require such information, and these indices are predominantly used in detailed studies of the mutation pressures that determine the mechanisms that shape the replication-related strand bias [72-74, 77-79]. The mechanism for replication-associated strand asymmetry has frequently been attributed to the cytosine deamination theory [21, 80, 81], which is based on evidence that single-stranded DNA is significantly more vulnerable to the spontaneous hydrolysis of cytosine to uracil, resulting in a C->T transition [82-84]. Due to the discontinuous strand synthesis of Okazaki fragments in the lagging strand, the leading strand is more prone to cytosine deamination, and this type of mutation had been generally accepted as one of the theories for strand-specific mutation. However, the aforementioned series of studies analyzed the mutation patterns using a large number of closely related chromosomes with an extremely careful selection of orthologous genes. By eliminating the effects of selection in combination with the strand bias measures, it is now clear that a C->T transition by cytosine deamination is not the only mutational pressure, and that different clades of bacteria are affected by different mutational biases [73] (see [23, 85, 86] for possible biological mechanisms causing these mutations). To summarize, these indices are designed for the careful analysis of mutational forces that shape the replication-related strand asymmetry and are needed to rule out the effects of selection and other mutational pressures; therefore, each of these indices represents a specific mutational effect. ΔGC skew and B_I both eliminate the effect of transcription-related mutation, and ΔGC skew only observes the G/C

Table 1. Summary of Strand Bias Measures

Index	Value range	Observing bias	Computation cost	*p-value*	Gene annotation	Replication origin and terminus	Circular genome
ΔGC skew	0 to 2	GC skew (GC3 only)	very low		required	required	
B_I	0 to $\sqrt{2}$	GC and AT skews (GC3 only)	very low		required	required	
LDA prediction accuracy	0.5 to 1	gene skew	high	yes	required		required
S/N of oligomer skew	1 to ∞	oligomer skew	very high				required
GCSI	0 to 1	GC skew (all regions)	low	yes			required

*GC3 denotes third codon positions.

mutation bias whereas B_I observes the combined effects on A/T and G/C. LDA is intended for the study of replication-related mutation bias on the nucleotide or codon composition of coding regions, and S/N of oligomer skew observes that for short oligonucleotides. GCSI does not separate these mutational effects, and quantifies the overall sum of replication-related mutations affecting on the genome.

The other two indices, S/N of oligomer skew and GCSI, do not require *a priori* knowledge of the coding regions or the positions of the replication origin and terminus, and they tend to indicate the overall contribution of different mutational forces, including the effects of nucleotide composition bias and gene orientation bias. Therefore, these measures are useful for assessing the significance of skew-based predictions of replication origins and termini, and to quantify the strength of strand asymmetry for comparative studies among hundreds of genomes. For example, Rocha and coworkers and later Worning and cowerkers showed that the degree of strand bias shows a characteristic distribution in different phyla of eubacteria (reproduced with modifications in Fig. (**1**)) and that the direction of the AT skew is correlated with the presence of the *polC* subunit of DNA polymerase, which results in an asymmetric proofreading system in the leading and lagging strands [27, 87]. The type of DNA polymerase holoenzyme subunit structure is also known to affect the gene strand bias in these genomes [27]. A comparative study using GCSI characterized the different types of replication, such as bidirectional replication in eubacteria, replication from multiple origins in archaea, and the theta and rolling circle methods of replication in bacterial plasmids [75]. Because the GCSI checks the conformity of the GC skew graph to the discrete sine curve that is a characteristic of bidirectional replication mechanisms, the GCSI is low when there are multiple origins of replication, as in archaeal species. Likewise, the rolling circle method of replication results in the entire strand of DNA being a leading strand, which results in a continuously rising cumulative GC skew graph, and subsequently, a high $dist_{norm}$ and a low *SA*. As a result, the GCSI becomes relatively high, but its *P*-value significance would be low. We have also showed a high correlation between the GCSI of plasmids and their corresponding host chromosomes [75], suggesting that amelioration within cells can be caused by replication-related mutation, in addition to reported amelioration for genomic signatures [88-91].

While the objectives of the indices of strand bias are different, as discussed so far, they are generally correlated to each other, as shown in Fig. (**2**) by clustering the indices according to their Pearson correlation using the calculated strand bias values for 1083 bacterial genomes. With the exception of the S/N of oligomer skew with several of the LDA prediction accuracies, almost all of the strand bias indices were highly correlated to each other, with $r > 0.70$. For comparison, the inverse of the doubling time collected from the literature for 254 species [92] and several other features that are related to the bacterial growth rate, including the copy number of tRNAs and rRNA operons as well as the Sharp's S-index for the strength of selected codon usage bias [93], were also used in the cluster analysis. While the strand bias indices are not as highly correlated to the growth rate as the known features, they nevertheless show a weak correlation ($r > 0.25$). This is consistent with the observations that slow growers, such as Cyanobacteria and *Mycoplasma* species, are known to have a very limited GC skew [76, 87] and that the fast growers tend to have a higher gene orientation bias, such as in Firmicutes, presumably to avoid the head-on collision of DNA and RNA polymerases [94, 95], although other mechanisms have also been suggested [23]. Note however, that the possible existence of multiple origins is also suggested as the reason for unclear GC skew [96].

STRAND BIAS STUDIES IN EUKARYOTES

The initiation of replication is highly diverse in eukaryotes. *Saccharomyces cerevisiae* is known to have highly conserved autonomously replicating sequences (ARSs) of about 200 bp in length [97] and replication-related strand asymmetry is only observed in the subtelomeric regions in this species [47]. The replication origins span a broader region of the chromosomes in *Schizosaccharomyces*

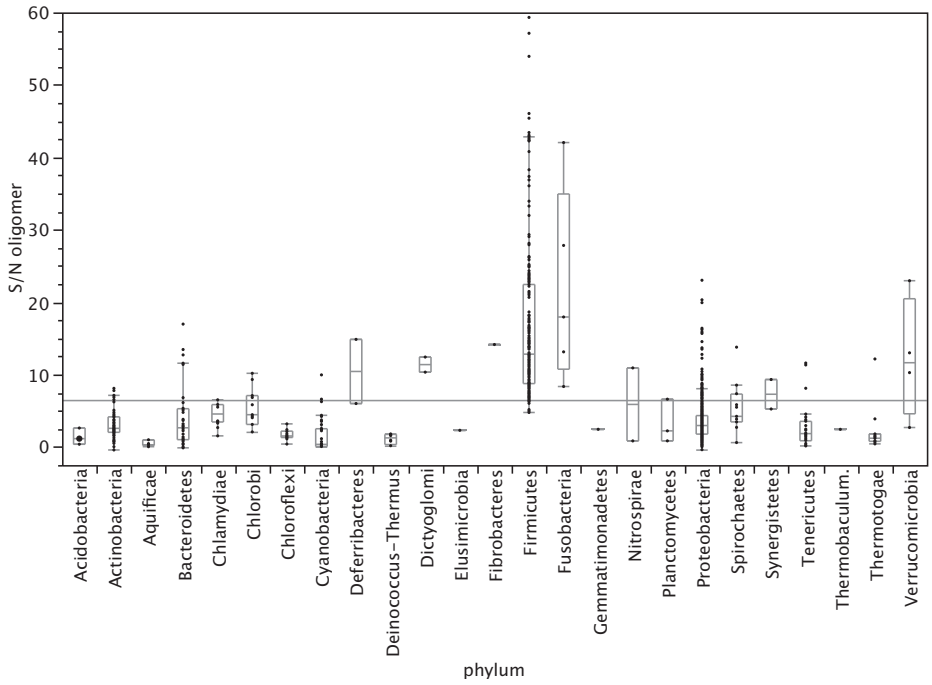


Fig. (1). Distribution of the S/N of oligomer skew in different phyla of eubacteria. The degree of replication-associated strand bias shows a characteristic distribution among the bacterial phyla, and Firmicutes is one of the most highly biased groups. The majority of the Firmicutes contain the *polC* subunit of the DNA polymerase, which results in a high gene strand bias. In comparison, proteobacteria belong to a group with a moderate degree of strand bias.

Fig. (2). Clustering of the measures of strand bias and growth rate by Pearson correlations in 1083 bacterial genomes (only 254 genomes were used for 1/generation time). All strand bias measures are highly correlated (most above $r > 0.70$) and are also weakly correlated with properties related to the growth rate ($r > 0.25$).

pombe [97]. In *Drosophila* and *Xenopus* embryo where the S-phase is extremely rapid [98-100], the replication origins are distributed randomly [101] and any regions of DNA seems to be able to function as a replicator. The initiation sites for replication in mammals are more strictly defined; however, their locations are not encoded in the nucleotide sequence as motifs, but rather, are epigenetically coded by the chromatin organization [102-104]. The availability of experimentally defined origins and more comprehensive origin mapping studies by chromatin immunoprecipitation (ChIP) analysis coupled with microarrays (ChIP on chip) or deep sequencing (ChIP-Seq) is opening the door to a number of computational studies regarding replication-related strand asymmetry in the human genome [105-109], and mechanisms leading to eukaryotic strand bias are getting reported [110, 111].

The study of strand asymmetry around human replication origins primarily uses alterations in the GC skew [19, 48-50, 112-115]:

$$S_{GC} = \frac{G-C}{G+C}$$
$$S_{AT} = \frac{T-A}{T+A} \quad (17)$$
$$S = S_{TA} + S_{GC}$$

When considering the regions upstream and downstream of replication origins, the degree of strand asymmetry around these sites can be defined as:

$$\Delta S = S_{upstream} - S_{downstream} \quad (18)$$

As with the GC skew analysis, the *S* values are plotted along the nucleotide sequence using a sliding window of 1 Kbp, and cumulative graphs are also used to identify the shift points. Unlike the GC skew in prokaryotes, where the nucleotide composition is uniform across a strand, the *S* in eukaryotes rapidly diminishes and approaches 0 as the distance from the origin of replication increases. This results in an "N-shaped" graph, unlike the GC skew graph of prokaryotes, which resembles a discrete sine curve. Based on this N-shaped *S* graph coupled with wavelet analyses, strand asymmetry is used to predict replication origins and to study the effects of gene orientation [50], transcription [48, 114], and chromatin organizations [49, 114].

The *S* and Δ*S* are basically equivalent to the ΔGC skew and B_I index. While the use of other indices of strand asymmetry in bacteria, such as the LDA prediction accuracy, the S/N of oligomer skew, and the GCSI, are limited due to their requirement for circular genomes, these indices should, in principle, be applicable to eukaryotic genomes and also to archaeal genomes with multiple origins [116, 117] by extracting a segment centered around the replication initiation sites, as in the calculation of Δ*S*. Modification would be necessary for the LDA prediction accuracy due to the scarce nature of coding regions in mammals, and for the GCSI due to the N-shape of the strand skew rather than a discrete sine curve. However, the availability of multiple indices that assess the strength-of-strand asymmetry by alternative means would provide a stronger basis for these analyses.

SOFTWARE FOR THE CALCULATION OF STRAND BIAS MEASURES

There are a wide variety of software and web tools for the graphical analysis of GC skew and other compositional asymmetries, such as the Integrated Microbial Genomes (IMG) system provided by the Joint Genome Institute [118], and a comprehensive review is also available elsewhere [61]. Therefore, in this review, we focus on the software for the calculation of strand bias measures described thusfar.

Analysis of strand asymmetry almost always requires the use of multiple indices or a combination of methods and careful selection or grouping of species or gene sets. Therefore, comprehensive analysis packages and environments are desirable for detailed studies. Two software packages are equipped with collections of tools for the study of compositional asymmetry – the SeqinR package in the R statistics language (http://www.r-project.org) [119] and the G-language Genome Analysis Environment (G-language GAE) in Perl [54, 120, 121]. Both packages support several formats of sequence flatfiles and can take advantage of the advanced statistical capabilities of programming environments. In this section, we illustrate the use of the G-language GAE to analyze strand bias because this software package contains almost all of the measures of strand bias that have been described thus far, and because the G-language GAE can be accessed *via* its web service interface [122]. The availability of the RESTful web service interface eliminates the requirement for software installation, setup, and maintenance, and users can access the service regardless of their platform.

The web services of the G-language GAE can be used by specifying a URL in a browser according to a set of rules. The most basic syntax is the following:

http://rest.g-language.org/[accession]/[program]/[option]=[value]/ (19)

where [accession] corresponds to the NCBI RefSeq accession number for the genome of interest, such as NC_000913 for *Escherichia coli* K12 MG1655, or an accession ID given by the G-language GAE after uploading a file at http://rest.g-language.org/upload/. The name of the program to use is [program], and multiple [option]=[value] pairs separated by a slash (/) can be appended for the configurable options. A list of programs, as well as the options that are related to the analysis of the replication-related strand bias, is shown in Table **2**, and a complete list of all programs in the G-language GAE and detailed documentation is available online at http://ws.g-language.org/gdoc/. For example, the calculation of the GCSI for the *E. coli* genome can be obtained by simply accessing the following URL

http://rest.g-language.org/NC_000913/gcsi/ (20)

which immediately returns a GCSI of 0.096 with *SA* and $dist_{norm}$. The options of programs can be configured to meet a variety of needs. Fig. (**3**) shows several examples of the GC skew analysis of *Escherichia coli*. Documentation for each of the programs, including the list of available options, can be viewed at

http://rest.g-language.org/help/[program]/ (21)

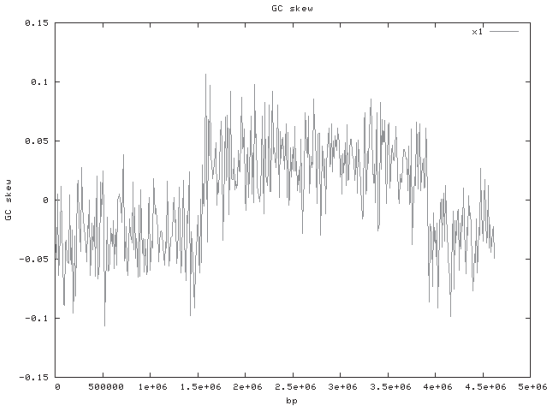
There are also several other programs that are suited to the study of strand asymmetry in addition to those listed in Table **2**, such as the extraction of leading and lagging strand sequences based on the position of the origin in dOriC database [123] and the *dif* positions [124], or for the search of DnaA box, Ter, and iteron sites. More detailed documentations and live examples are available at http://www.g-language.org/wiki/rest. The G-language REST web service provides an intuitive and rapid method to use a variety of tools to study strand bias without the need for installation and setup. However, in order to automate an analysis pipeline that encompasses a number of tools applied to hundreds of genomes and to take advantage of the maximal efficiency of local computers, installation of the G-language GAE software is recommended. The latest software package and detailed documentation and tutorials are available at the project's web site, http://www.g-language.org/. G-language GAE is free software licensed under the GNU General Public License version 2.

Table 2. Programs and Options for Strand Bias Analysis in the G-Language Genome Analysis Environment

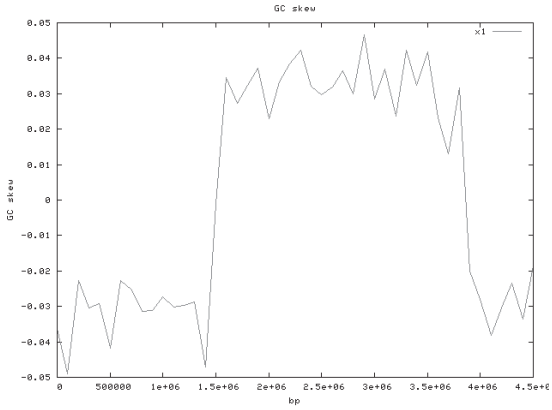
Name	Option	Description
B1		B_I index
B2		B_{II} index
delta_gcskew	method=degenerate (default)	ΔGC skew using four-fold degenerate GC3
	method=gc3	ΔGC skew using GC3
	method=all	ΔGC skew using all bases
	at=1	ΔAT skew
	purine=1	ΔPurine skew
	keto=1	ΔKeto skew
gcsi		GC skew index
	at=1	AT skew index
	purine=1	Puine skew index
	keto=1	Keto skew index
lda_bias	variable=codon (default)	LDA prediction accuracy using 61 codons
	variable=base	LDA prediction accuracy using 4 bases
	variable=codonbase	LDA prediction accuracy using 12 bases/codon positions
	variable=amino	LDA prediction accuracy using 20 amino acids
gcskew		GC skew graph
	cumulative=1	cumulative GC skew graph
	at=1	AT skew graph
	purine=1	Purine skew graph
	keto=1	Keto skew graph
gcwin		GC content graph
	at=1	AT content graph
	purine=1	Purine content graph
	keto=1	Keto content graph
geneskew		gene skew graph
	cumulative=1	cumulative gene skew graph
	gc3=1	GC/AT/Purine/Keto skew graph in GC3 (specified with "base" option)
genomicskew		GC skew graph of coding/non-coding/GC3 regions
	at=1	AT skew graph of coding/non-coding/GC3 regions
dnawalk		DNA walk graph
find_ori_ter		origin / terminus prediction using cumulative GC skew
	at=1	origin / terminus prediction using cumulative AT skew
	purine=1	origin / terminus prediction using cumulative Purine skew
	keto=1	origin / terminus prediction using cumulative Keto skew
	filter=95	origin / terminus prediction using low-pass filtering with FFT
rep_ori_ter	gcskew=1	origin / terminus prediction using cumulative GC skew
	oriloc=1	origin / terminus prediction using Oriloc algorithm
	dbonly=1	origin / terminus prediction using dOriC and dif prediction data

*GC3 denotes third codon positions.

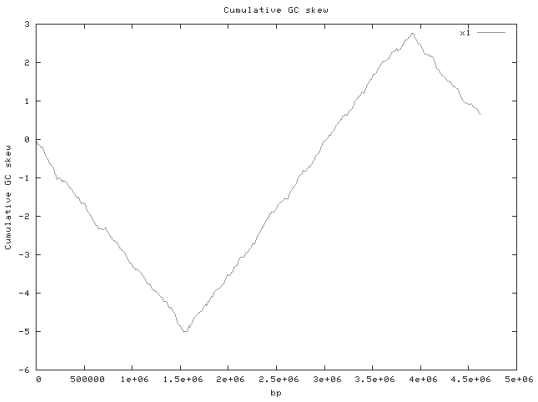
http://rest.g-language.org/NC_000913/gcskew



http://rest.g-language.org/NC_000913/gcskew/window=100000



http://rest.g-language.org/NC_000913/gcskew/cumulative=1



http://rest.g-language.org/NC_000913/gcskew/at=1/window=10000

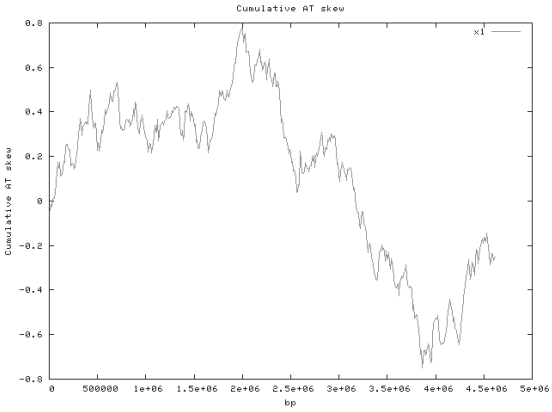


Fig. (3). Examples of GC skew graphs using the G-language GAE web service.

G-language REST web service can be easily used by typing a URL in a browser starting with http://rest.g-language.org/ followed by a set of commands. Here, the GC skew graphs of *Escherichia coli* (NC_000913) are visualized using different sets of options, such as normal or wider windows (window=100000), a cumulative graph (cumulative=1), and the AT skew at a window size of 10000 bp (at=1, window=10000). Neither registration nor setup is required to use these services, allowing the user to readily make use of these tools for their research.

CONCLUSIONS

Quantitative measures of replication-related strand bias are necessary to quantify the different mutational pressures, to study complex biological causes of strand asymmetry, and to compare the degree of strand bias in genomes among the diverse clades of bacteria. Such studies require an extremely careful selection of the set of target species, genes, and genetic features to rule out potential biases due to pseudo-correlations. Therefore, the use of suitable measures of strand bias and possibly combining multiple methods based on different algorithms can be a critical part of the research. The availability of intuitive, RESTful services of these indices as part of the G-language GAE allows for quick, heuristic checking. Replication-related strand bias in eukaryotes has yet to be explored in detail, and measures of strand bias that are suitable for the analysis of eukaryotic origins would be an interesting area of research.

ACKNOWLEDGEMENTS

This research was supported by funds from the Yamagata Prefectural Government and Tsuruoka City.

REFERENCES

[1] Nakabachi, A.; Yamashita, A.; Toh, H.; Ishikawa, H.; Dunbar, H. E.; Moran, N. A.; Hattori, M. The 160-kilobase genome of the bacterial endosymbiont Carsonella. *Science,* **2006,** *314*(5797), 267.

[2] Hildebrand, F.; Meyer, A.; Eyre-Walker, A. Evidence of selection upon genomic GC-content in bacteria. *PLoS Genet.,* **2010,** *6*(9), e1001107.

[3] Rocha, E. P.; Danchin, A. Base composition bias might result from competition for metabolic resources. *Trends Genet.,* **2002,** *18*(6), 291-294.

[4] Akashi, H.; Gojobori, T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. *Proc. Natl. Acad. Sci. U S A,* **2002,** *99*(6), 3695-3700.

[5] Mitchell, D. GC content and genome length in Chargaff compliant genomes. *Biochem. Biophys. Res. Commun.,* **2007,** 353(1), 207-210.

[6] McEwan, C. E.; Gatherer, D.; McEwan, N. R. Nitrogen-fixing aerobic bacteria have higher genomic GC content than non-fixing species within the same genus. *Hereditas,* **1998,** *128*(2), 173-178.

[7] Naya, H.; Romero, H.; Zavala, A.; Alvarez, B.; Musto, H. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. *J. Mol. Evol.,* **2002,** *55*(3), 260-264.

[8] Wang, H. C.; Susko, E.; Roger, A. J. On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors. *Biochem. Biophys. Res. Commun.,* **2006,** *342*(3), 681-684.

[9] Forsdyke, D. R.; Mortimer, J. R. Chargaff's legacy. *Gene,* **2000,** *261*(1), 127-137.

[10] Chargaff, E. Structure and function of nucleic acids as cell constituents. *Fed. Proc.,* **1951,** *10*(3), 654-659.

[11] Watson, J. D.; Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature,* **1953,** *171*(4356), 737-738.

[12] Rudner, R.; Karkas, J. D.; Chargaff, E. Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. *Proc. Natl. Acad. Sci. U S A,* **1968,** *60*(3), 921-922.

[13] Mitchell, D.; Bridge, R. A test of Chargaff's second rule. *Biochem. Biophys. Res. Commun.,* **2006,** *340*(1), 90-94.

[14] Sueoka, N. Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J. Mol. Evol.,* **1995,** *40*(3), 318-325.

[15] Nikolaou, C.; Almirantis, Y. Deviations from Chargaff's second parity rule in organellar DNA Insights into the evolution of organellar genomes. *Gene,* **2006,** *381*, 34-41.

[16] Szybalski, W.; Kubinski, H.; Sheldrick, P. Pyrimidine clusters on the transcribing strand of DNA and their possible role in the initiation of RNA synthesis. *Cold Spring Harb Symp. Quant. Biol.,* **1966,** *31*, 123-127.

[17] Bell, S. J.; Forsdyke, D. R. Deviations from Chargaff's second parity rule correlate with direction of transcription. *J. Theor. Biol.,* **1999,** *197*(1), 63-76.

[18] Lao, P. J.; Forsdyke, D. R. Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. *Genome Res.,* **2000,** *10*(2), 228-236.

[19] Touchon, M.; Arneodo, A.; d'Aubenton-Carafa, Y.; Thermes, C. Transcription-coupled and splicing-coupled strand asymmetries in eukaryotic genomes. *Nucleic Acids Res.,* **2004,** *32*(17), 4969-4978.

[20] Rocha, E. P. Order and disorder in bacterial genomes. *Curr. Opin. Microbiol.,* **2004,** *7*(5), 519-527.

[21] Lobry, J. R.; Louarn, J. M. Polarisation of prokaryotic chromosomes. *Curr. Opin. Microbiol.,* **2003,** *6*(2), 101-108.

[22] Lobry, J. R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.,* **1996,** *13*(5), 660-665.

[23] Rocha, E. P. The replication-related organization of bacterial genomes. *Microbiology* **2004,** *150*(Pt 6), 1609-1627.

[24] Rocha, E. P. The organization of the bacterial genome. *Annu. Rev. Genet.,* **2008,** *42*, 211-233.

[25] McLean, M. J.; Wolfe, K. H.; Devine, K. M. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *J. Mol. Evol.,* **1998,** *47*(6), 691-696.

[26] Mrazek, J.; Karlin, S. Strand compositional asymmetry in bacterial and large viral genomes. *Proc. Natl. Acad. Sci. USA,* **1998,** *95*(7), 3720-3725.

[27] Rocha, E. Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? *Trends Microbiol.,* **2002,** *10*(9), 393-395.

[28] Guo, F. B.; Ning, L. W. Strand-specific Composition Bias in Bacterial Genomes. In *DNA Replication-Current Advances*, Seligmann, H., Ed. InTech: 2011.

[29] Guo, F. B.; Yu, X. J. Separate base usages of genes located on the leading and lagging strands in Chlamydia muridarum revealed by the Z curve method. *BMC Genomics,* **2007,** *8*, 366.

[30] Guo, F. B.; Yuan, J. B. Codon usages of genes on chromosome, and surprisingly, genes in plasmid are primarily affected by strand-specific mutational biases in Lawsonia intracellularis. *DNA Res.,* **2009,** *16*(2), 91-104.

[31] Rocha, E. P.; Danchin, A.; Viari, A. Universal replication biases in bacteria. *Mol. Microbiol.,* **1999,** *32*(1), 11-16.

[32] Wei, W.; Guo, F. B. Strong Strand Composition Bias in the Genome of Ehrlichia canis Revealed by Multiple Methods. *Open Microbiol. J.,* **2010,** *4*, 98-102.

[33] Salzberg, S. L.; Salzberg, A. J.; Kerlavage, A. R.; Tomb, J. F. Skewed oligomers and origins of replication. *Gene,* **1998,** *217*(1-2), 57-67.

[34] Arakawa, K.; Uno, R.; Nakayama, Y.; Tomita, M. Validating the significance of genomic properties of Chi sites from the distribution of all octamers in Escherichia coli. *Gene,* **2007,** *392*(1-2), 239-246.

[35] Kowalczykowski, S. C.; Dixon, D. A.; Eggleston, A. K.; Lauder, S. D.; Rehrauer, W. M. Biochemistry of homologous recombination in Escherichia coli. *Microbiol. Rev.,* **1994,** *58*(3), 401-465.

[36] Uno, R.; Nakayama, Y.; Arakawa, K.; Tomita, M. The orientation bias of Chi sequences is a general tendency of G-rich oligomers. *Gene,* **2000,** *259*(1-2), 207-215.

[37] Bigot, S.; Saleh, O. A.; Cornet, F.; Allemand, J. F.; Barre, F. X. Oriented loading of FtsK on KOPS. *Nat. Struct. Mol. Biol.,* **2006,** *13*(11), 1026-1028.

[38] Bigot, S.; Saleh, O. A.; Lesterlin, C.; Pages, C.; El Karoui, M.; Dennis, C.; Grigoriev, M.; Allemand, J. F.; Barre, F. X.; Cornet, F. KOPS: DNA motifs that control E. coli chromosome segregation by orienting the FtsK translocase. *EMBO J.,* **2005,** 24, (21), 3770-80.

[39] Saleh, O. A.; Perals, C.; Barre, F. X.; Allemand, J. F. Fast, DNA-sequence independent translocation by FtsK in a single-molecule experiment. *EMBO J.,* **2004,** *23*(12), 2430-2439.

[40] Chedin, F.; Kowalczykowski, S. C. A novel family of regulated helicases/nucleases from Gram-positive bacteria: insights into the initiation of DNA recombination. *Mol. Microbiol.,* **2002,** *43*(4), 823-834.

[41] Hendrickson, H.; Lawrence, J. G. Selection for chromosome architecture in bacteria. *J. Mol. Evol.,* **2006,** *62*(5), 615-629.

[42] Rocha, E. P.; Danchin, A. Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res.,* **2003,** *31*(22), 6570-6577.

[43] Omont, N.; Kepes, F. Transcription/replication collisions cause bacterial transcription units to be longer on the leading strand of replication. *Bioinformatics,* **2004,** *20*(16), 2719-2725.

[44] Price, M. N.; Alm, E. J.; Arkin, A. P. Interruptions in gene expression drive highly expressed operons to the leading strand of DNA replication. *Nucleic Acids Res.,* **2005,** *33*(10), 3224-3234.

[45] Arakawa, K.; Tomita, M. Selection effects on the positioning of genes and gene structures from the interplay of replication and transcription in bacterial genomes. *Evol. Bioinform. Online,* **2007,** *3*, 279-286.

[46] Valens, M.; Penaud, S.; Rossignol, M.; Cornet, F.; Boccard, F. Macrodomain organization of the Escherichia coli chromosome. *EMBO J.* **2004,** *23*(21), 4330-4341.

[47] Gierlik, A.; Kowalczuk, M.; Mackiewicz, P.; Dudek, M. R.; Cebrat, S. Is there replication-associated mutational pressure in the Saccharomyces cerevisiae genome? *J. Theor. Biol.,* **2000,** *202*(4), 305-314.

[48] Huvet, M.; Nicolay, S.; Touchon, M.; Audit, B.; d'Aubenton-Carafa, Y.; Arneodo, A.; Thermes, C. Human gene organization driven by the coordination of replication and transcription. *Genome Res.,* **2007,** *17*(9), 1278-1285.

[49] Necsulea, A.; Guillet, C.; Cadoret, J. C.; Prioleau, M. N.; Duret, L. The relationship between DNA replication and human genome organization. *Mol. Biol. Evol.,* **2009,** *26*(4), 729-741.

[50] Touchon, M.; Nicolay, S.; Audit, B.; Brodie of Brodie, E. B.; d'Aubenton-Carafa, Y.; Arneodo, A.; Thermes, C. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. *Proc. Natl. Acad. Sci. USA,* **2005,** *102*(28), 9836-9841.

[51] Calistri, E.; Livi, R.; Buiatti, M. Evolutionary trends of GC/AT distribution patterns in promoters. *Mol. Phylogenet. Evol.,* **2011,** *60*(2), 228-235.

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

[53] Tatarinova, T.; Brover, V.; Troukhan, M.; Alexandrov, N. Skew in CG content near the transcription start site in Arabidopsis thaliana. *Bioinformatics,* **2003,** *19* Suppl 1, i313-4.

[54] Arakawa, K.; Suzuki, H.; Tomita, M. Computational Genome Analysis Using The G-language System. *Genes, Genomes and Genomics,* **2008,** *2*(1), 1-13.

[55] Frank, A. C.; Lobry, J. R. Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics,* **2000,** *16*(6), 560-561.

[56] Mackiewicz, P.; Zakrzewska-Czerwinska, J.; Zawilak, A.; Dudek, M. R.; Cebrat, S. Where does bacterial replication start? Rules for predicting the oriC region. *Nucleic Acids Res.,* **2004,** *32*(13), 3781-3791.

[57] Perna, N. T.; Kocher, T. D. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. *J. Mol. Evol.,* **1995,** *41*(3), 353-358.

[58] Freeman, J. M.; Plasterer, T. N.; Smith, T. F.; Mohr, S. C. Patterns of Genome Organization in Bacteria. *Science* **1998,** *279*(5358), 1827.

[59] Bell, S. J.; Forsdyke, D. R. Accounting units in DNA. *J. Theor. Biol.,* **1999,** *197*(1), 51-61.

[60] Palleja, A.; Guzman, E.; Garcia-Vallve, S.; Romeu, A. In silico prediction of the origin of replication among bacteria: a case study of Bacteroides thetaiotaomicron. *OMICS,* **2008,** *12*(3), 201-210.

[61] Touchon, M.; Rocha, E. P. From GC skews to wavelets: a gentle guide to the analysis of compositional asymmetries in genomic data. *Biochimie,* **2008,** *90*(4), 648-659.

[62] Lobry, J. R. A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie,* **1996,** *78*(5), 323-326.

[63] Peng, C. K.; Buldyrev, S. V.; Goldberger, A. L.; Havlin, S.; Sciortino, F.; Simons, M.; Stanley, H. E. Long-range correlations in nucleotide sequences. *Nature,* **1992,** *356*(6365), 168-170.

[64] Zhang, C. T.; Zhang, R.; Ou, H. Y. The Z curve database: a graphic representation of genome sequences. *Bioinformatics,* **2003,** *19*(5), 593-599.

[65] Zhang, R.; Zhang, C. T. Z curves, an intutive tool for visualizing and analyzing the DNA sequences. *J. Biomol. Struct. Dyn.,* **1994,** *11*(4), 767-782.

[66] Grigoriev, A. Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.,* **1998,** *26*(10), 2286-2290.

[67] Tillier, E. R.; Collins, R. A. The contributions of replication orientation, gene direction, and signal sequences to base-composition asymmetries in bacterial genomes. *J. Mol. Evol.,* **2000,** *50*(3), 249-257.

[68] Arakawa, K.; Saito, R.; Tomita, M. Noise-reduction filtering for accurate detection of replication termini in bacterial genomes. *FEBS Lett.,* **2007,** *581*(2), 253-258.

[69] Song, J.; Ware, A.; Liu, S. L. Wavelet to predict bacterial ori and ter: a tendency towards a physical balance. *BMC Genomics,* **2003,** *4*(1), 17.

[70] Nicolay, S.; Brodie Of Brodie, E. B.; Touchon, M.; Audit, B.; d'Aubenton-Carafa, Y.; Thermes, C.; Arneodo, A. Bifractality of human DNA strand-asymmetry profiles results from transcription. *Phys. Rev. E Stat. Nonlin Soft Matter Phys.,* **2007,** *75*(3 Pt 1), 032902.

[71] Kowalczuk, M.; Mackiewicz, P.; Mackiewicz, D.; Nowicka, A.; Dudkiewicz, M.; Dudek, M. R.; Cebrat, S. DNA asymmetry and the replicational mutational pressure. *J. Appl. Genet.,* **2001,** *42*(4), 553-577.

[72] Rocha, E. P.; Danchin, A. Ongoing evolution of strand composition in bacterial genomes. *Mol. Biol. Evol.,* **2001,** *18*(9), 1789-1799.

[73] Rocha, E. P.; Touchon, M.; Feil, E. J. Similar compositional biases are caused by very different mutational effects. *Genome Res.,* **2006,** *16*(12), 1537-1547.

[74] Lobry, J. R.; Sueoka, N. Asymmetric directional mutation pressures in bacteria. *Genome Biol.,* **2002,** *3*(10), RESEARCH0058.

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

[76] Arakawa, K.; Tomita, M. The GC Skew Index: A Measure of Genomic Compositional Asymmetry and the Degree of Replicational Selection. *Evol. Bioinform. Online,* **2007,** *3*, 159-168.

[77] Chen, C.; Chen, C. W. Quantitative analysis of mutation and selection pressures on base composition skews in bacterial chromosomes. *BMC Genomics,* **2007,** *8*, 286.

[78] Marin, A.; Xia, X. GC skew in protein-coding genes between the leading and lagging strands in bacterial genomes: new substitution models incorporating strand bias. *J. Theor. Biol.,* **2008,** *253*(3), 508-513.

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

[80] Frank, A. C.; Lobry, J. R. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene,* **1999,** *238*(1), 65-77.

[81] Reyes, A.; Gissi, C.; Pesole, G.; Saccone, C. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol. Biol. Evol.,* **1998,** *15*(8), 957-966.

[82] Coulondre, C.; Miller, J. H.; Farabaugh, P. J.; Gilbert, W. Molecular basis of base substitution hotspots in Escherichia coli. *Nature,* **1978,** *274*(5673), 775-780.

[83] Frederico, L. A.; Kunkel, T. A.; Shaw, B. R. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry,* **1990,** *29*(10), 2532-2537.

[84] Lindahl, T.; Nyberg, B. Heat-induced deamination of cytosine residues in deoxyribonucleic acid. *Biochemistry,* **1974,** *13*(16), 3405-3410.

[85] Francino, M. P.; Ochman, H. Strand asymmetries in DNA evolution. *Trends Genet.,* **1997,** *13*(6), 240-245.

[86] Karlin, S. Bacterial DNA strand compositional asymmetry. *Trends Microbiol.,* **1999,** *7*(8), 305-308.

[87] Worning, P.; Jensen, L. J.; Hallin, P. F.; Staerfeldt, H. H.; Ussery, D. W. Origin of replication in circular prokaryotic chromosomes. *Environ. Microbiol.,* **2006,** *8*(2), 353-361.

[88] Campbell, A.; Mrazek, J.; Karlin, S. Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. *Proc. Natl. Acad. Sci. U S A,* **1999,** *96*(16), 9184-9189.

[89] Lawrence, J. G.; Ochman, H. Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.,* **1997,** *44*(4), 383-397.

[90] Ochman, H.; Lawrence, J. G.; Groisman, E. A. Lateral gene transfer and the nature of bacterial innovation. *Nature,* **2000,** *405*(6784), 299-304.

[91] Suzuki, H.; Sota, M.; Brown, C. J.; Top, E. M. Using Mahalanobis distance to compare genomic signatures between bacterial plasmids and chromosomes. *Nucleic Acids Res.,* **2008,** *36*(22), e147.

[92] Vieira-Silva, S.; Rocha, E. P. The systemic imprint of growth and its uses in ecological (meta)genomics. *PLoS Genet.,* **2010,** 6(1), e1000808.

[93] Sharp, P. M.; Bailes, E.; Grocock, R. J.; Peden, J. F.; Sockett, R. E. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.,* **2005,** *33*(4), 1141-1153.

[94] Deshpande, A. M.; Newlon, C. S. DNA replication fork pause sites dependent on transcription. *Science,* **1996,** *272*(5264), 1030-1033.

[95] French, S. Consequences of replication fork movement through transcription units *in vivo*. *Science,* **1992,** *258*(5086), 1362-1365.

[96] Nikolaou, C.; Almirantis, Y. 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 Res.,* **2005,** *33*(21), 6816-6822.

[97] Bell, S. P.; Dutta, A. DNA replication in eukaryotic cells. *Annu. Rev. Biochem.,* **2002,** *71*, 333-374.

[98] Coverley, D.; Laskey, R. A. Regulation of eukaryotic DNA replication. *Annu. Rev. Biochem.,* **1994,** *63*, 745-776.

[99] Hyrien, O.; Mechali, M. Chromosomal replication initiates and terminates at random sequences but at regular intervals in the ribosomal DNA of Xenopus early embryos. *EMBO J.,* **1993,** *12*(12), 4511-4520.

[100] Sasaki, T.; Sawado, T.; Yamaguchi, M.; Shinomiya, T. Specification of regions of DNA replication initiation during embryogenesis in the 65-kilobase DNApolalpha-dE2F locus of Drosophila melanogaster. *Mol. Cell Biol.,* **1999,** *19*(1), 547-555.

[101] Gilbert, D. M. Making sense of eukaryotic DNA replication origins. *Science,* **2001,** *294*(5540), 96-100.

[102] Demeret, C.; Vassetzky, Y.; Mechali, M. Chromatin remodelling and DNA replication: from nucleosomes to loop domains. *Oncogene,* **2001,** *20*(24), 3086-3093.

[103] McNairn, A. J.; Gilbert, D. M. Epigenomic replication: linking epigenetics to DNA replication. *Bioessays,* **2003,** *25*(7), 647-656.

[104] Mechali, M. DNA replication origins: from sequence specificity to epigenetics. *Nat. Rev. Genet.,* **2001,** *2*(8), 640-645.

[105] The ENCODE consortium, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature,* **2007,** *447*(7146), 799-816.

[106] Cadoret, J. C.; Meisch, F.; Hassan-Zadeh, V.; Luyten, I.; Guillet, C.; Duret, L.; Quesneville, H.; Prioleau, M. N. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc. Natl. Acad. Sci. USA,* **2008,** *105*(41), 15837-15842.

[107] Cayrou, C.; Coulombe, P.; Vigneron, A.; Stanojcic, S.; Ganier, O.; Peiffer, I.; Rivals, E.; Puy, A.; Laurent-Chabalier, S.; Desprat, R.; Mechali, M., Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites

defined by conserved features. *Genome Res.,* **2011***, 21*(9), 1438-1449.

[108] Hansen, R. S.; Thomas, S.; Sandstrom, R.; Canfield, T. K.; Thurman, R. E.; Weaver, M.; Dorschner, M. O.; Gartler, S. M.; Stamatoyannopoulos, J. A. Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. *Proc. Natl. Acad. Sci. U S A,* **2010***, 107*(1), 139-144.

[109] Lucas, I.; Palakodeti, A.; Jiang, Y.; Young, D. J.; Jiang, N.; Fernald, A. A.; Le Beau, M. M. High-throughput mapping of origins of replication in human cells. *EMBO Rep.,* **2007,** *8*(8), 770-777.

[110] 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. Replication-associated mutational asymmetry in the human genome. *Mol. Biol. Evol.,* **2011,** *28*(8), 2327-2337.

[111] Polak, P.; Arndt, P. F. Long-range bidirectional strand asymmetries originate at CpG islands in the human genome. *Genome Biol. Evol.,* **2009,** *1*, 189-197.

[112] Audit, B.; Nicolay, S.; Huvet, M.; Touchon, M.; d'Aubenton-Carafa, Y.; Thermes, C.; Arneodo, A. DNA replication timing data corroborate in silico human replication origin predictions. *Phys. Rev. Lett.,* **2007,** *99*(24), 248102.

[113] Audit, B.; Zaghloul, L.; Vaillant, C.; Chevereau, G.; d'Aubenton-Carafa, Y.; Thermes, C.; Arneodo, A. Open chromatin encoded in DNA sequence is the signature of 'master' replication origins in human cells. *Nucleic Acids Res.,* **2009,** *37*(18), 6064-6075.

[114] Baker, A.; Nicolay, S.; Zaghloul, L.; d'Aubenton-Carafa, Y.; Thermes, C.; Audit, B.; Arneodo, A. Wavelet-based method to disentangle transcription- and replication-associated strand asymmetries in mammalian genomes. *Appl. Comput. Harmon. Anal.,* **2010,** *28*(2), 150-170.

[115] Touchon, M.; Nicolay, S.; Arneodo, A.; d'Aubenton-Carafa, Y.; Thermes, C. Transcription-coupled TA and GC strand asymmetries in the human genome. *FEBS Lett.,* **2003,** *555*(3), 579-582.

[116] Lundgren, M.; Andersson, A.; Chen, L.; Nilsson, P.; Bernander, R. Three replication origins in Sulfolobus species: synchronous initiation of chromosome replication and asynchronous termination. *Proc. Natl. Acad. Sci. U S A,* **2004,** *101*(18), 7046-7051.

[117] Robinson, N. P.; Dionne, I.; Lundgren, M.; Marsh, V. L.; Bernander, R.; Bell, S. D. Identification of two origins of replication in the single chromosome of the archaeon Sulfolobus solfataricus. *Cell,* **2004,** *116*(1), 25-38.

[118] Markowitz, V. M.; Chen, I. M.; Palaniappan, K.; Chu, K.; Szeto, E.; Grechkin, Y.; Ratner, A.; Anderson, I.; Lykidis, A.; Mavromatis, K.; Ivanova, N. N.; Kyrpides, N. C. The integrated microbial genomes system: an expanding comparative analysis resource. *Nucleic Acids Res.,* **2010,** *38*(Database issue), D382-90.

[119] Charif, D.; Lobry, J. R. SeqinR 1.0-2: a contributed package to the R-project for statistical computing devoted to biological sequences retrieval and analysis. In *Structural approaches to sequence evolution: Molecules, networks, populations*, Bastolla, U.; Porto, M.; H.E., R.; Vendruscolo, M., Eds. Springer Verlag: New York, **2007**; pp 207-232.

[120] Arakawa, K.; Mori, K.; Ikeda, K.; Matsuzaki, T.; Kobayashi, Y.; Tomita, M. G-language Genome Analysis Environment: a workbench for nucleotide sequence data mining. *Bioinformatics,* **2003,** *19*(2), 305-306.

[121] Arakawa, K.; Tomita, M. G-language System as a platform for large-scale analysis of high-throughput omics data. *Journal of Pesticide Science,* **2006,** *31*(3), 282-288.

[122] Arakawa, K.; Kido, N.; Oshita, K.; Tomita, M. G-language genome analysis environment with REST and SOAP web service interfaces. *Nucleic Acids Res.,* **2010,** *38*(Web Server issue), W700-5.

[123] Gao, F.; Zhang, C. T. DoriC: a database of oriC regions in bacterial genomes. *Bioinformatics,* **2007,** *23*(14), 1866-1867.

[124] Kono, N.; Arakawa, K.; Tomita, M. Comprehensive prediction of chromosome dimer resolution sites in bacterial genomes. *BMC Genomics,* **2011,** *12*, 19.