The Nucleic Acid Sequences of the Genome Are Highly Structured on a Genome-Wide Scale in Terms of Nucleic Acid Composition Indices Such as TA Skew and GC Skew

Genshiro Esumi
Department of Pediatric Surgery, Hospital of the University of Occupational and Environmental Health, Kitakyushu, Japan

Abstract

Nucleic acid sequences in the genome are often assumed to be random because their mutations are thought to occur randomly. In fact, several analyses have shown that the sequences have a certain structural nature. However, the role of these structures in biological activity has rarely been reported.

In a previous report, I showed that the high and low TA skew of the gene correspond to the transmembrane domains and the intrinsically disordered regions of the protein, respectively. Therefore, in this paper, I examined the variation of TA skew in combination with other nucleic acid composition indices, GC skew and GC content, by one base pair step over the entire genome sequence length at the 1000 base pair sequence windows, which is close to the size of most genes.

In this study, these three indices were calculated for the genomes of the three bacterial species and the genome sequence of human chromosome 1. The results showed that the distribution of GC content was different for each species, but the distributions of TA skew and GC skew were positively and negatively symmetric with zero at the center. In addition, scatter plots of the two indices, TA skew and GC skew, showed a rotationally symmetric distribution in each species.

It has been previously reported that the numbers of T and A and the numbers of G and C in the nucleic acid composition of the genome are almost equal in sequences above a certain length, and this is called Chargaff's second parity rule. However, there has been no report on the correlated behavior of TA skew and GC skew on the genome, and this is the first report on this phenomenon.

The nucleic acid sequences on the genome, which are often thought to be random, are actually highly structured on a genome-wide scale in terms of nucleic acid composition. And I speculated that organisms exploit the cooperative association between this structural nature of the genome and the functional assignment of the genetic code to achieve functional protein synthesis. However, how such large-scale genome structures are built and maintained remains a mystery.

Keywords: genome structure, TA skew, GC skew, GC content, symmetry

Email: esumi@clnc.uoeh-u.ac.jp

1. Background

Nucleic acid sequences in the genome are often assumed to be random because their mutations are thought to occur randomly. In fact, oligonucleotide analyses known as k-mer signature analyses have shown that the genome sequence of each species has certain characteristics, also known as genome fingerprints [1]. Thus, there is a certain degree of structure in the genome sequences of species. However, it has rarely been shown how these structural features are actually involved in biological activities.

In my previous work, I showed that high and low TA skew genes correspond to membrane proteins with many transmembrane domains and intrinsically disordered proteins with many intrinsically disordered regions, respectively [2]. From these, I assumed that the TA skew of the sequence corresponds to the functional properties of the protein. Therefore, in this paper, I examined the variation of TA skew in combination with other two nucleic acid composition indices, GC skew and GC content, by one base pair step over the entire genome sequence length at the 1000 base pair sequence windows, which is close to the size of most genes.

2. Materials and methods

This study used genome sequence data from NCBI and other public databases. I used three bacterial genome sequences with different whole genome GC content [3,4,5] and the chromosome 1 sequence of the human CHM13 gapless genome [6]. The three selected bacteria were Anaeromyxobacter dehalogenans 2CP-C as a representative of high GC content genome bacteria [3], E. coli k-12 as a representative of medium GC content genome bacteria [4], and Candidatus Zinderia insecticola CARI as a representative of low GC content genome bacteria [5]. These bacteria were selected from the list of bacteria with different GC content genome published in a paper [7]. Each TA skew, GC skew and GC content was calculated in 1000 base pair windows, moving one base pair step across the entire genome or chromosome.

In this study, I define three indices: TA skew = (T-A)/(T+A), GC skew = (G-C)/(G+C), and GC content = (G+C)/(T+A+G+C), where each capital letter indicates the number of each nucleic acid in the sequence window. As a result, TA skew and GC skew take values between -1 and 1, and GC contents take values between 0 and 1.

First, the distributions of TA skew, GC skew, and GC content were examined by species.

Second, scatter plots were generated for TA skew with the other two indices to examine their correlations. For visual recognition of the plot distribution, each plot sheet was displayed as a heat map with colors highlighted according to plot densities.

In this study, I used Microsoft® Excel for Mac v16.74 (Microsoft Corporation, Redmond, WA, USA) to generate nucleic acid composition indices and other calculation results. I also used JMP® 17.1.0 (SAS Institute Inc., Chicago, IL, USA) to generate graphs and figures.

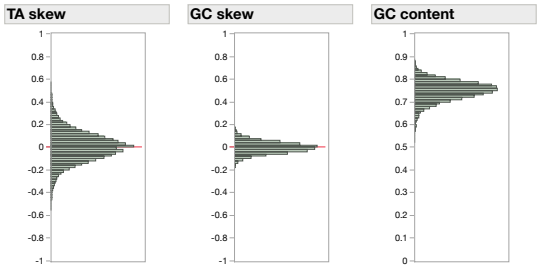
3. Results

3.1. Results of the distributions

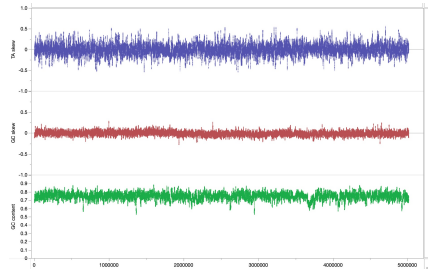
The distributions of the three indices, TA skew, GC skew, and GC content (left column), and these indices by the positions of each genome (right column), are both shown by species in Figure 1. The distributions of GC content in each species differed by species, but the distributions of both TA skew and GC skew were positively and negatively symmetric with zero at the center in all bacterial species, although the shape of the distributions themselves differed. The distributions of TA skew and GC skew were unimodal in bacteria with genomes of high and medium GC content, but they were bimodal in bacteria with genomes of low GC content. The distributions of TA skew and GC skew in human chromosome 1 were both unimodal on a large scale, but each had a sub-peak. These sub-peaks corresponded to repeated sequences of the chromosomal centromere located in the center of the chromosome.

Figure 1

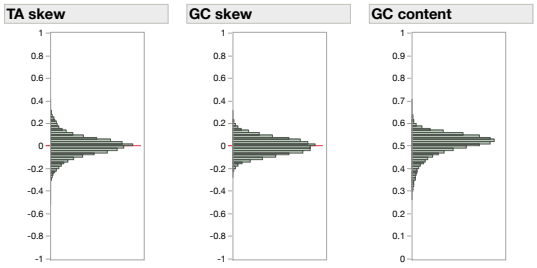
Anaeromyxobacter dehalogenans 2CP-C



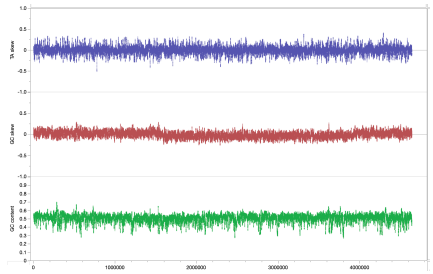
1000bp moving window, n = 5013479



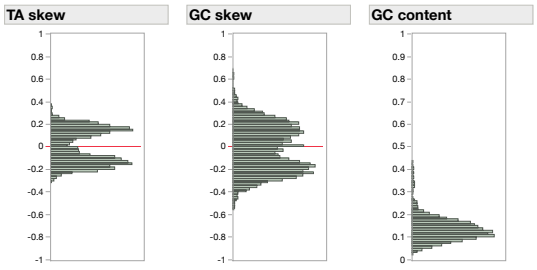
Escherichia coli str. K-12



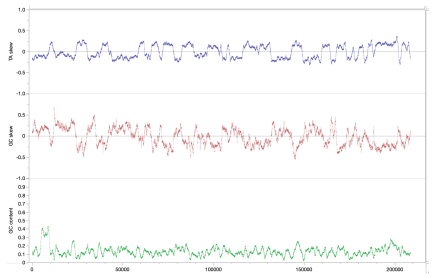
1000bp moving window, n = 4641652



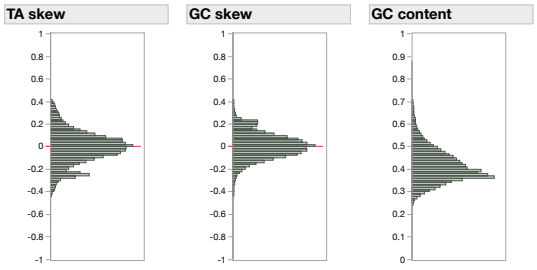
Candidatus Zinderia insecticola CARI



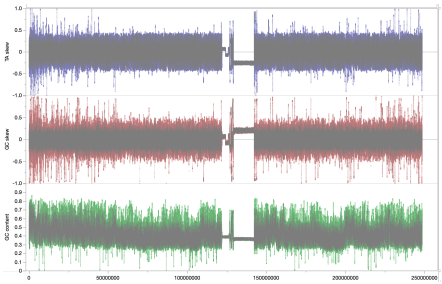
1000bp moving window, n = 208564



Homo sapiens (human), Chromosome 1



1000bp moving window, n = 248386329

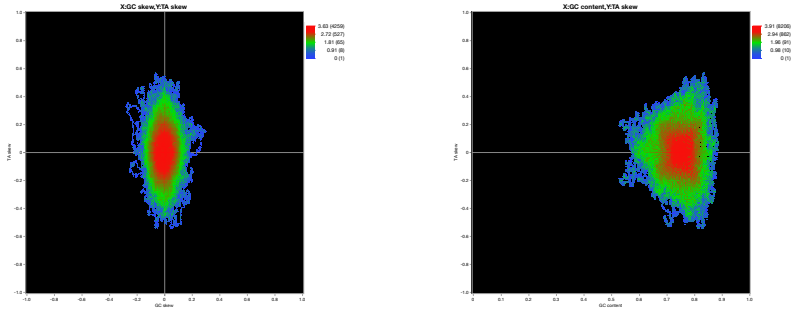


3.2. Results of index correlations on scatter plots

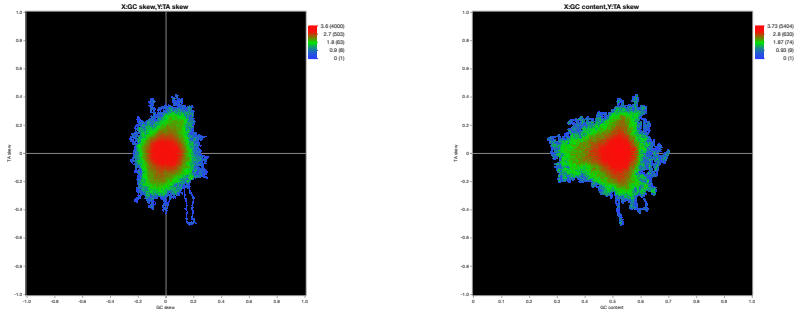
Scatter plots of TA skew and GC skew and scatter plots of TA skew and GC content are shown by species in Figure 2. All plots of TA skew and GC skew were found to be 180° rotationally symmetric (left column), and plots of TA skew and GC content were all found to be horizontal mirror images (right column). The distribution shapes of the plots were different for each of the three bacterial species with different genomic GC content. In addition, the bimodal distribution of TA and GC skew observed in the low-GC bacteria resulted in two clusters in the plot. In contrast, their distributions in humans were significantly broader and more complex than those in bacteria. However, the former was still rotationally symmetric and the latter was still mirror image.

Figure 2

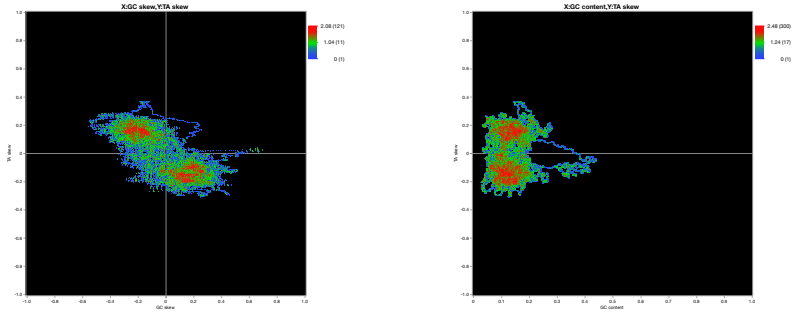
Anaeromyxobacter dehalogenans 2CP-C 1000bp moving window, n = 5013479



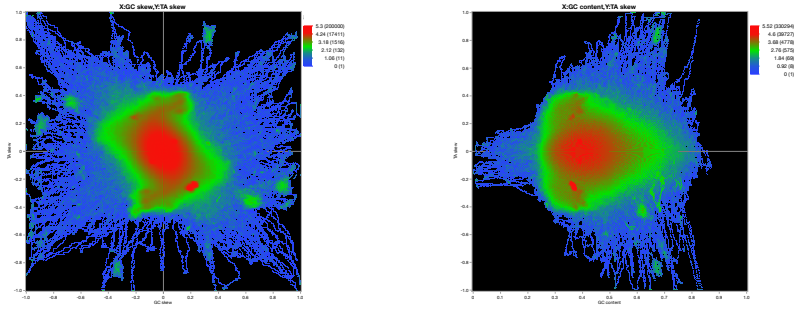
Escherichia coli str. K-12 1000bp moving window, n = 4641652



Candidatus Zinderia insecticola CARI 1000bp moving window, n = 208564



Homo sapiens (human), Chromosome 1 1000bp moving window, n = 248386329



4. Discussions

It has already been reported that the numbers of T and A and the numbers of G and C in a sequence in a genome are approximately equal for sequences above a certain length, which is called Chargaff's second parity rule [8]. It has also been reported that in bacterial genomes, half of the genome is known to have a slightly higher GC skew than the other half, and the boundaries between these compartments have been found to coincide with the origin and terminus of chromosome replication [9]. However, the correlation between TA skew and GC skew in the genome has not been reported.

The results of this study showed that each distribution of TA skew and GC skew in the genome is mirror-image symmetric with zero at the center, and their combined scatter plots showed 180° rotationally symmetric distributions resulting from their elaborate correlation. In the present analyses, only one of the two strands of the genome was analyzed, but when the other strand is analyzed, the result will be a 180° rotation of the present result. Therefore, it can be concluded that the TA and GC skews are distributed with the same correlation in both strands of the genome.

In contrast, the scatter plots of TA skew and GC content showed mirror images. This suggests that the positive and negative mirror image pairs of the TA skews are composed of sequences with similar GC contents. Furthermore, although not shown in the data, the genomic locations of these mirror image pairs were relatively close. These results give the impression that a sequence always has its complement on the same side chain nearby, although in reality such sequences are rarely found.

Meanwhile, if an organism's genome follows Chargaff's second parity rule, the simplest behavior would be for both TA and GC skew to peak at zero [8]. However, in the present analysis, both TA and GC skew were correlated in low-GC bacteria, and their plots formed two clusters at a distance from the origin. These results suggest that sequences of high TA skew correspond to those of low GC skew and vice versa in such a genome. Figure 3 shows 85788 protein genes from 25 bacterial species with genomes of different GC content plotted by their TA skew and GC skew and colored by their GC content [3, 5, 10]. In this plot, genes with low GC content are mainly plotted in the lower right quadrant, the area of lower TA skew and higher GC skew. Figure 4 shows that as the genomic GC content decreases in the bacterial genome, the TA skew decreases and the GC skew increases (arrows) [3, 5, 10]. From these results, I concluded that the origin of the two clusters in the low-GC bacteria could be explained by the low-GC bacteria requiring low TA skew and high GC skew protein genes. This could be an important finding in terms of amino acid assignment in the standard genetic code.

Figure 3

Total 85788 bacterial protein genes

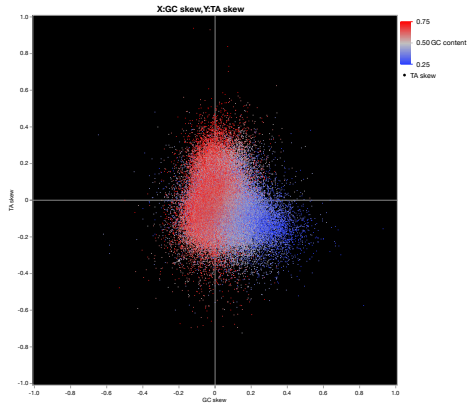
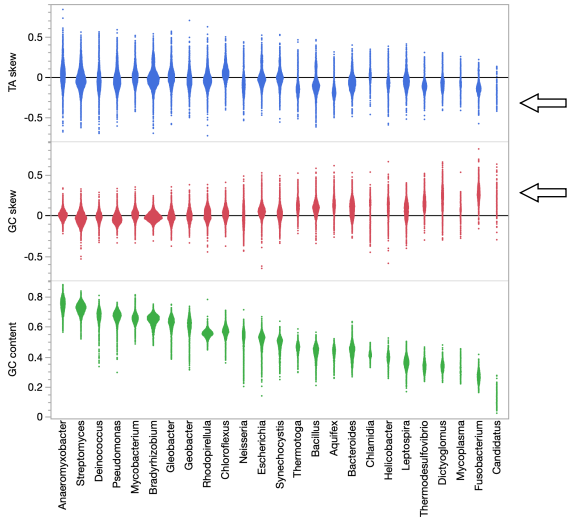


Figure 4

Each index distribution of 25 bacteria



In a previous report, I showed a relationship between the TA skew of genes and their protein functions [2]. However, based on Chargaff's second parity rule, it has also been suggested that both TA skew and GC skew approach zero for sequences above a certain length [8]. From the results of the current study, I have shown that the TA skew varies quite aggressively with the sequence length of a gene, and it is then suggested that different gene functions are encoded by genes with corresponding TA skews.

In this study, I showed that variation in TA skew in the genome sequence is associated with GC skew and GC content, and that these associations constitute the genome sequence and genes. If these inferences from the phenomenon are correct, the genome cannot be random, but is highly and extensively structured. Finally, I speculated that organisms use the cooperative relationship between the structural properties of the genome and the functional properties of the genetic code to achieve structured functional protein synthesis. This report is the first to show that the nucleic acid composition of the genome is structured on a genome-wide scale, and it is also the first English-language report to mention that genome structure may play a role in the synthesis of functional proteins.

Originally, there was no clear explanation as to why Chargaff's second parity rule was observed in the genomes of almost all organisms. However, the present results inductively suggest that this second parity rule may have resulted from the coordinated mirror-image correspondence between TA skew and GC skew, and that this mirror-image correspondence may have been necessary for functional protein synthesis. From this perspective, the mirror-image correspondence between TA skew and GC skew on the genomic sequence may be an upper concept of Chargaff's second parity rule.

5. Conclusion

In this paper, I showed that nucleic acid sequences on the genome, often considered random, are highly structured on a genome-wide scale in terms of nucleic acid composition. And I speculated that organisms exploit the cooperative relationship between the structural properties of the genome and the functional properties of the genetic code to achieve structured functional protein synthesis. On the other hand, how such large-scale genomic structures are built and maintained remains a mystery. The abyss of life's mysteries seems infinitely deep, I think.

6. References

1. Gautier, C. (2000). Compositional bias in DNA. *Current Opinion in Genetics & Development*, *10*(6), 656–661. https://doi.org/10.1016/S0959-437X(00)00144-1
2. Esumi, G. (2023). The α-helical transmembrane domains and intrinsically disordered regions on the human proteins are coded for by the skews of their genes' nucleic acid composition with the "universal" assignment of the genetic code table. *Jxiv*. https://doi.org/10.51094/jxiv.247
3. Anaeromyxobacter dehalogenans 2CP-C, complete genome, GenBank: CP000251.1, NCBI website, https://www.ncbi.nlm.nih.gov/nuccore/CP000251
4. Escherichia coli str. K-12 substr. MG1655, complete genome, NCBI Reference Sequence: NC_000913.3, NCBI website, https://www.ncbi.nlm.nih.gov/nuccore/556503834
5. Candidatus Zinderia insecticola CARI, complete genome, NCBI Reference Sequence: NC_014497.1, NCBI website, https://www.ncbi.nlm.nih.gov/nuccore/NC_014497.1
6. Genome dataset, “Homo sapiens (human) / Genome assembly T2T-CHM13v2.0", NCBI website, https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_009914755.1/
7. Almpanis, A., Swain, M., Gatherer, D., & McEwan, N. (2018). Correlation between bacterial G+C content, genome size and the G+C content of associated plasmids and bacteriophages. *Microbial Genomics*, *4*(4). https://doi.org/10.1099/mgen.0.000168
8. Rudner, R., Karkas, J. D., & Chargaff, E. (1968). Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. *Proceedings of the National Academy of Sciences*, *60*(3), 921–922. https://doi.org/10.1073/pnas.60.3.921
9. Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biology and Evolution*, *13*(5), 660–665. https://doi.org/10.1093/oxfordjournals.molbev.a025626
10. Esumi, G. (2022). Synonymous codon usage and its bias in the bacterial proteomes primarily offset guanine and cytosine content variation to maintain optimal amino acid compositions. *Jxiv*. https://doi.org/10.51094/jxiv.99