Synonymous codon usage and its bias in the bacterial proteomes primarily offset GC content variation to maintain optimal amino acid compositions

Genshiro Esumi
University of Occupational and Environmental Health, Japan

Abstract

Codon usage bias is the preferential or non-random synonymous codon usage among species. A recent review concluded that their biases are a complex phenomenon influenced by numerous factors, including genome composition, GC content, expression level, length of genes, and recombination rates. In this paper, I present a new plot chart and show a more straightforward explanation of the primary function of the synonymous codon usage and its bias.

First, I calculated each protein's amino acid compositions and its gene's nucleotide compositions from the publicly available proteome cording sequence data set of 23 different bacteria. Next, I calculated the maximum and minimum GC contents of the possible gene variations of the amino acid composition of each protein. And then, they were plotted together by their actual GC content on a scatter plot.

The plot showed a clear tendency. Proteins with lower actual GC content genes are coded for by genes closer to the minimum possible GC content. On the other hand, proteins with higher actual GC content genes are coded for by genes closer to the maximum possible GC content. This tendency indicates that synonymous codon usage bias is uniformly working toward offsetting the variation in GC content. Meanwhile, all plots of maximum and minimum values were aligned in a row within a narrow band for each. Therefore, I considered that the optimal range of amino acid composition of the proteome is relatively limited and that organisms use this GC offset function to meet the range conditions.

Synonymous codons are part of the genetic code table. Therefore, if synonymous codons and their usage bias have a GC offset function to maintain the optimal amino acid composition, it must be considered a fundamental function of the genetic code table assignment.

Keywords: synonymous codon, codon usage bias, amino acid composition, GC content

E-mail: esumi@clnc.uoeh-u.ac.jp

The authors declare no conflicts of interest associated with this manuscript.

Background

Codon usage bias is the preferential or non-random synonymous codon usage among species. A recent review concluded that their biases are a complex phenomenon influenced by numerous factors, including genome composition, GC content, expression level, length of genes, and recombination rates [1]. In this paper, I present a new plot chart and show a more straightforward explanation of the primary function of the synonymous codon usage and its bias.

Materials and methods

First, I calculated each protein's amino acid compositions and its gene's nucleotide compositions from the publicly available proteome cording sequence (CDS) data set of 23 different bacteria [2-24]. Next, I calculated the maximum and minimum GC contents of the possible gene variations of the amino acid composition of each protein. And then, they were plotted together by their actual GC content on a scatter plot. Eventually, 81,237 proteins from 23 different bacteria were plotted.

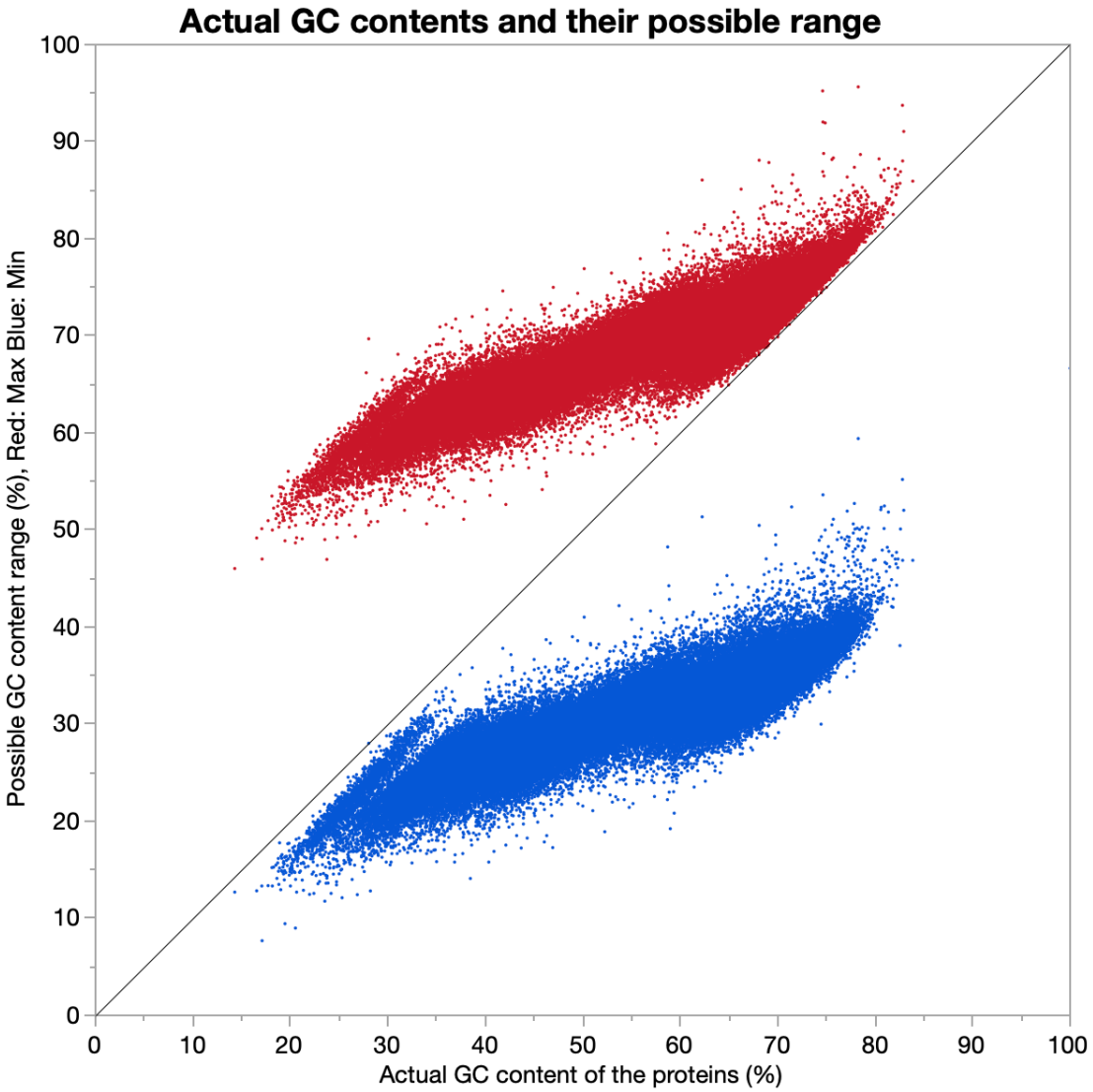
I selected the 23 bacteria by referring to the list of reference proteomes for the "Quest For Orthologs" project on the EMBL-EBI website to avoid the influence of selection bias [25]. Since this list does not include genetic information, the datasets of the bacteria of the same species were downloaded from the NCBI website and used [2-24].
I used Microsoft® Excel for Mac Ver16.62 (Microsoft Corporation, USA) for composition calculations and JMP® 16.2.0 (SAS Institute Inc. USA) for making the scatter plot.

Results

The scatter plot of the possible GC range (max and min) and the actual GC content of the gene is shown as a figure.
The plot showed a clear tendency. Proteins with lower actual GC content genes are coded for by genes closer to the minimum possible GC content. On the other hand, proteins with higher actual GC content genes are coded for by genes closer to the maximum possible GC content.

Figure

The scatter plot of the possible GC range (max and min) and the actual GC content of the gene



The scatter plot of the range of possible GC content of each protein gene (y-axis) and its actual GC content (x-axis). The upper red dots indicate maximum GC content, and the lower blues indicate minimum. The slashed line on the chart represents the actual GC contents, which lie between the maximum and minimum values. There is a clear tendency for proteins with lower actual GC content to lie closer to the minimum and proteins with higher actual GC content to lie closer to the maximum.

Discussion

The plot showed a clear tendency. Describing this tendency in another way, "Genes with higher GC content dominantly use higher GC content codons among the synonymous codons.", "Genes with lower GC content dominantly use lower GC content codons among the synonymous ones.". Therefore, I considered that the tendency of synonymous codon usage bias is uniformly working toward offsetting the variation in GC content.
Meanwhile, all plots of maximum and minimum values were aligned in a row within a narrow band for each. Therefore, I considered that the optimal range of amino acid composition of the proteome is relatively limited and that organisms use this GC offset function to meet the range conditions.

On examining CDS data set, I also found that the variations of GC content are relatively small in some bacteria species. If the primary purpose of codon usage bias is to offset GC content variation, organisms with such small GC content variation would have less necessity to adjust their codon usage biases for the GC content variations. It might be possible for them to save intracellular energy and physical resources by concentrating on using a particular synonymous codon and suppressing others. Therefore, through this discussion, I speculated that this cost reduction with these synonymous codon selections might be the origin of bacterial codon usage bias.

In this study, I made a rather new explanation about codon usage bias with a unique chart. But what was unique in my research? Here I make some consideration about it. Previous studies examined how organisms choose a codon among several synonymous codons [1]. But codon usage bias is a part of the translation function of the genetic code table. Therefore in this study, I investigated the translation function itself, in other words, the relationship between input nucleic acids and output amino acids. For this investigation, there are two possible approaches. One is to examine all the possible output amino acid compositions from a given nucleic acid composition and compare them with its actual output composition. Another is to examine all the possible input nucleic acid compositions from a given output amino acid composition and compare them with its actual input composition. The calculation amount of the former would exponentially increase according to the increased amino acid residue numbers. On the other hand, I could easily calculate the latter when only examining the maximum and minimum possible GC content values. That is how I could take this new approach and reached this finding.

Conclusion

In this paper, I showed that synonymous codon usage and its bias in the bacterial proteomes primarily offset GC content variation to maintain optimal amino acid compositions.
Synonymous codons are part of the genetic code table. Therefore, if synonymous codons and their usage bias have a GC offset function to maintain the optimal amino acid composition, it must be considered a fundamental function of the genetic code table assignment.

Reference

1. Parvathy, S. T., Udayasuriyan, V., & Bhadana, V. (2022). Codon usage bias. Molecular Biology Reports, 49(1), 539–565. https://doi.org/10.1007/s11033-021-06749-4
2. Genome dataset, "Fusobacterium nucleatum subsp. nucleatum", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_003019295.1/
3. Genome dataset, "Mycoplasma genitalium G37", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000027325.1/
4. Genome dataset, "Dictyoglomus turgidum DSM 6724", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000021645.1/
5. Genome dataset, "Thermodesulfovibrio yellowstonii DSM 11347", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000020985.1/
6. Genome dataset, "Leptospira interrogans", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_001569005.1/
7. Genome dataset, "Helicobacter pylori 26695", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000307795.1/
8. Genome dataset, "Chlamydia trachomatis D/UW-3/CX", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000008725.1/
9. Genome dataset, "Bacteroides thetaiotaomicron", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_014131755.1/
10. Genome dataset, "Aquifex aeolicus VF5", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000008625.1/
11. Genome dataset, "Bacillus subtilis subsp. subtilis str. 168", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000009045.1/
12. Genome dataset, "Thermotoga maritima MSB8", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000230655.2/
13. Genome dataset, "Synechocystis sp. PCC 6803", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000009725.1/
14. Genome dataset, "Escherichia coli str. K-12 substr. MG1655", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000005845.2/
15. Genome dataset, "Neisseria meningitidis MC58", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000008805.1/
16. Genome dataset, "Chloroflexus aurantiacus J-10-fl", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000018865.1/
17. Genome dataset, "Rhodopirellula baltica SH 1", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000196115.1/
18. Genome dataset, "Geobacter sulfurreducens PCA", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000007985.2/
19. Genome dataset, "Gloeobacter violaceus PCC 7421", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000011385.1/
20. Genome dataset, "Bradyrhizobium diazoefficiens USDA 110", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000011365.1/
21. Genome dataset, "Mycobacterium tuberculosis H37Rv", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000195955.2/
22. Genome dataset, "Pseudomonas aeruginosa PAO1", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000006765.1/
23. Genome dataset, "Deinococcus radiodurans ATCC 13939", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_020546685.1/
24. Genome dataset, "Streptomyces coelicolor A3(2)", NCBI website, https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_008931305.1/
25. "Reference proteomes - Primary proteome sets for the Quest For Orthologs", EMBL-EBI website. https://www.ebi.ac.uk/reference_proteomes/