Collagens achieve highly biased amino acid composition not only by random mutation but also by slanted assignment of the genetic code table

Genshiro Esumi

Pediatric Surgery, University of Occupational and Environmental Health, Kitakyushu, Japan

Abstract

Collagens are proteins that are ubiquitous in animal bodies. They have unique triple-helix domains mainly consisting of glycines, prolines, and hydroxyprolines, resulting in collagens being highly biased amino acid compositions. Traditionally, we have considered that only the accumulation of random mutations develops the structures and compositions of proteins. However, the results of the present study suggest that it is not only random mutations but also the slanted assignment of the genetic code table that assisted in forming the biased amino acid composition of collagens.

In my previous paper, I showed that the synonymous codon usage selections in all proteins of various 23 bacteria species primarily offset the influences of guanine and cytosine (GC) content variation on their amino acid compositions. In this report, I did the same analysis on all the human proteins and found two things. First, the human proteins' coding sequences have a broader GC content range than each bacteria species. Second, most human proteins' synonymous codon usage selections offset the influences of GC content variation like bacteria, but those of some proteins like collagens did not. Instead, synonymous codon usage selections of these proteins emphasized the feature of their high GC content.

These findings suggest that the slanted genetic code table assignment assisted animals in forming collagens, highly biased amino acid composition proteins, and assisted us be multicellular organisms in our evolutions.

Keywords: collagen, amino acid composition, GC content, synonymous codon, genetic code table

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

*The author has no conflicts of interest relevant to the content of this article.

Introduction

Collagens are proteins that are ubiquitous in animal bodies. They have unique triple-helix domains mainly consisting of glycines, prolines, and hydroxyprolines, resulting in collagens being highly biased and characteristic amino acid compositions. Traditionally, we have considered that only the accumulation of random mutations develops the structures and compositions of proteins. However, the results of the present study suggest that it is not only random mutations but also the slanted assignment of the genetic code table that assisted in forming the biased amino acid composition of collagens.

Materials and methods

In this paper, I examined the publicly available protein coding sequence (CDS) dataset of humans (Homo sapiens) [1] with a scatter plot used in my previous paper [2]. First, from a total of 123411 proteins on the list, I excluded proteins with unusual initiation codons other than ATG (807 proteins) and proteins whose sequences did not end with the normal termination codons TGA TAG and TAA (1238 proteins) for their data cleaning. Next, following the approach I took in the previous paper [2], I calculated each protein's amino acid compositions and its corresponding nucleotide compositions from their CDSs. Then I calculated the maximum and minimum GC contents they could take from their amino acid compositions. Finally, I plotted them with their actual GC content on a scatter plot. Eventually, I plotted a total of 121731 proteins. Then I compared these plots with my previous paper [2].

I used Microsoft® Excel for Mac v16.61 (Microsoft Corporation, Redmond, WA, USA) for composition calculations and JMP® 16.2.0 (SAS Institute Inc., Chicago, IL, USA) for creating the scatter plot.

Results

GC content ranges

Figures show the scatter plots of the possible GC range (maximum and minimum) with the actual GC contents and their distributions (Figures 1 and 2). Figure 1 shows those of human proteins, and Figure 2 shows those of various 23 bacteria species from the previous paper [2]. The human protein-coding sequences have a broader range of GC contents than each bacteria species.

Figure 1 Human protein plots and their GC content distribution

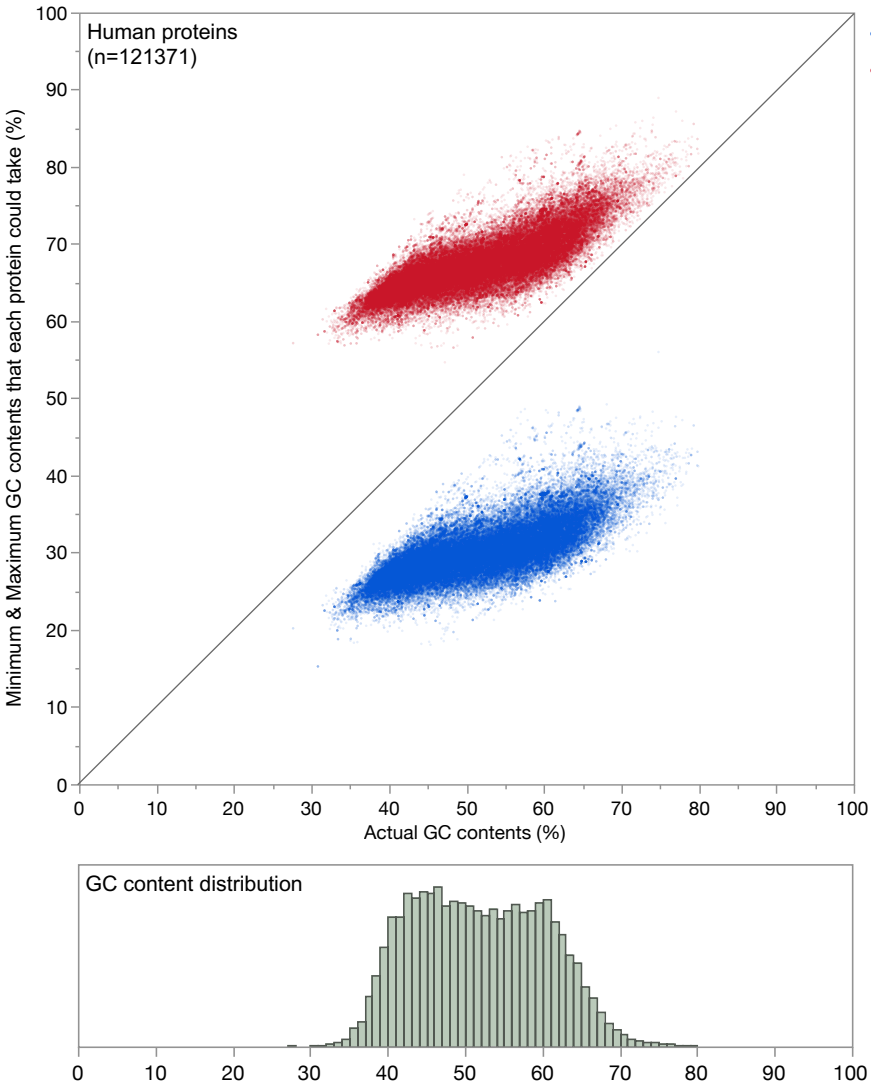


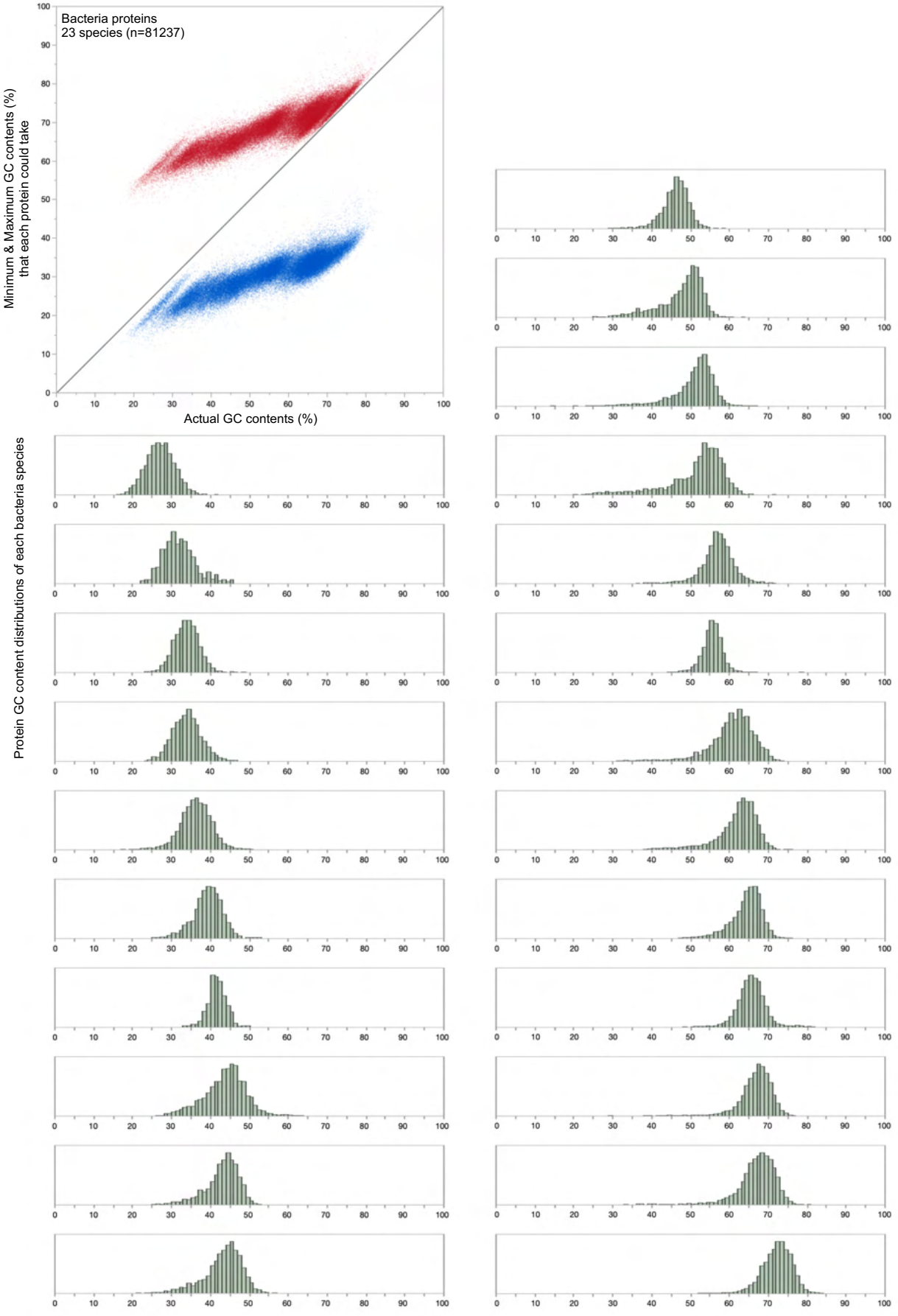
Figure 1

Figure 1 shows human proteins (n=121731). The upper graph shows the scatter plots of the possible GC content range (minimum and maximum) and the actual GC contents, and the lower shows their distribution.

Figure 2

Figure 2 shows various 23 bacteria species proteins (n=81237) [2]. The upper shows the scatter plots of the possible GC content range (minimum and maximum) and the actual GC contents of all their proteins in their genomes, and the lowers show each distribution. Each of them is narrower than that of humans. In other words, human distribution has a broader range than any other bacteria species listed in our previous examination [2].

Figure 2 Bacteria protein plots and their GC content distributions



Comparison of human and bacterial protein plots

Figures show the scatter plots of the possible GC range (maximum and minimum) and the actual GC contents (Figures 3a-d). Figures 3a and 3b are identical to Figures 1 and 2, respectively, except for adding two rectangles placed diagonally to indicate the areas where these plots are most concentrated. Since the sizes and locations of the rectangles in each figure are the same, the areas of greatest plot concentration were considered common among the figures of humans and bacteria.

In Figure 3c, I highlighted collagen proteins and found that many collagen plots are outside the crowded rectangles.

In Figure 3d, I added plots of the type I collagen complex. Type I collagen is a complex protein synthesized by combining two α1 chains and one α2 chain with the propeptides at both ends removed. The type I collagen complex showed an extremely outlier position from all other proteins.

Figure 3a-d Scatter plots of Human proteins, Bacteria proteins, and Human collagens

Figure 3a

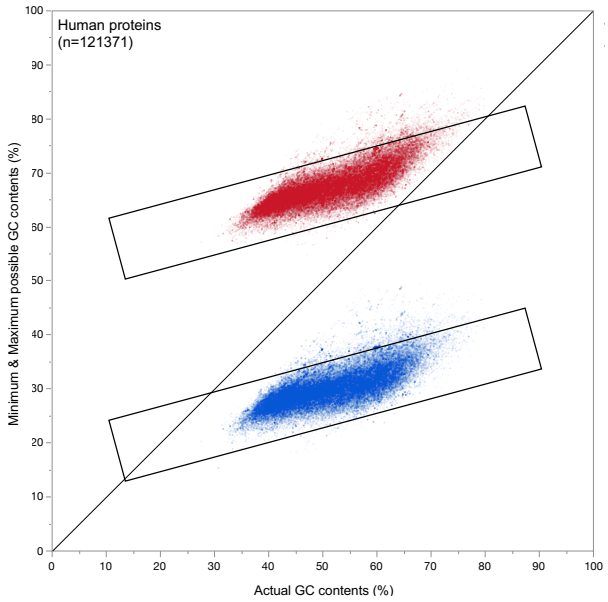


Figure 3b

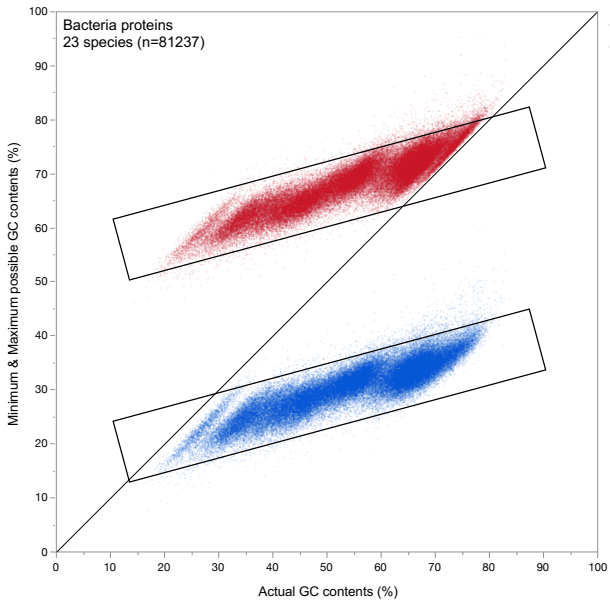


Figure 3c

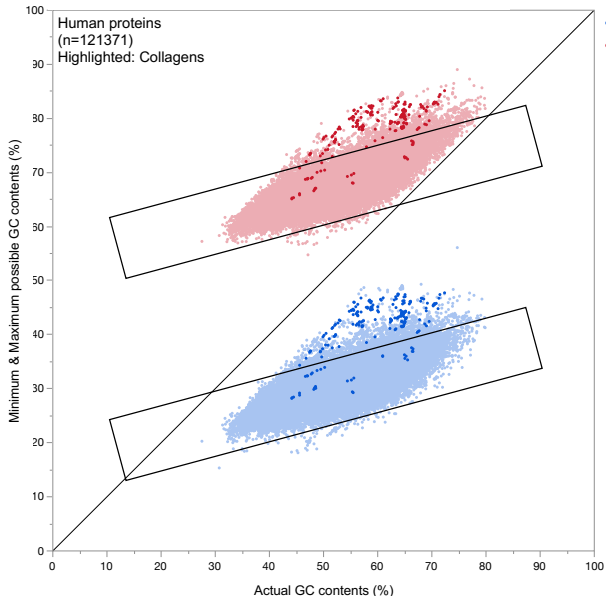


Figure 3d

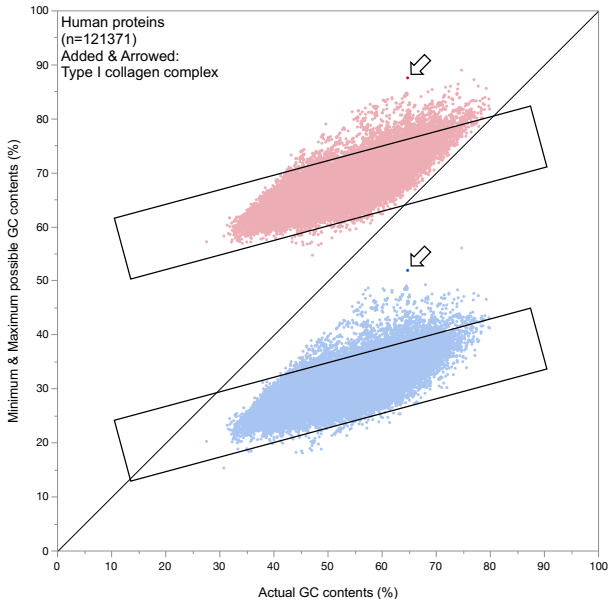


Figure 3a-d

Figures show the scatter plots of the possible GC ranges (maximum and minimum), calculated from their amino acid compositions, and their actual GC contents (Figures 3a-d). Figures 3a and 3b are identical to Figures 1 and 2, except for adding two identical rectangles placed diagonally to indicate the areas where these plots are most concentrated. In Figure 3c, I highlighted collagen proteins and found that most collagen plots are outside the other proteins' concentrated area. In Figure 3d, I added plots of the type I collagen complex (arrowed). The type I collagen complex showed an outlier position from all other proteins.

Discussion

GC offset rule

In the previous paper, I showed a basic rule: bacteria proteins encoded by genes with lower GC content use lower GC codons, and proteins encoded by genes with higher GC content use higher GC codons among their synonymous codon variations [2]. This rule indicates that the bacteria proteins select their synonymous codons in a direction that offsets their GC content. This paper calls this rule "the GC offset rule".

Collagens do not follow the GC offset rule.

In the current examination of human proteins, I found that most proteins follow the GC offset rule, but some proteins, such as collagens, do not follow it (Figure 3c). Inversely, collagens seem to have used the GC offset rule in the opposite direction and enhanced the feature of their high GC content gene. This finding could explain how collagens achieve their highly biased amino acid compositions.

Why do almost all proteins follow the GC offset rule?

A previous paper on organisms' amino acid composition distribution showed that each organism's distance distribution from their mean amino acid composition is relatively narrow [3]. However, the genome GC contents vary widely among species (Figure 2). From a general perspective, significant differences in GC content will inevitably affect the amino acid composition. Therefore, organisms cannot choose synonymous codons dispersedly, for they must keep their relatively narrow amino acid composition distributions. As a result, in most proteins of various organisms, their GC contents primarily determine their selection of synonymous codons, and they have no choice but to follow the GC offset rule.

How could collagens break the GC offset rule?

Here we discuss the conditions for organisms to break the GC offset rule and synthesize proteins like collagens significantly:

1. The organisms must have genes with higher GC content.
2. The genes must use lower GC synonymous codons.
3. The translation systems corresponding to the lower GC synonymous codons (of lower GC content genes) efficiently translate these higher GC content genes.

I assumed that only in that case will the amino acid composition characteristics of the high GC content genes be further emphasized. So, the broad GC content distribution of the genes in the organism may be the first condition for the ability to synthesize proteins by utilizing the CG offset rule in the reverse direction. Consequently, the broader distribution of GC content in humans might not be coincidental but inevitable.

In humans, contrary to collagens, some proteins emphasized features of low GC content. These were mainly G protein-coupled receptor proteins, but their extents were lesser than collagens [data not shown].

How did the humans achieve broader GC range distribution of their genes?

In the current examination, the calculated minimum possible GC content of the collagen complex protein (52.0%) is much higher than the average GC content of the entire human genome (40.5%) [1]. If the human genome consisted of uniform GC content, all protein-coding sequences would have a uniform GC content. And proteins with characteristic amino acid compositions such as collagens would not be synthesized. Meanwhile, vertebrates, including humans, have islands of high GC content regions in the genome sequences, called isochores [4]. These isochores could explain how humans keep broader distributions of GC contents in their genes. In the current examination, the human distribution of GC content in the genes showed two peaks at around 40% and around 60%. The former peak can be explained by the average GC of the entire human genome (40.5%), while the latter peak could be explained by assuming that they are coded in the isochore regions. In the literature, isochores have a higher density of genes with more rapid rates of evolution [5,6]. Maybe these are because isochore genes have phenotypic advantages and are increasing by selection. Besides, isochore genes might be in the process of evolution, for they are relatively new. Therefore, I speculated that animals, including humans, acquired the genomic isochore region for synthesizing collagens and utilizing them as the main components of their bodies.

How the amino acid composition of type I collagen was formed:

A possible explanation from the genetic code

In type I collagen complex protein, glycine is the most abundant amino acid in the amino acid composition at 33%, followed by proline (including hydroxyproline) at 22%. These two amino acids are essential components of the triple helix domain of collagens [7,8]. However, the reason why the next most abundant is alanine at 11%, followed by arginine at 5%, has never been directly explained. Meanwhile, if you check the standard genetic code table, you will see that all those four amino acids perfectly correspond to the four amino acids with the highest GC content codons. Therefore, since collagens use codons that emphasize their high GC content feature, these findings might indicate that the genetic code assignment could explain these four amino acids being collagen's most abundant compositions. Perhaps the amino acid composition of collagen was already determined when the current standard genetic code table appeared in the history of life evolution.

Are there any other proteins like collagens?

In the current human protein analysis, elastin, loricrin, some mucins, and some basic salivary proteins also showed similar plot positions of collagens. This finding could indicate that collagens and these proteins achieved their amino acid composition characters by breaking the GC offset rule, which also

explains why they have similar amino acid compositions. Additional analyses for other organisms showed that spider silk threads and silkworm silk fibers showed similar plot positions [9, data not shown]. Furthermore, the added analyses of the bacterial proteins in the previous paper [2] showed that the PE_PGRS family proteins in Mycobacterium tuberculosis were the only protein group that appeared to fall into this category. However, the extent was lesser than those of humans and other eukaryotic cellular organisms [data not shown].

Conclusion

My findings in this report suggest that the slanted genetic code table assignment assisted animals in forming collagens, highly biased amino acid composition proteins, and assisted us be multicellular organisms in our evolutions. This kind of protein synthesis assistance could be a function of the standard genetic code table assignment.

Reference

1. National Center for Biotechnology Information (NCBI). (2022). Genome assembly GRCh38.p14. National Library of Medicine (NIH) website. https://www.ncbi.nlm.nih.gov/data-hub/genome/GCF_000001405.40/
2. Esumi, G. (2022). Synonymous codon usage and its bias in the bacterial proteomes primarily offset GC content variation to maintain optimal amino acid compositions. Jxiv. https://doi.org/10.51094/jxiv.99
3. Esumi, G. (2022). Proteome and cellular amino acid compositions may be mutually constrained and in a state of narrow convergence. Jxiv. https://doi.org/10.51094/jxiv.95
4. Costantini, M., Cammarano, R., & Bernardi, G. (2009). The evolution of isochore patterns in vertebrate genomes. *BMC Genomics, 10*, 146. https://doi.org/10.1186/1471-2164-10-146
5. Bernardi, G. (1993). The vertebrate genome: isochores and evolution. *Molecular Biology and Evolution*, *10*(1), 186–204. https://doi.org/10.1093/oxfordjournals.molbev.a039994
6. Kiktev, D. A., Sheng, Z., Lobachev, K. S., & Petes, T. D. (2018). GC content elevates mutation and recombination rates in the yeast Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(30), E7109–E7118. https://doi.org/10.1073/pnas.1807334115
7. UniProt consortium. (2022). P02452 · CO1A1_HUMAN. Entry version 250. UniProtKB website. https://www.uniprot.org/uniprotkb/P02452/entry
8. UniProt consortium. (2022). P08123 · CO1A2_HUMAN. Entry version 237. UniProtKB website. https://www.uniprot.org/uniprotkb/P08123/entry
9. Zhang, Y., Shimizu, K., Shiomi, K., Kajiura, Z., & Nakagaki, M. (2008). cDNA cloning of *Nephila clavata* dragline silk (MaSp1) gene and comparison with the sequence of *Bombyx mori* fibroin heavy chain. *Sanshi-Konchu Biotec*, *77*(1), 39–46. https://doi.org/10.11416/konchubiotec.77.1_39