ELSEVIER

Gene 238 (1999) 59–64

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

www.elsevier.com/locate/gene

Intra-strand biases in bacteriophage T4 genome

Tamiko Kano-Sueoka [a,*], Jean R. Lobry [b], Noboru Sueoka [a]

[a] *University of Colorado, Department of Molecular, Cellular, and Developmental Biology, Boulder, CO 80309-0347, USA*
[b] *Université Claude Bernard, Laboratoire BGBP-CNRS UMR 5558, 43 Bd. De 11 Novembre 1918, F-69622 Villeurbanne Cedex, France*

Received 5 March 1999; received in revised form 14 June 1999; accepted 6 July 1999; Received by G. Bernardi

Abstract

In bacteriophage T4, a major portion of DNA replication is initiated at random along the map, although several proven and putative origins have been described for early replication. In order to analyze the contribution of transcription and translation as well as DNA replication to intra-strand bias from $A=T$ and $G=C$, we examined the pattern of the intra-strand biases in the first, second, and third codon positions of the coding regions as well as the intergenic regions of the T4 genome. We found, along the map, characteristic biases both from $A=T$ and $G=C$ for each codon position and the intergenic regions. The bias patterns were closely associated with the location of the sense and anti-sense segments in the genome. The results suggest that: (1) transcription-associated mutation is likely a significant cause of the bias, which is suggested by the pattern of the AT bias (bias from $A=T$) in the third codon position; (2) DNA replication coupled bias may also exist, which is suggested by the pattern of the GC bias (bias from $G=C$) in the third codon position and the intergenic regions; and (3) the bias patterns of the first and second codon positions of the sense segments are consistent with universal properties of the coding sequence that G is in excess and T is deficient in the first codon position, and G is deficient in the second codon position.

Keywords: Asymmetric mutation pressure; Bias from Parity Rule 2; Sense versus anti-sense strand; Transcription- and translation-associated biases

1. Introduction

Bacteriophage T4 is a virulent DNA phage that grows in *Escherichia coli*. The genome consists of about 169 kb (see Kutter et al., 1994) and the complete genome sequence is available at an ncbi ftp site. The packaged DNA molecules are linear duplex. However, their ends are permuted over the circular map with terminal redundancies of approximately 3% (Streisinger, 1966). The mode of replication of the phage DNA is unique in that: (a) there is a multiple number of origins of DNA replication; (b) the origins are used only for the first round of replication (origin-dependent); and (c) the majority of replication is initiated from recombinational intermediates presumably at any place in the genome (origin-independent, recombination-dependent) (see Kreuzer and Morrical, 1994; Mosig, 1998). The T4 genes can be classified into two functional groups. The early genes that are active during the early phase of infection are mostly for T4 DNA replication and mRNA synthesis, and the late genes that are active during the late phase of infection are for the structural proteins and their assemblies (see Kutter et al., 1994). The early and late genes are clustered along the map. In addition, the early genes are transcribed counterclockwise using the early and middle promoters, whereas the late genes are transcribed clockwise using the late promoters (see Kutter et al., 1994).

In bacteria, the intra-strand biases from $A=T$ and $G=C$ [Parity Rule 2 (PR2) (Lobry, 1995; Sueoka, 1995, 1999a)] are clearly seen as coupled with DNA replication (Lobry, 1996a,b; Lobry and Sueoka, 1999). The varied degrees and nature of the biases have been observed that are associated with DNA replication due to asymmetric directional mutation pressure in leading and lagging strands (Mclean et al., 1998; Lobry and Sueoka, 1999). In bacteriophage T4, since the bulk of DNA replication does not initiate from fixed origins, the contribution of DNA replication in intra-strand bias may be small, and instead we may be able to detect biases associated with transcription or translation. We have therefore analyzed the intra-strand bias pattern in the first, second, and third codon positions of the coding regions as well as the intergenic regions of a contiguous

* Corresponding author. Tel.: +1-303-492-8451; fax: +1-303-492-7799.
E-mail address: tamiko@stripe.colorado.edu (T. Kano-Sueoka)

0378-1119/99/$ – see front matter
PII: S0378-1119(99)00296-6

strand. The bias patterns thus obtained were analyzed in relation to the origins of DNA replication and the direction of transcription.

2. Materials and methods

Completely sequenced and annotated T4 genome was retrieved from an ncbi ftp site (ftp://ncbi.nlm.nih.gov/repository/t4phage/old files/T4.gb). The genomic map of T4 phage and the direction of transcription of the genome are as described by Kutter et al. (1994). All probable coding sequences (187 of them) were used for our computation. 147 of them are longer than 200 bp. The contiguous strand that starts clockwise 5′–3′ direction from the map position 0 (the published sequence) was used for calculation. To estimate the intra-strand bias from $A=T$ and $G=C$, the cumulative values for $C-G$ [$\Sigma(C-G)$], and those for $T-A$ [$\Sigma(T-A)$], were calculated for coding regions and non-coding (intergenic) regions. Cumulative values have been used by Grigoriev (1998) to express the intra-strand GC bias. Here, A, T, G, and C are the number of corresponding nucleotides in each sequence. For coding regions, the values were calculated separately for the first, second, and third codon positions. The position of a data point for a coding sequence presented in Fig. 1A and B represents the middle of an open reading frame. The position of a data point for an intergenic region represents the 5′ end of the region. The value for the preceding intergenic region was used where there was no intergenic sequence between the two genes. When the segments of the strand were anti-sense, codon positions of the corresponding sense strand were used to calculate the values. The G+C content of P_3 for each coding region was calculated using the sense strand. P_3 has been defined as the G+C content of the third codon position of the total codons minus ATG (methionine), TGG (tryptophane), ATA (isoleucine), and the termination codon (TAA, TAG, or TGA) (Sueoka, 1995).

3. Results and discussion

3.1. Pattern of intra-strand bias

The analysis of intra-strand bias from $A=T$ and $G=C$ in the T4 genome showed a distinct bias pattern along the map. Fig. 1A shows the cumulative values of $T-A$ in a contiguous strand of the T4 genome for the three codon positions as well as for the intergenic regions. The cumulative values of $C-G$ are shown in a similar manner in Fig. 1B. Along the map, characteristic, distinct biases both from $A=T$ (the AT bias) and $G=C$ (the GC bias) were detected for each codon position. There are several distinct reflection points where the bias pattern changes such that the pattern is symmetrical at both sides of the reflection points. In the first as well as second codon positions, both A/T and G/C pairs show this bias pattern, whereas in the third codon position only the A/T pair shows this pattern. Based on the bias pattern, the genome can be divided into six segments of DNA (160–78, 78–107, 107–116, 116–123, 123–150 and 150–160 kb) (the assignment of these nucleotide positions is approximate), where the pattern of the three segments (160–78, 107–116, 123–150 kb) is similar and that of the other three (78–107, 116–123, 150–160 kb) is similar. The second group of segments is the mirror image of the first three segments in their bias pattern. This means that the bias pattern of a strand in the first three segments is similar to that of the complementary strand of the other three segments. For example, in the first codon position the $T-A$ values for the segment 0–78, 78–106, 123–150 kb are approximately +11, −12, and +11 bases/100 bases, respectively. Likewise, in the third codon position, the $T-A$ values for the same set of segments are around −12, +11, and −12 bases/100 bases, respectively.

The exception is the GC bias in the third codon position where there are more C's than G's in the analyzed strand throughout the genome, although the extent of bias is relatively small (~2.5 bases/100 bases) compared to those in other codon positions or the AT bias in the third codon position (~12 bases/100 bases). Intergenic sequences showed little bias from $A=T$, whereas small, but significant, bias was observed for the G/C pair (~3.7 bases/100 bases). The nature of the GC bias in the third codon position and the intergenic regions was further compared by calculating the values for PR2 of the G/C pair. Interestingly, both positions had very similar degrees of bias. Namely, the third codon position had 54.9% C (7094 C's and 5829 G's) whereas the intergenic regions had 55.2% C (2591 C's and 2106 G's). The result suggests that biases from $G=C$ (~5%) in these two positions may be caused by a similar mechanism(s).

3.2. Correlation of intra-strand bias patterns with origin of DNA replication and direction of transcription

In order to understand the causes for the observed intra-strand biases their correlation with the direction of DNA replication and also that with transcription were examined. The results described in Section 3.1 are summarized in Fig. 2 in the form of a circular genomic map that includes the origins of DNA replication as well as the direction of transcription.

As described in the Introduction, there are a multiple number of origins of DNA replication in the T4 genome that are used only for the first round of replication (Fig. 2). Two of them (oriF and oriG) have been shown

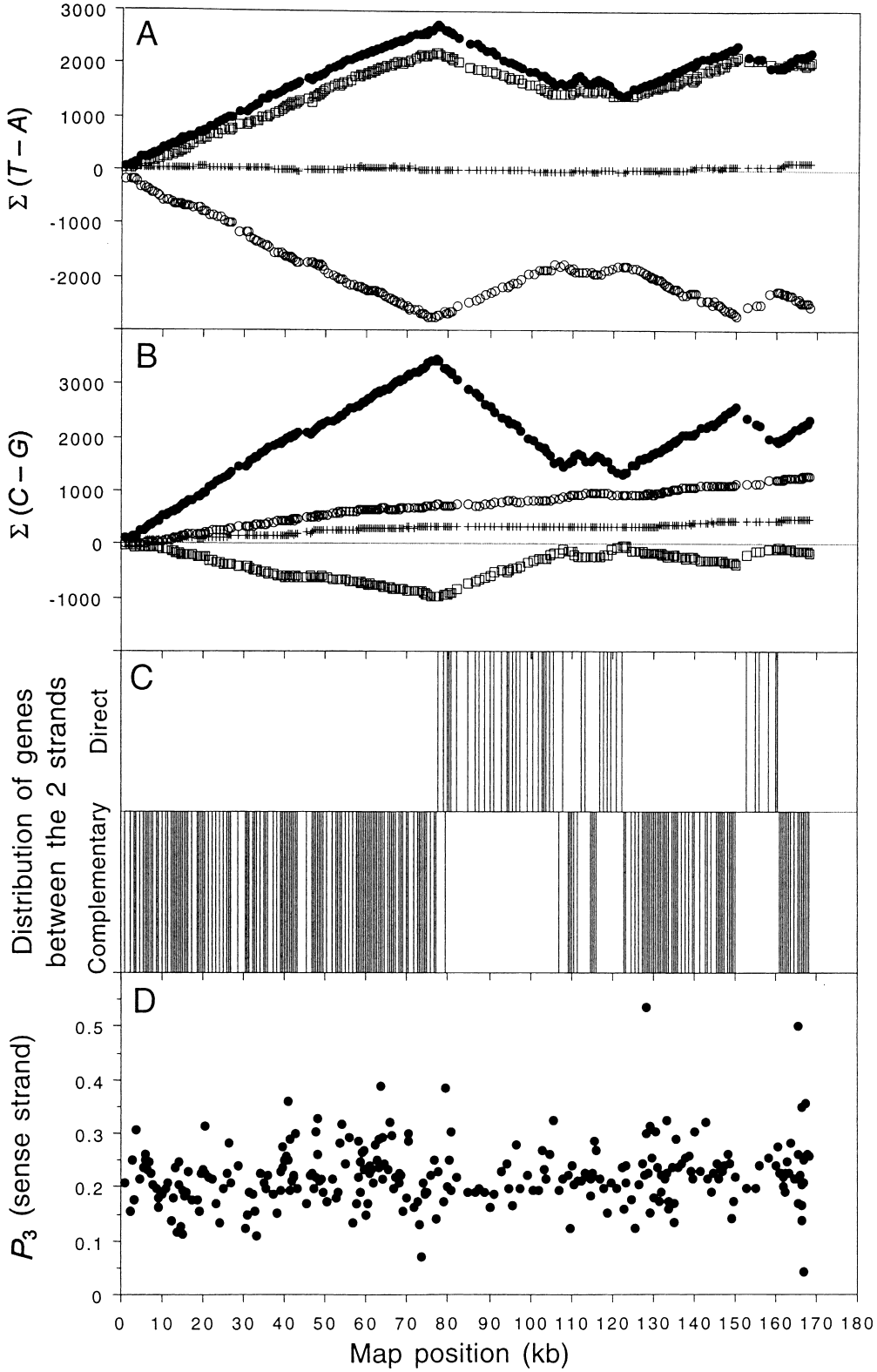

Fig. 1. Intra-strand biases in relation to sense versus anti-sense regions and P_3 values. (**A**) The cumulative values of $T-A$ in a contiguous strand of the T4 genome for the first, second, and third codon positions as well as for the intergenic regions were plotted against the map position. The contiguous strand that starts clockwise 5′–3′ direction from the map position 0 (published sequence) was used for the calculation. ●, First codon position; □, second codon position; ○, third codon position; +, intergenic region. (**B**) The cumulative values of $C-G$ were plotted as in the case of (**A**). (**C**) Vertical lines show the distribution of genes between the two strands of DNA. 'Direct' is the strand with which the analysis was carried out, and 'Complementary' is the complementary strand to Direct strand. (**D**) P_3 value (defined in Section 2) of each gene was plotted against its map position.

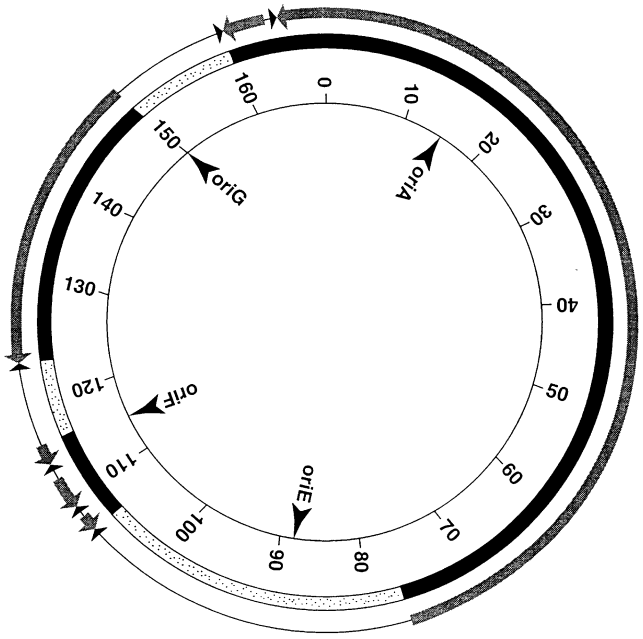

Fig. 2. The genomic map of bacteriophage T4 and distribution of two types of segments with regard to the pattern of intra-strand biases. The inner circle represents the circular map of the T4 genome with the nucleotide positions and the origins of DNA replication. The middle circle presents the areas of the genome where two types of segments similar in intra-strand bias pattern are located. ■, The segments in which the sequenced strand contains anti-sense sequences. These segments have literally the same bias pattern. ⚄, The segments in which the sequenced strand contains sense sequences. Their bias pattern is the mirror image of the segments indicated by the filled segments except for the GC bias in the third codon position. The outer circle indicates the direction of transcription along the map.

to initiate the replication in vivo (Kreuzer and Alberts, 1985, 1986). Two additional ones (oriA and oriE) that are highly likely origins are also indicated in Fig. 2 (see Kreuzer and Morrical, 1994). The three additional regions that are not shown in Fig. 2, around 29–35 kb (oriB), 62–64 kb (oriC), and 73 kb (oriD), may possibly be origins also (see Kreuzer and Morrical, 1994). There is no known terminus of replication. Since these origins are used for limited rounds of DNA replication, the contribution of replication coupled, asymmetric mutations (that may take place at the opposite side of the origins) on the intra-strand bias must be rather small, if any. The results indeed indicate that not all origins are at the reflection points, and not all reflection points have origins in the vicinity. The origins E, F, and G are in the vicinity of the reflection points at 78, 116 and 150 kb, respectively. However, these reflection points also coincide with the place where the direction of transcription switches, as described in detail below. The origins A, B, C, and D are not in the vicinity of any reflection points, and no origins are found near to or at the reflection points 107, 123, and 160 kb.

In contrast, there is a remarkable coincidence between the bias pattern and the direction of transcription. Fig. 1C shows the distribution of coding sequences between the two strands of DNA. By comparing Fig. 1A, B and C, one can see clearly that at the map positions 78 kb, 107 kb, 116 kb, 123 kb, 150 kb, and 160 kb, the direction of transcription changes and so does the bias pattern. Fig. 2 summarizes the findings along the circularly permuted map of the T4 genome. In the filled black segments the sequenced strand (5′–3′ direction clockwise from the map position 0) contains anti-sense sequences, whereas in the dotted segments the sequenced strand contains sense sequences. In the first codon position, whichever the strand is, the sense regions always have $A>T$ and $G>C$, whereas the anti-sense regions have $A<T$ and $G<C$ that are directly opposite to the sense regions. Likewise, in the second codon position, the sense regions have $A>T$ and $G<C$ whereas the anti-sense regions have $A<T$ and $G>C$. In the third codon position the sense regions have $A<T$ whereas the anti-sense regions have $A>T$. Contrary to the above, the GC bias in the third codon position ($G<C$) is consistent in a contiguous strand throughout the map whether a given segment is sense or anti-sense. The 105–115 kb segment contains both early and late genes. However, the majority are early genes whose direction of transcription is counterclockwise, and the bias pattern of this segment seems to be consistent with that of other areas. Table 1 summarizes the nature of the bias pattern at different codon positions and intergenic regions.

The first and second codon positions in the sense segments have a similar bias pattern throughout the genome. The bias patterns are consistent with properties of the coding sequences found among a wide variety of organisms where $G>C$ and $A>T$ in the first codon position and $G<C$ in the second codon position (Trifonov, 1987). The third codon position is presumed to be most neutral (see Lee, 1997; Sueoka, 1999a). Similar to the case of the first and second codon positions, the bias pattern of the A/T pair in the third codon position is well correlated with whether the section of DNA is sense or anti-sense, indicating that the direction of transcription is correlated with the AT bias. This is expected if the mutations and repairs during transcription lead to more T's than A's in the sense strand. Based on the fact that in the transcribed strand C→T transition occurs more frequently than G→A in *E. coli*, Francino and Ochman (1997) proposed that

Table 1
Intra-strand bias from $A=T$ and $G=C$ in the sequenced strand

Position	Sense segment	Anti-sense segment	ITG[a]
First codon	$A>T$, $G>C$	$A<T$, $G<C$	
Second codon	$A>T$, $G<C$	$A<T$, $G>C$	
Third codon	$A<T$, $G<C$	$A>T$, $G<C$	
			$A=T$, $G<C$

[a] Intergenic region.

damages and repairs that occur in the transcribed strand (anti-sense) may cause the biases. They also proposed that deamination of C (that becomes T) in the sense regions may be more frequent than in the anti-sense regions. As far as the third codon position in the T4 genome is concerned, transcription coupled repair seems to play a more significant role on bias than deamination, since we find $C > G$ and $T < A$ in the transcribed (anti-sense) strand (Fig. 1 and Table 1). Cowe and Sharp (1991) found that the pattern of codon usage in T4 phage is different between the early and late genes as well as between highly and lowly expressed late genes. Changes in the pattern are correlated with preferred use of phage encoded tRNAs that recognize codons having A's at the third codon position. The result shown in Fig. 1A cannot detect such change in the early and late genes because the amount of overall change in codons having A's in the third codon position is not large enough to be detected by our analysis.

The GC bias in the third codon position is quite different from the AT bias or the biases of other codon positions, in that the bias is consistent within a strand. The intergenic regions also exhibit a similar pattern of GC bias, suggesting that the third codon position and the intergenic regions are under asymmetric mutation pressure. Interestingly, as shown in Section 3.1, the degree of PR2 bias is literally the same between the intergenic regions and the third codon positions $[C/(C+G)=0.55]$. These results indicate that there is a preference of a strand that is used to initiate recombination-dependent replication, causing the strand-specific GC bias.

The G+C content (P_3) is more or less constant throughout the genome although there is a considerable fluctuation (Fig. 1D). This indicates that in the T4 genome strand-independent, directional mutation pressure (symmetric mutation pressure) is similar throughout the genome irrespective of the difference in PR2 in the sense and anti-sense strands. This result supports the notion that PR2 bias is independent of the G+C content in bacteria (Lobry and Sueoka, 1999; Sueoka, 1999a,b).

It will be of interest to examine the effect of transcription on intra-strand biases in other systems where the sense and anti-sense segments are well organized as in the case of the T4 genome. Indeed, bacteriophage λ shows a similar type of intra-strand bias patterns where the biases are in general specific for the sense or the anti-sense strand (Kano-Sueoka et al., unpublished results).

4. Conclusions

1. The analysis of intra-strand bias from $A = T$ and $G = C$ showed a distinct bias pattern in the T4 genome.
2. The pattern of the AT bias in the third codon position was very similar in all sense regions and that of the anti-sense regions was the mirror image of the sense regions. It is very likely that this bias is coupled with transcription.
3. The GC bias in the third codon position and the intergenic regions is significant, although the extent of the bias is smaller than in other instances. Their bias pattern suggests the possible existence of replication coupled bias.
4. The AT as well as the GC bias patterns in the first and second codon positions were literally the same in all sense regions, and the patterns of the anti-sense regions were the mirror image of the sense regions. The bias patterns of these positions are similar to those found in other organisms.
5. The T4 genome is the first case where transcription may be the main factor involved in the intra-strand bias of the third codon position from PR2.

Acknowledgements

This work was initiated while T.K-S. and N.S. were visiting Université Claude Bernard.

References

Cowe, E., Sharp, P.M., 1991. Molecular evolution of bacteriophages: discrete patterns of codon usage in T4 genes are related to the time of gene expression. J. Mol. Evol. 33, 13–22.

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

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

Kreuzer, K.N., Alberts, B.M., 1985. A defective phage system reveals bacteriophage T4 replication origins that coincide with recombination hot spots. Proc. Natl. Acad. Sci. USA 82, 3345–3349.

Kreuzer, K.N., Alberts, B.M., 1986. Characterization of a defective phage system for the analysis of bacteriophage T4 DNA replication origins. J. Mol. Biol. 188, 185–198.

Kreuzer, K.N., Morrical, S.W., 1994. Initiation of DNA replication. In: Karam, J.D. (Ed.), Molecular Biology of Bacteriophage T4. American Society of Microbiologists, Washington, DC, pp. 28–48.

Kutter, E., Stidham, T., Guttman, B., Kutter, E., Batts, D., et al., 1994. Genomic map of bacteriophage T4. In: Karam, J.D. (Ed.), Molecular Biology of Bacteriophage T4. American Society of Microbiologists, Washington, DC, pp. 491–519.

Lee, W.H., 1997. Molecular Evolution. Sinauer Associate, Sunderland, MA.

Lobry, J.R., 1995. Properties of a general model of DNA evolution under no-strand-bias conditions. J. Mol. Evol. 40, 326–330. Erratum 41, 680.

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

Lobry, J.R., 1996b. Origin of replication of *Mycoplasma genitalium*. Science 272, 745–746.

Lobry, J.R., Sueoka, N., 1999. Asymmetric directional mutation pressure in ten eubacterial genomes. Proc. Natl. Acad. Sci. USA.

submitted.

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

Mosig, G., 1998. Recombination and recombination-dependent DNA replication in bacteriophage T4. Annu. Rev. Genet. 32, 379–413.

Streisinger, G., 1966. Terminal redundancy or all's well that ends well. In: Cairns, J., Stent, G.C., Watson, J.D. (Eds.), Phage and the Origins of Molecular Biology. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 335–340.

Sueoka, N., 1995. Intra-strand parity rules of DNA base composition and usage biases of synonymous codons. J. Mol. Evol. 40, 318–325. Erratum 42, 323.

Sueoka, N., 1999a. Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A=T and G=C. J. Mol. Evol. 49, 49–62.

Sueoka, N., 1999b. DNA G+C content and violation of Parity Rule 2 in human genes are two mostly independent evolutionary events. Gene. in press.

Trifonov, E.N., 1987. Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16SrRNA nucleotide sequences. J. Mol. Biol. 194, 647–652.