A Proposed Solution to the Historic Puzzle of Chargaff's Second Parity Rule

Kenji Sorimachi*

Educational Support Center, Dokkyo Medical University, Mibu, Tochigi 321-0293, Japan

Abstract: Chargaff's first parity rule for the contents of the four nucleotides in DNA is easily understood based on the double-stranded DNA structure. However, the second parity rule, based on similar nucleotide relationships in single-stranded DNA, has been a puzzle in molecular biology, because it is impossible to imagine how pairs of G and C, and A and T are formed in the single DNA strand. In the present study, Chargaff's second parity rule can be solved based on nucleotide contents and the correlation between the two strands in double-stranded DNA.

INTRODUCTION

The first parity rule for the contents of the four nucleotides in DNA, G = C, A = T, and G + A = T + C, was discovered experimentally by Chargaff in1950 [1]. This finding seems to have contributed to the discovery of the double helical structure of DNA by Watson and Crick in 1953 [2]. In addition, Chargaff and his colleagues also reported that similar nucleotide relationships were applicable to the single DNA strand and this became known as Chargaff's second parity rule [3]. Despite being proposed in 1968 and although many complete genomes have been found to obey this rule, its basis remains unknown.

Recently, Mitchell and Bridge (2006) examined 1,495 viral, 835 organelle, 231 bacterial and 20 archaeal genomes, and 164 sequences from 15 eukaryotes, to determine whether these DNAs fit Chargaff 's second parity rule [4]. They reported that only single DNA strands that form cell genome double-stranded DNA obeyed Chargaff's second parity rule [4] whereas some organelle genomes deviated from this rule [5,6].

In general, molecular biology, including genome biology, has progressed based on an understanding of the relationships between gene functions and nucleotide or amino acid sequences. Previously, we determined the ratio of nucleotides to the total number of nucleotides in the coding region on the genome or that of amino acids to the total number of amino acid presumed to be encoded by the genome. This analysis showed that the genome is homogeneously constructed from putative small units consisting of various genes displaying almost the same codon usages [7] and amino acid compositions [8]. These units, each encoding 3,000 – 7,000 amino acid residues, represent the characteristics of the complete genome, even though each gene has a different nucleotide sequence. In addition, symmetry is observed in long nucleotide sequences [9] and in complete genomes [10]. Based on these results, the genomes of both prokaryotes and eukaryotes are homogeneously constructed from certain units consisting of coding and non-coding DNA.

*Address correspondence to this author at the Educational Support Center, Dokkyo Medical University, Mibu, Tochigi 321-0293, Japan;
E-mail: kenjis@dokkyomed.ac.jp

RESULTS AND DISCUSSION

The present study was designed to solve the second parity rule. To analyze the nucleotide contents in double-stranded DNA of the complete genome, the strands were schematically drawn, as shown in Fig. (**1**). The size of open reading frame (ORF) 1, consisting of numerous genes on the forward strand, is almost equal to that of ORF2, consisting of numerous genes on the reverse strand. Indeed, the nucleotide contents in the coding region were almost the same between the forward and reverse strands, and no significant difference was also observed in the non-coding region between the two strands [11]. The non-coding region can be divided simply into two equal parts (non-open reading frame (NORF) 1 and NORF2). Thus, the nucleotide contents in ORF2 can be expressed as a function of the nucleotide contents in ORF1, because they belong to the same genome and have almost the same coding size: Gb ≈ Ga, Cb ≈ Ca, Tb ≈ Ta, and Ab ≈ Aa. Additionally, as b' is complementary to b, the nucleotide contents in the complementary ORF2 can be expressed as a function of the nucleotide contents in ORF1 *via* those in ORF2, as follows: Gb' = Cb ≈ Ca, Cb' = Gb ≈ Ga, Tb' = Ab ≈ Aa and Ab' = Tb ≈ Ta.

In the non-coding region, consisting of two equal parts, the same relationships hold:

Gd ≈ Gc, Cd ≈ Cc, Td ≈ Tc, Ad ≈ Ac, Gd' = Cd ≈ Cc, Cd' = Gd ≈ Gc, Td' = Ad ≈ Ac and Ad' = Td ≈ Tc.

The total contents of G and C in the single DNA strand consisting of four units are:

Ga + Gb' + Gc + Gd' ≈ Ga + Ca + Gc +Cc

Ca + Cb' + Cc + Cd' ≈ Ca + Ga + Cc + Gc

In these two equations, the right-hand sides of the equations are equal.

Therefore,

Ga + Gb' + Gc + Gd' ≈ Ca + Cb' + Cc + Cd'

Finally,

G ≈ C in the single DNA strand.

Similarly,

Ta + Tb' + Tc + Td' ≈ Aa + Ab' + Ac + Ad'

1875-693X/09

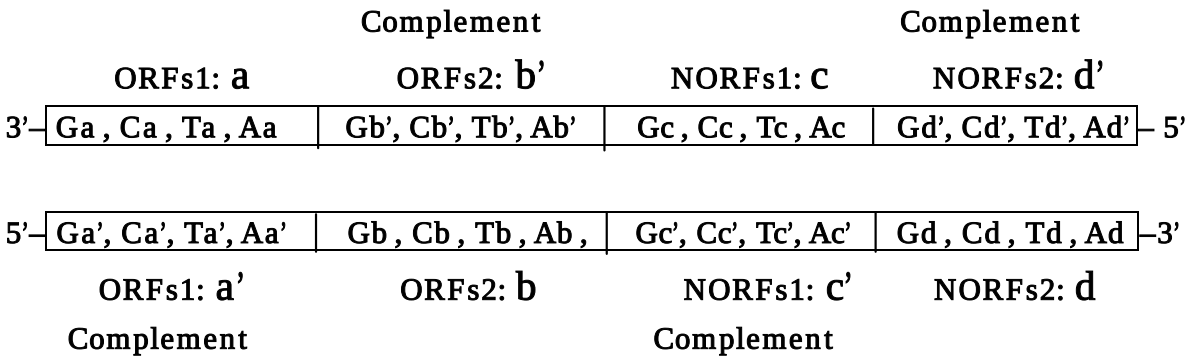

Fig. (1). Schema of the double-stranded DNA of the complete genome. The complete genome is divided into coding and non-coding regions. The coding region is further divided into two parts consisting of ORF1 and complement ORF2, and the non-coding region is also divided into NORF1 and NORF2 in the forward strand. Similar divisions are carried out in the reverse strand. The size of ORF1 is almost equal to that of complement ORF2, and similarly that of NORF1 is equal to that of NORF2. The size of coding region is independent on that of non-coding region. G, C, T and A represent nucleotide contents.

Thus T ≈ A in the single DNA strand.

These results lead us to the assumption that G + A ≈ C + T.

Based on the present result, mitochondria consisting of H and L chains, with completely different gene numbers, appear not to be subject to the second parity rule [5,12]. When the complete genome is divided into two almost equal sized parts, the content of each nucleotide can be expressed as follows (Fig. **2**):

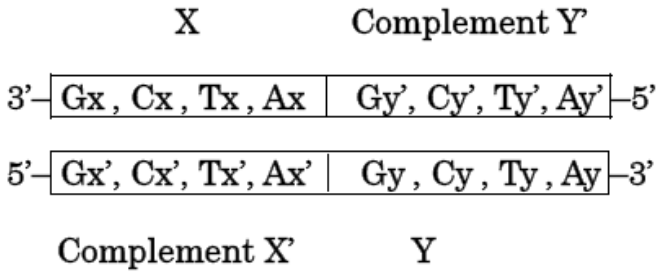

Fig. (2). More detailed schema of the double-stranded DNA of the simplified complete genome. The complete genome is divided into two equal parts. The forward strand consists of X and the complement Y', and the reverse strand consists of X' and Y.

X is almost equal to Y, and X' and Y' are the complements for X and Y, respectively.

Therefore,

Gy ≈ Gx, Cy ≈ Cx, Gy' = Cy ≈ Cx and Cy' = Gy ≈ Gx

The G and C contents in the genome fragments X and Y' are:

Gx + Gy' ≈ Gx + Cx

Cx + Cy' ≈ Cx + Gx

The right-hand sides of these equations are equal.

Thus,

Gx + Gy' ≈ Cx + Cy'

Finally,

G ≈ C

Similarly,

Ty ≈ Tx, Ay ≈ Ax, Ty' = Ay ≈ Ax and Ay' = Ty ≈ Tx

Therefore,

Tx + Ty' ≈ Tx + Ax

Ax + Ay' ≈ Ax + Tx

Thus,

Tx + Ty' ≈ Ax + Ay'

Finally,

T ≈ A, and G + A ≈ C + T

The double-helical structure of DNA is well known [2]. This double-helical structure is based on the physicochemical characteristics of each nucleotide: hydrogen bonding between G and C, and T and A. However, the significance of this structure has not yet been elucidated. It is well known that the replication of DNA and the transcription of mRNA occur based on a single DNA strand after unwinding of the double helix. Thus the function of the double-helical structure is unclear. Recently, we showed that codon evolution, in terms of codon usage, is governed by linear formulas expressing nucleotide contents, not only in prokaryotes, but also in eukaryotes [11]. The present study clearly reveals that the correlationship of nucleotide contents between the double strands is extremely important to solve the second parity rule, and that the double-helical structure of DNA plays an important role in nucleotide substitutions. Thus I conclude that the formation of double-stranded DNA is involved in biological evolution. In addition, it seems that the double-helical structure of DNA might have been formed synchronously with genome establishment during the formation of primitive life [13,14].

REFERENCES

[1] Chargaff E. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. Experimentia 1950; VI: 201-9.

[2] Watson JD, Crick FHC. Genetical implications of the structure of deoxyribonucleic acid. Nature 1953; 171: 964-7.

[3] Rudner R, Karkas JD, Chargaff E. Separation of *B. subtilis* DNA into complementary strands. 3. Direct analysis. Proc Natl Acad Sci USA 1968; 60: 921-2.
[4] Mitchell D, Bridge R. A test of Chargaff's second rule. Biochem Biophys Res Commun 2006; 340: 90-4.
[5] Nikolaou C, Almirantis Y. Deviations from Chargaff's second parity rule in organelle DNA insights into the evolution of organelle genomes. Gene 2006; 381: 34-41.
[6] Bell SJ, Forsdyke DR. Deviations from Chargaff's second parity rule with direction of transcription. J Theor Biol 1999; 197: 63-76.
[7] Sorimachi K, Okayasu T. An evaluation of evolutionary theories based on genomic structures in *Saccharomyces cerevisiae* and *Encephalitozoon cuniculi.* Mycoscience 2004; 45: 345-50.
[8] Sorimachi K, Okayasu T. Gene assembly consisting of small units with similar amino acid composition in the *Saccharomyces cerevisiae* genome. Mycosicence 2003; 44: 415-7.
[9] Prabhu C. Symmetry observations in long nucleotide sequences. Nucleic Acids Res 1993; 21: 2797-800.
[10] Qi D, Cuticcia AJ. Compositional symmetries in complete genomes. Bioinformatics 2001; 17: 557-9.
[11] Sorimachi K, Okayasu T. Codon evolution is governed by linear formulas. Amino Acids 2008; 34: 661-8.
[12] Sorimachi K, Okayasu T. Universal rules governing genome evolution expressed by linear formulas. Open Genomics J 2008; 1**:** 33-43.
[13] Sorimachi K. Evolutionary changes reflected by the cellular amino acid composition. Amino Acids 1999; 17: 207-26.
[14] Sorimachi K, Itoh T, Kawarabayasi Y, Okayasu T, Akimoto K, Niwa A. Conservation of the basic pattern of cellular amino acid composition during biological evolution and the putative amino acid composition of primitive life forms. Amino Acids 2001; 21: 393-9.

Received: February 23, 2009 Revised: March 12, 2009 Accepted: March 12, 2009