Nucleotide composition of the genomic and protein-coding sequences in the two DNA strands

Nucleotide Composition of the Genomic and Protein-Coding Sequences in the two DNA Strands

Hiroshi Nakashima

Division of Health Science, Kanazawa University, Kanazawa, Japan

Abstract— **The nucleotide composition of protein-coding genes in the two DNA strands of *Escherichia coli, Bacillus subtilis, Methanococcus jannaschii* and mitochondrial genes of human and fruit fly was studied. *E. coli, B. subtilis* and *M. jannaschii* indicated compositional asymmetry in their genomic sequences. The protein-coding genes of *E. coli* and *B. subtilis* showed the influence of compositional asymmetry in their compositions, however, no influence was observed in *M. jannaschii.* Mitochondrial protein-coding genes showed significant difference in composition in the two DNA strands. The deviation of nucleotide composition in the two DNA strands is discussed.**

Keywords— **nucleotide composition, compositional asymmetry, leading and lagging strands, cumulative GC skew**

I. Introduction

Comparison of nucleotide composition is a simple way to analyze the character of nucleotide sequences. In a double stranded DNA, adenine pairs with thymine and guanine pairs with cytosine, therefore, the amount of adenine is equal to thymine and guanine is equal to cytosine in DNA. This is known as Chargaff's first rule. This rule also roughly holds for single stranded DNA in some species. This is known as Chargaff's second rule. It is reported that guanine plus cytosine (G+C) content is fairly constant in a bacterial genome. However, G+C content differs considerably among species, and ranges from 25% to 75% [1]. Karlin's group has reported that genes from bacteria have their species specific nucleotide compositions based on the relative ratio between observed and expected dinucleotide frequencies [2-4]. Even genes encoding homologous proteins from different species were discriminated by their dinucleotide frequencies [5]. The protein-coding genes from nine genomes were classified into their species with accuracy of 80% in terms of dinucleotide frequencies [6].

The compositional asymmetry in a genomic sequence has been reported [7-10]. It is considered that the biased mutational occurrences due to the different replication mechanism between leading and lagging strands might be the cause of compositional asymmetry. Plots of cumulative GC skew of *E. coli* genomic sequence indicated maximum and minimum points [9]. The minimum point is consistent with the site of replication origin and the maximum with terminus, and the nucleotide composition asymmetry reversed at these two points in the genomic sequence. It is known that the leading strand contains more guanine than cytosine compared to lagging strand. The deviation between adenine and thymine is smaller than that between guanine and cytosine. Due to the compositional asymmetry in a genomic sequence, it is anticipated that the nucleotide composition of protein-coding genes would be different between the genes in leading and lagging strands. The deviation of nucleotide composition of protein-coding genes in the two DNA strands was investigated in this study. It is reported that the nucleotide composition of genes in the light (L) and heavy (H) strands of mitochondrial DNA (mtDNA) is different [11]. Mammalian mtDNA encodes thirteen proteins, which are components of enzyme complexes of the inner mitochondrial membrane that function in electron transport chain and hence oxidative phosphorylation. The proteins encoded on mtDNA are rich in hydrophobic amino acids [12]. mtDNAs of human, *Homo sapiens* [13] and fruit fly, *Drosophila melanogaster* [14] were also studied.

II. Materials and methods

The genome sequences of *E. coli* [15], *B. subtilis* [16], and *M. jannaschii* [17] were obtained from the National Center for Biotechnology Information (NCBI) (ftp://ftp.ncbi.nih.gov/genomes/), and mitochondrial genomes of human and fruit fly were retrieved from the site of eukaryote organelles of NCBI. The protein-coding nucleotide sequences, which are included in the NCBI ffn file, were used to calculate nucleotide compositions. The nucleotide composition of genomic sequence was calculated using genomic sequence, NCBI gbk file. The protein-coding nucleotide sequences of mtDNAs were obtained according to the annotation of the genomes and their compositions were calculated. The TA and GC skews of genomic sequence were calculated as

TA skew = (T-A)/(T+A)

GC skew = (G-C)/(G+C)

using 1-kbp non-overlapping windows.

The clockwise genomic sequence from the origin to terminus and that from the terminus to origin were obtained. The nucleotide composition of the two genomic sequences

was calculated. The strand which had more guanine than cytosine was regarded as leading strand.

The protein-coding genes in a clockwise strand from the origin to terminus and those in the counterclockwise strand from the terminus to origin were combined together and their average nucleotide composition and that at the third codon position were calculated. Similarly, the protein-coding genes in a counterclockwise strand from the origin to terminus and those in the clockwise strand from the terminus to origin were combined together and their average compositions were calculated. The genes of less than 150 base pairs were excluded in the calculation. The nucleotide composition of whole mtDNA of human and fruit fly was calculated, separately.

III. Results

A. Cumulative GC skew

The cumulative AT/GC skew profiles were obtained by cumulative addition of the skew values along the sequences. The GC skew profile of *E. coli* genomic sequence is shown in Fig.1. The maximum and minimum sites are located at 1.58 Mbp and 3.92 Mbp, they correspond to the terminus and origin of DNA replication [9]. Similarly, GC skew profiles of *B. subtilis* and *M. jannaschii* were obtained using their genomic sequences. The profiles also indicated the maximum and minimum points. However, the AT/GC skew profiles of mtDNAs of human and fruit fly gave no maximum nor minimum.

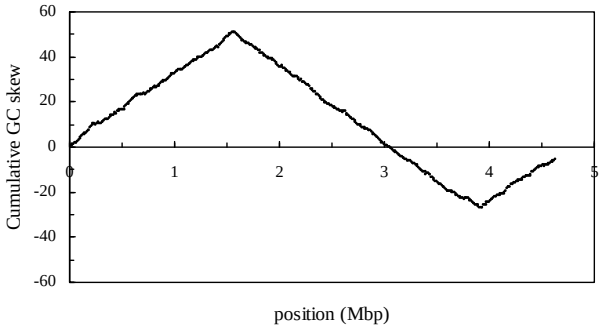

Fig. 1. Cumulative GC skew of *E. coli* genomic sequence.

In *E. coli* and *B. subtilis*, the minimum and maximum points were in correspondence with the experimentally determined origin and terminus of DNA replication [9]. In *M. jannaschii*, the sites of origin and terminus are not experimentally determined. Assuming the minimum and maximum points are the sites of origin and terminus, the nucleotide composition was calculated.

B. Nucleotide composition of genomic sequences

The nucleotide composition of genomic sequences is listed in Table 1. In *E. coli*, guanine was 1.6% greater than cytosine in clockwise genomic sequence from the origin to terminus. Conversely, cytosine was 1.6% greater than guanine in clockwise genomic sequence from the terminus to origin. It is known that DNA replication proceeds bidirectionally from origin to terminus in *E. coli*, and leading strand has more guanine than cytosine. So, the clockwise sequence from origin to terminus and counterclockwise sequence from terminus to origin must be leading sequence. Conversely, cytosine was greater than guanine in lagging strand.

Similarly, the genomic sequence which has more guanine than cytosine was regarded as leading sequence both in *B. subtilis* and *M. jannaschii*. Consequently, the genomic sequence which has more cytosine than guanine was lagging sequence. The deviation between guanine and cytosine was greater than that between adenine and thymine in three species. This result reconfirmed that the asymmetry of GC is greater than that of AT [9]. The deviation between guanine and cytosine and that between adenine and thymine in leading strand was close to the corresponding one in lagging strand with opposite sign.

The largest deviation of nucleotide composition between leading and lagging genomic sequences was 1.8% of guanine in *E. coli*, 3.9% of guanine in *B. subtilis*, and 1.6% of cytosine in *M. jannaschii*.

The nucleotide composition of human mtDNA was significantly deviated from the Chargaff's second rule. Guanine was deficient compared to cytosine and thymine was lesser compared to adenine. The fruit fly mtDNA has high A+T content of 82.2%.

C. Nucleotide composition of protein-coding sequences

The nucleotide composition of protein-coding sequences in the leading and lagging strands is listed in Table 1. In *E. coli*, guanine > cytosine and adenine ≈ thymine were observed both in the leading and lagging strands for the protein-coding whole sequence. At the third codon position, guanine > cytosine and thymine > adenine in the leading strand, and cytosine > guanine and thymine > adenine in the lagging strand was observed. The richness of cytosine in the lagging strand is consistent with that in the genomic sequence. Thymine was preferred to adenine at the third codon position both in leading and lagging strands in *E. coli*.

In *B. subtilis*, the trend of guanine > cytosine and adenine > thymine was observed both in the leading and lagging strands for the protein-coding whole sequence. At the third codon position, guanine > cytosine and thymine > adenine was observed in the leading strand, and cytosine > guanine and thymine > adenine in the lagging strand. The richness of cytosine in the lagging strand was consistent with that in the genomic sequence. Thymine was preferred to adenine at the third codon position both in leading and lagging strands in *B. subtilis* as observed in *E. coli*.

Table 1. Nucleotide composition of genomic and protein-coding sequences.

species	feature	composition (%)			
		A	T	G	C
E. coli	genomic sequence				
	from origin to terminus				
	clockwise, leading strand	24.4	24.6	26.3	24.7
	from terminus to origin				
	clockwise, lagging strand	24.8	24.6	24.5	26.1
	protein-coding sequences				
	2313 genes in leading strand				
	whole sequence	24.3	24.3	27.7	23.7
	codon third position	18.4	26.8	29.8	25.0
	1915 genes in lagging strand				
	whole sequence	24.7	24.2	26.3	24.8
	codon third position	19.0	26.1	27.3	27.6
B. subtilis	genomic sequence				
	from origin to terminus				
	clockwise, leading strand	29.3	26.7	23.8	20.2
	from terminus to origin				
	clockwise, lagging strand	27.2	29.7	19.9	23.2
	protein-coding sequences				
	3004 genes in leading strand				
	whole sequence	30.5	25.9	24.3	19.3
	codon third position	27.4	29.0	24.0	19.6
	1040 genes in lagging strand				
	whole sequence	30.3	26.4	22.3	21.0
	codon third position	27.5	28.4	21.0	23.1
M. jannaschii	genomic sequence				
	from origin to terminus				
	clockwise, leading strand	35.0	33.9	16.2	14.9
	from terminus to origin				
	clockwise, lagging strand	33.7	34.3	15.5	16.5
	protein-coding sequences				
	950 genes in leading strand				
	whole sequence	38.2	29.9	20.8	11.1
	codon third position	39.9	35.2	15.8	9.1
	751 genes in lagging strand				
	whole sequence	38.4	29.8	20.6	11.2
	codon third position	40.3	35.0	15.3	9.4
H. sapiens	mtDNA				
	genomic sequence of L strand	30.9	24.6	13.2	31.3
	12 genes in L strand				
	whole sequence	29.7	25.7	11.7	32.9
	codon third position	37.3	15.5	5.6	41.6
	1 gene in H strand				
	whole sequence	19.4	37.5	35.8	7.3
	codon third position	19.4	41.2	36.0	3.4
D. melanogaster	mtDNA				
	genomic sequence of L strand	41.8	40.4	7.6	10.2
	9 genes in L strand				
	whole sequence	34.1	43.4	10.0	12.5
	codon third position	46.8	47.7	1.1	4.4
	4 genes in H strand				
	whole sequence	30.7	49.0	12.7	7.6
	codon third position	42.4	52.0	4.8	0.8

In *M. jannaschii*, the trend of guanine > cytosine and adenine > thymine was observed both in leading and lagging strands, and their deviations were almost identical. The same trend was observed for nucleotide occurrence at the third codon position. These results indicated that the composition of protein-coding sequences is independent of the compositional asymmetry in genomic sequence of *M. jannaschii.* The richness of purines, adenine/guanine over pyrimidines, thymine/cytosine was significant in the protein-coding sequences of *M. jannaschii.*

The largest deviation of nucleotide composition of the protein-coding whole sequence between leading and lagging strands was 1.4% of guanine in *E. coli*, 2.0% of guanine in *B. subtilis*, and 0.2% of guanine/adenine in *M. jannaschii.* Similarly, the largest deviation at the third codon position was 2.6% of cytosine in *E. coli*, 3.5% of cytosine in *B. subtilis*, and 0.5% of guanine in *M. jannaschii.* The deviation was larger at the third codon position than the whole sequence.

The leading strand has more protein-coding genes than the lagging strand in three species, the leading strand in *B. subtilis* has 3004 genes about three times more genes than the lagging strand of 1040 genes.

The fruit fly mtDNA has high A+T content and the guanine is 7.6%. The deficiency of guanine/cytosine at the third codon position of protein-coding sequence was significant. Guanine content was 1.1% on the L strand and cytosine was 0.8% on the H strand.

IV. Discussion

The average nucleotide compositions of protein-coding genes in the leading and lagging strands were simply compared to detect compositional differences. The distribution of nucleotide composition of A, T, G and C in protein-coding genes indicated normal distribution with a single peak both in the leading and lagging strands in three species.

The average nucleotide composition of protein-coding genes in a clockwise strand from the origin to terminus and that in the counterclockwise strand from the terminus to origin were almost identical in three species. This holds for the genes in the alternative strands.

The G-C and T-A deviations were different for leading and lagging coding sequences, and the difference was greater at the third codon position in *E. coli* and *B. subtilis*. This result indicated that the compositional deviation is increased when the constraint on amino acids is smaller. To exclude the constraint of amino acids, the composition at the third codon position of eight kinds of four-fold degenerate codons (Ala: GCN, Arg: CGN, Gly: GGN, Leu: CUN, Pro: CCN, Ser: UCN, Thr: ACN and Val: GUN) in leading and

lagging sequences was investigated. The genes which have less than 50 four-fold degenerate codons were excluded to calculate average composition. The average composition is listed in Table 2.

Table 2. Average composition at the third position of eight four-fold codons.

species	feature	composition (%)			
		A	T	G	C
E. coli	2079 genes in leading strand	14.0	23.9	35.8	26.3
	1718 genes in lagging strand	14.9	22.4	32.7	30.0
B. subtilis	2515 genes in leading strand	25.2	26.4	27.6	20.8
	843 genes in lagging strand	25.8	25.8	23.2	25.2
M. jannaschii	685 genes in leading strand	43.7	39.8	8.1	8.4
	525 genes in lagging strand	43.9	39.9	7.7	8.5
H. sapiens	mtDNA				
	11 genes in L strand	39.3	12.9	5.4	42.4
	1 gene in H strand	21.6	35.2	39.8	3.4
D. melanogaster	mtDNA				
	7 genes in L strand	50.0	44.2	2.3	3.5
	3 genes in H strand	28.4	64.7	6.2	0.7

The richness of guanine over cytosine in leading strand was clear in *E. coli* and *B. subtilis.* The largest deviation of nucleotide composition at the third position of eight four-fold codons of protein coding sequences between leading and lagging strands was 3.7% of cytosine in *E. coli*, 4.4% of guanine/cytosine in *B. subtilis*, and 0.4% of guanine in *M. jannaschii*. This deviation was larger than that at the third codon position (Table 1) in *E. coli* and *B. subtilis.* However, the deviation in genomic sequence was largest in *M. jannaschii*. *M. jannaschii* belongs to thermophilic archaea with optimum growth temperature of 85°C [18]. To get thermal stability for their DNA and proteins, the nucleotide composition may have no room to obey compositional asymmetry. The protein-coding sequences occupy 87% of genomic sequence of *M. jannaschii* [19], therefore, the remaining 13% of non-coding sequence must be the cause of the observed compositional asymmetry.

Mammalian mtDNAs have two origins of DNA replication for the H and L strands. The replication of mtDNA starts at L strand and H strand remained for hours as single-stranded. Deamination occurs much more frequently on single-stranded than on double-stranded DNA, and this yields many mutations in mtDNA [20]. Therefore, the nucleotide composition of mtDNA genes in the two DNA strands showed a large deviation.

REFERENCES

1. Muto A, Osawa S (1987) The guanine and cytosine content of geomic DNA and bacterial evolution. Proc Natl Acad Sci USA 84: 166-169
2. Karlin S, Cardon LR (1994) Computational DNA sequence analysis. Annu Rev Microbiol 48: 619-654
3. Karlin S, Burge C (1995) Dinucleotide relative abundance extremes: a genomic signature. Trends Genet 11: 283-290
4. Karlin S, Mrázek J, Campbell AM (1997) Compositional biases of bacterial genomes and evolutionary implications. J Bacteriol 179: 3899-3913
5. Nakashima H, Nishikawa K, Ooi T (1997) Differences in dinucleotide frequencies of human, yeast, and *Escherichia coli* genes. DNA Res 4: 185-192
6. Nakashima H, Ota M, Nishikawa K et al. (1998) Genes from nine genomes are separated into their organisms in the dinucleotide composition space. DNA Res 5: 251- 259
7. Lobry JR (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. Mol Biol Evol 13: 660-665
8. Mrázek J, Karlin S (1998) Strand compositional asymmetry in bacterial and large viral genomes. Proc Natl Acad Sci USA 95: 3720-3725
9. Grigoriev A (1998) Analyzing genomes with cumulative skew diagrams. Nucleic Acids Res 26: 2286-2290
10. Touchon M, Nicolay S, Audit B et al. (2005) Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. Proc Natl Acad Sci USA 102: 9836-9841
11. Martin AP (1995) Metabolic rate and directional nucleotide substitution in animal mitochondrial DNA. Mol Biol Evol 12: 1124-1131
12. Nakashima H, Nishikawa K, Ooi T (1990) Distinct character in hdrophobicity of amino acid compositions of mitochondrial proteins. Proteins 8: 173-178
13. Anderson S, Bankier AT, Barrell GB et al. (1981) Sequence and organization of the human mitochondrial genome. Nature 290: 457-465
14. Garesse R (1988) *Drosophila melanogaster* mitochondrial DNA: gene organization and evolutionary considerations. Genetics 118: 649-663
15. Blattner FR, Plunkett GIII, Bloch CA et al. (1997) The complete genome sequence of *Escherichia coli* K-12. Science 277: 1453-1462
16. Kunst F, Ogasawara N, Moszer I et al. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390: 249-256
17. Bult CJ, White O, Olsen GJ et al. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii.* Science 273: 1058-1073
18. Nakashima H, Fukuchi S, Nishikawa K (2003) Compositional changes in RNA, DNA and proteins for bacterial adaptation to higher and lower temperatures. J Biochem 133: 507-513
19. Nakashima H, Nishikawa K (2000) The genomic DNA sequences of various species are distinctively distributed in nucleotide composition space. Res Commun Biochem Cell Mol Biol 4: 25-45
20. Xia X (2005) Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. Gene 345: 13-20

The address of the corresponding author:

Author: Hiroshi Nakashima
Institute: Department of Clinical Laboratory Science,
Graduate Course of Medical Science and Technology,
Division of Health Science, Kanazawa University
Street: Kodatsuno 5-11-80
City: Kanazawa 920-0942
Country: Japan
Email: naka@kenroku.kanazawa-u.ac.jp