The Synonymous Codon Usage of a Protein Gene Is Primarily Determined by the Guanine + Cytosine Content of the Individual Gene Rather Than the Species to Which It Belongs To Synthesize Proteins With a Balanced Amino Acid Composition

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

Abstract

In the genetic code, most amino acids have multiple corresponding codons, and codons corresponding to the same amino acid are called synonymous codons. The use of synonymous codons in protein genes is known to be biased rather than random, and such bias is often explained by species differences in the guanine + cytosine (GC) content of their genomes, or in the abundance of tRNAs that intervene between codons and amino acids in the translation process.

In this study, I statistically analyzed the synonymous codon usage in protein genes of the proteomes of 79 species from 3 domains, published as Reference Proteome, and found that the GC content of the individual gene, rather than the species to which it belongs, primarily determines its synonymous codon usage.

Why then does the GC content of the individual gene determine its codon usage selection? Some papers have already mentioned that the GC content of the third letters of codons is even higher in genes with high total GC content and even lower in genes with low total GC content, and these were usually explained by evolutionary pressure on their genomic GC content and its subsequent shift. However, while this explanation explained the behavior of the third letter of the codon, it did not explain the behavior of the first and second letters. To provide a new explanation for the overall behavior of synonymous codon usage, I added an analysis. Since in previous work the amino acid composition distributions of organisms appeared to be in a state of narrow convergence, and since we know that most organisms share some highly conserved proteins across species boundaries, an additional analysis, based on the assumption that the organism maintains a proteome close to the amino acid composition of a particular conserved protein, suggests that this codon selection counteracts the effect of the GC content of the gene on the amino acid composition of the protein and behaves in the direction of counterbalancing and maintaining a constant and balanced amino acid composition.

From the results of this study, I concluded that synonymous codon selection in a protein gene primarily counterbalances its GC content to maintain a balanced amino acid composition for the proteome. The ability to generate proteins with balanced amino acid composition from genes with different ranges of GC content is considered to be one of the basic functions achieved by the genetic code itself.

Keywords: codon usage bias, standard genetic code, GC content, amino acid composition, proteome

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

1. Background

In the genetic code, most amino acids have multiple corresponding codons, and codons corresponding to the same amino acid are called synonymous codons. The use of synonymous codons in protein genes is known to be biased rather than random, and this bias is called "codon usage bias". Since the codon usage bias in bacterial proteomes mainly differs between species, these biases are often explained by species differences in the guanine + cytosine (GC) content of their genomes, or in the abundance of tRNAs that intervene between codons and amino acids in the translation process. [1, 2]

In a previous report, I showed the possibility that codon usage bias in bacterial proteomes primarily counterbalances the GC content of their individual genes [3], so in the current report, I performed more detailed statistical analyses not only on bacterial proteomes, but on the proteomes of all three domains, archaea, bacteria, and eukaryotes.

2. Materials and Methods

In this study, I used protein genes from a protein dataset published on the Internet as "reference proteomes" consisting of more than one million protein entries [4]. This reference proteome dataset has the amino acid sequence data of each protein and the corresponding gene nucleotide sequences, but because I found that their gene sequences did not completely match their amino acid sequences, I first excluded protein genes that did not match the length of the protein amino acid sequence length. In this first exclusion, 249,042 proteins out of 1,023,125 proteins are excluded, leaving 857,750 proteins. Then, I excluded protein genes whose amino acid sequences did not match those in the UniProt database [5], whose amino acid sequence data had missing or abnormal alphabetic codes, whose nucleotide sequence data had missing or abnormal descriptions, and whose start and stop codons did not follow known deviations [6]. Finally, protein genes with mismatches to amino acid sequences other than the start codon were excluded. As a result, 774,083 protein genes in 79 species proteomes of 3 domains were included in this analysis [Table 1, Supplemental Data.csv].

Since the purpose of this analysis was to analyze how the bias in codon usage behaves, I first counted the number of each codon in each gene, and then calculated the codon composition by gene, by corresponding amino acid. Then, I performed a principal component analysis on all the synonymous codon compositions calculated, and then examined the results.

For the additional analysis mentioned in the Discussion, I calculated the average amino acid composition of both the total protein genes and the proteome of each organism as representative central amino acid compositions of the total proteins and each proteome, respectively, and then calculated the L1 distances (Manhattan distances) of each protein amino acid composition from each of the two central compositions.

Before calculating L1 distances of protein amino acid compositions, each amino acid composition of protein genes is generated by counting each amino acid residue number on each gene and dividing it by the sum of all 20 amino acids on the gene, resulting in each amino acid composition taking values between 0 and 1 and the total sum being 1.

The L1 distances between the amino acid compositions of protein x and protein y were calculated using the following equation, where x_i represents each amino acid composition of protein x. It is the sum of all 20 amino acid composition differences between these two proteins, resulting in a value between 0 and 2, the lower the closer.

$$\text{L1 distance}(x, y) = \sum_{i=1}^{20} |x_i - y_i|$$

In this study, I used Microsoft® Excel for Mac v16.79.1 (Microsoft Corporation, Redmond, WA, USA) to generate compositions and other computational results. I also used JMP® 17.2.0 (SAS Institute Inc., Chicago, IL, USA) to perform principal component analysis and to generate graphs and figures.

Table 1.

Taxonomy ID	Domain	Organism Name	Listed Proteins	Target Proteins	Proportions
64091	Archaea	Halobacterium salinarum	2423	2332	0.96244325
69014	Archaea	Thermococcus kodakarensis	2301	2298	0.99869622
188937	Archaea	Methanosarcina acetivorans	4468	4344	0.97224709
243232	Archaea	Methanocaldococcus jannaschii	1787	1666	0.93228875
273057	Archaea	Saccharolobus solfataricus	2937	2871	0.97752809
374847	Archaea	Korarchaeum cryptofilum	1602	1601	0.99937578
436308	Archaea	Nitrosopumilus maritimus	1795	1795	1
83332	Bacteria	Mycobacterium tuberculosis	3995	3821	0.95644556
83333	Bacteria	Escherichia coli	4403	4323	0.98183057
85962	Bacteria	Helicobacter pylori	1554	1501	0.96589447
100226	Bacteria	Streptomyces coelicolor	8035	7967	0.99153703
122586	Bacteria	Neisseria meningitidis serogroup B	2001	1971	0.9850075
189518	Bacteria	Leptospira interrogans serogroup Icterohaemorrhagiae serovar Lai	3676	3645	0.99156692
190304	Bacteria	Fusobacterium nucleatum subsp. nucleatum	2046	2022	0.98826979
208964	Bacteria	Pseudomonas aeruginosa	5564	5532	0.99424874
224308	Bacteria	Bacillus subtilis	4260	4201	0.98615023
224324	Bacteria	Aquifex aeolicus	1553	1529	0.98454604
224911	Bacteria	Bradyrhizobium diazoefficiens	8253	8191	0.99248758
226186	Bacteria	Bacteroides thetaiotaomicron	4782	4768	0.99707235
243090	Bacteria	Rhodopirellula baltica	7271	7194	0.98940998
243230	Bacteria	Deinococcus radiodurans	3084	2946	0.95525292
243231	Bacteria	Geobacter sulfurreducens	3402	3387	0.99559083
243273	Bacteria	Mycoplasma genitalium	483	132	0.27329193
243274	Bacteria	Thermotoga maritima	1852	1818	0.98164147
251221	Bacteria	Gloeobacter violaceus	4406	4385	0.99523377
272561	Bacteria	Chlamydia trachomatis	895	882	0.98547486
289376	Bacteria	Thermodesulfovibrio yellowstonii	1982	1977	0.9974773
324602	Bacteria	Chloroflexus aurantiacus	3850	3846	0.99896104
515635	Bacteria	Dictyoglomus turgidum	1743	1737	0.99655766
1111708	Bacteria	Synechocystis sp.	3507	3407	0.9714856
3055	Eukaryota	Chlamydomonas reinhardtii	17614	17533	0.99540139
3218	Eukaryota	Physcomitrium patens	31359	29631	0.9448962
3702	Eukaryota	Arabidopsis thaliana	27481	26051	0.94796405
4577	Eukaryota	Zea mays	39225	36352	0.9267559
5664	Eukaryota	Leishmania major	8038	8031	0.99912914
5888	Eukaryota	Paramecium tetraurelia	39461	98	0.00248346
6239	Eukaryota	Caenorhabditis elegans	19827	18885	0.95248903
6412	Eukaryota	Helobdella robusta	23328	19395	0.83140432
6945	Eukaryota	Ixodes scapularis	20496	11496	0.56088993
7070	Eukaryota	Tribolium castaneum	16568	16362	0.98756639
7165	Eukaryota	Anopheles gambiae	13016	2187	0.16802397
7227	Eukaryota	Drosophila melanogaster	13821	12954	0.93726937
7719	Eukaryota	Ciona intestinalis	16680	8343	0.50017986
7739	Eukaryota	Branchiostoma floridae	26627	25265	0.94884891
7918	Eukaryota	Lepisosteus oculatus	18321	10547	0.57567818
7955	Eukaryota	Danio rerio	26249	20683	0.78795383
8090	Eukaryota	Oryzias latipes	23617	21181	0.89685396
8355	Eukaryota	Xenopus laevis	35860	31829	0.88759063
8364	Eukaryota	Xenopus tropicalis	22229	20072	0.9029646
9031	Eukaryota	Gallus gallus	18369	1998	0.10877021
9595	Eukaryota	Gorilla gorilla gorilla	21783	19329	0.88734334
9598	Eukaryota	Pan troglodytes	23051	21334	0.92551299
9606	Eukaryota	Homo sapiens	20586	935	0.04541922
9615	Eukaryota	Canis lupus familiaris	20972	3957	0.18868014
9913	Eukaryota	Bos taurus	23841	17832	0.7479552
10090	Eukaryota	Mus musculus	21957	4724	0.21514779
10116	Eukaryota	Rattus norvegicus	22870	9162	0.40061216
13616	Eukaryota	Monodelphis domestica	21223	8094	0.38137869
35128	Eukaryota	Thalassiosira pseudonana	11717	9246	0.78910984
36329	Eukaryota	Plasmodium falciparum	5372	5361	0.99795235
39947	Eukaryota	Oryza sativa subsp. japonica	43672	35690	0.81722843
44689	Eukaryota	Dictyostelium discoideum	12726	12384	0.97312588
45351	Eukaryota	Nematostella vectensis	24427	13315	0.54509354
81824	Eukaryota	Monosiga brevicollis	9188	8289	0.90215498
164328	Eukaryota	Phytophthora ramorum	15349	13304	0.86676656
184922	Eukaryota	Giardia intestinalis	4900	4896	0.99918367
214684	Eukaryota	Cryptococcus neoformans var. neoformans serotype D	6604	6515	0.98652332
237561	Eukaryota	Candida albicans	6035	2029	0.33620547
237631	Eukaryota	Ustilago maydis	6788	6723	0.99042428
284591	Eukaryota	Yarrowia lipolytica	6449	6415	0.99472786
284812	Eukaryota	Schizosaccharomyces pombe	5122	5051	0.98613823
321614	Eukaryota	Phaeosphaeria nodorum	15998	15891	0.99331166
330879	Eukaryota	Aspergillus fumigatus	9647	9518	0.98662797
367110	Eukaryota	Neurospora crassa	9759	9673	0.99118762
412133	Eukaryota	Trichomonas vaginalis	50190	43646	0.86961546
418459	Eukaryota	Puccinia graminis f. sp. tritici	15688	15488	0.9872514
559292	Eukaryota	Saccharomyces cerevisiae	6060	6006	0.99108911
665079	Eukaryota	Sclerotinia sclerotiorum	14445	14384	0.99577709
684364	Eukaryota	Batrachochytrium dendrobatidis	8610	7910	0.91869919
Total			1023125	774083	0.75658693

This table shows the total number of protein entries listed in the reference proteomes and the number of target proteins selected for this analysis [4].

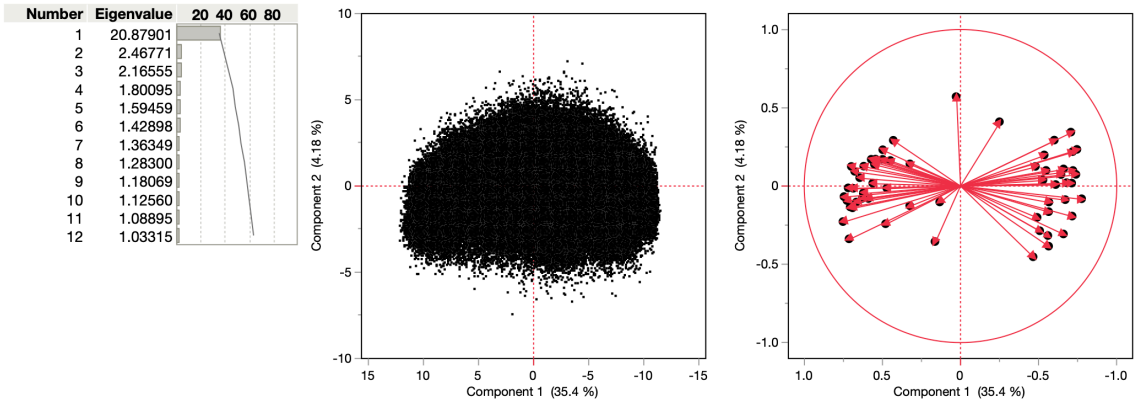
For organisms with known deviations from the standard genetic code (IDs 243273, 5888, 237561), the majority of genes were excluded due to their deviations. In addition, in some proteomes of multicellular prokaryotic organisms, the majority of genes were excluded due to their sequence length mismatch (reasons unknown).

As a result, about 75% of the total protein entries were included.

3. Results

The results of the principal component analysis of the synonymous codon compositions are presented [Figure 1]. The contribution of the first principal component was very large with 35.4% compared to 4.18% for the second principal component. In addition, most codons were bipolarized in the eigenvector of the first principal component [Figure 1].

Figure 1.

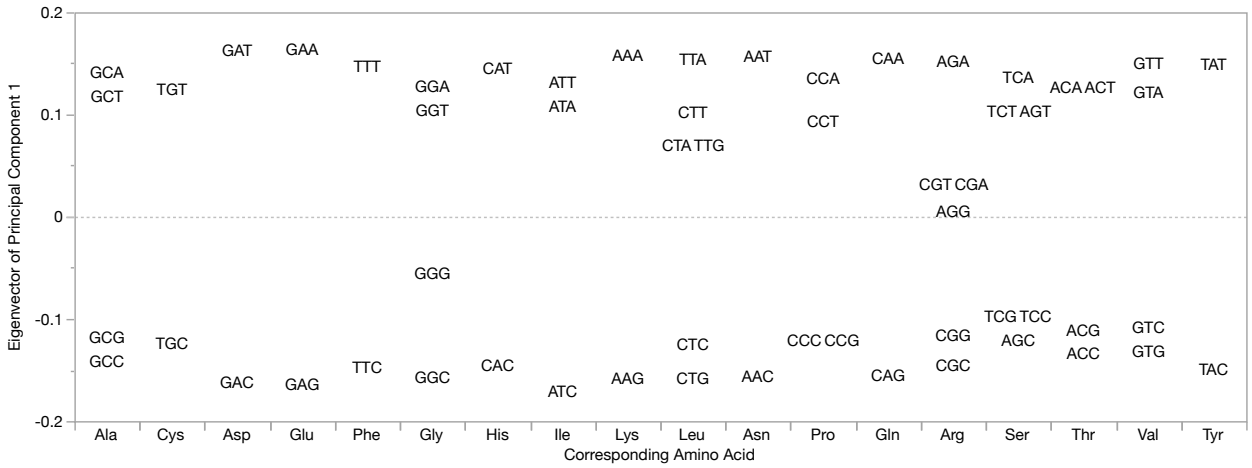


This figure shows the results of the principal component analysis of the synonymous codon composition, with the eigenvalues of the principal components on the left, the protein plot of the first and second principal components in the middle, and the eigenvector plot of the first and second principal components on the right.

The contribution of the first principal component (proportional to the eigenvalues) was very large at 35.4% compared to 4.18% for the second principal component (middle). In addition, most codons were bipolarized in the eigenvector of the first principal component (right).

To analyze the structure of bipolarization found in Figure 1, I compared the eigenvalues of the first principal component of the synonymous codons of each amino acid and found that the bipolar split occurred between synonymous codons with low GC content (the same with high TA content) and those with high GC content. Furthermore, those that were not at either extreme were all those with intermediate GC content when three or more synonymous codons were present for an amino acid [Figure 2].

Figure 2.



This figure shows each eigenvector value of the first principal component of the synonymous codons by each amino acid.

The upper and lower extremes of each amino acid codon were uniformly composed of those with lower and higher GC content of the synonymous codons, respectively, and the codons not at either extreme were uniformly composed of codons with intermediate GC content when there were three or more synonymous codons.

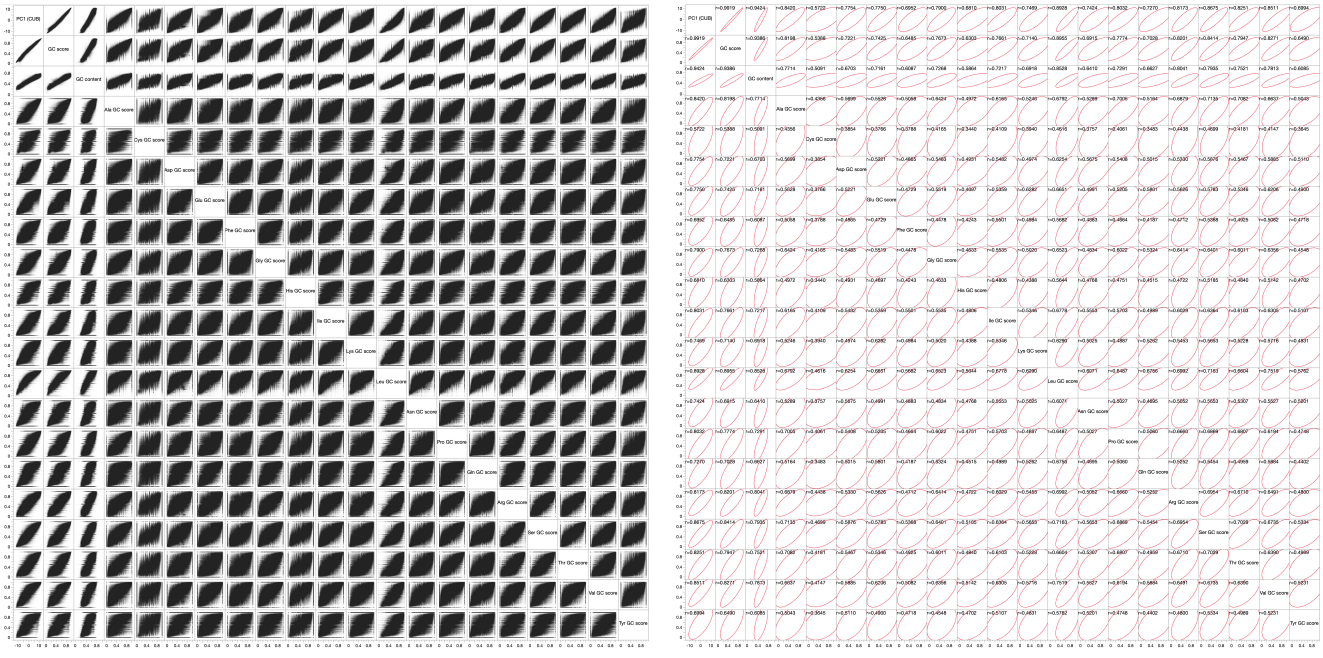
The above uniform bipolarization of codons suggested that the first principal component of synonymous codon composition was related to the GC content of synonymous codons. To verify this, I next evaluated the character of bipolarized GC-rich and GC-poor synonymous codons in the first principal component by introducing the value of "GC score". Specifically, I first defined "GC content" as the proportion of the number of guanine and cytosine in the gene sequence, and it takes values from 0 to 1. Then, I defined "max GC" and "min GC" as the GC content of amino acids on the target protein when only the most GC-rich and most GC-poor synonymous codons were used, respectively. Finally, the following equation was used to calculate the GC score.

$$\text{GC score} = \frac{\text{GC content} - \text{min GC}}{\text{max GC} - \text{min GC}}$$

Consequently, this GC score takes values from 0 to 1, and a higher GC score means that the synonymous codons with higher GC content are used in the protein gene, and vice versa. And because this GC score describes the behavior of synonymous codon usage in terms of GC content, it could also be used as an index for codon usage selection.

Comparing the GC score with the first principal component of synonymous codon composition, I found that they correlated with each other with a correlation coefficient of 0.99 [Figure 3]. They also correlated with the GC content of each protein gene with correlation coefficients of 0.94 and 0.93, respectively [Figure 3]. In addition, by calculating the GC score for each amino acid in each protein independently, I found that they were all positively correlated with each other with correlation coefficients of 0.34 - 0.75 [Figure 3].

Figure 3.



These figures show the correlations between the individual amino acid GC scores calculated for all target proteins in this study. On the left are scatter plots of their first principal components of codon usage, GC score, GC content, and individual GC scores by amino acid of all target proteins, and on the right are their density ellipses ($\alpha = 0.90$) and their correlation coefficients.

The GC score correlated with the first principal component of synonymous codon composition with a correlation coefficient of 0.99. These two also correlated with their GC content with correlation coefficients of 0.94 and 0.93, respectively. In addition, their independently calculated GC scores for each amino acid were all positively correlated with each other, with correlation coefficients ranging from 0.34 to 0.75.

4. Discussions

In the results of this study, the first principal component of synonymous codon usage composition correlated with the GC score, a newly defined index of GC content balance in synonymous codon usage, with a correlation coefficient of 0.99, so that these two were considered almost identical [Figure 3]. At the same time, the eigenvectors of the first principal component were bipolarized into those with high GC content and those with low GC content in each synonymous codon triplet, and those in the middle of the two extremes were those with intermediate GC content when three or more synonymous codons were present [Figure 2]. In addition, I found that not only was the GC score correlated with GC content, but the individual GC scores calculated independently for each amino acid were not only positively correlated with total GC content, but were all positively correlated with each other (correlation coefficient 0.34 - 0.75) [Figure 3]. Because the correlation coefficients between individual amino acid GC scores and gene GC content are smaller than that between GC score and GC content, I speculate that the background of these correlations is likely due to selection pressure on overall amino acid composition rather than selection pressure on individual codon usage of each amino acid.

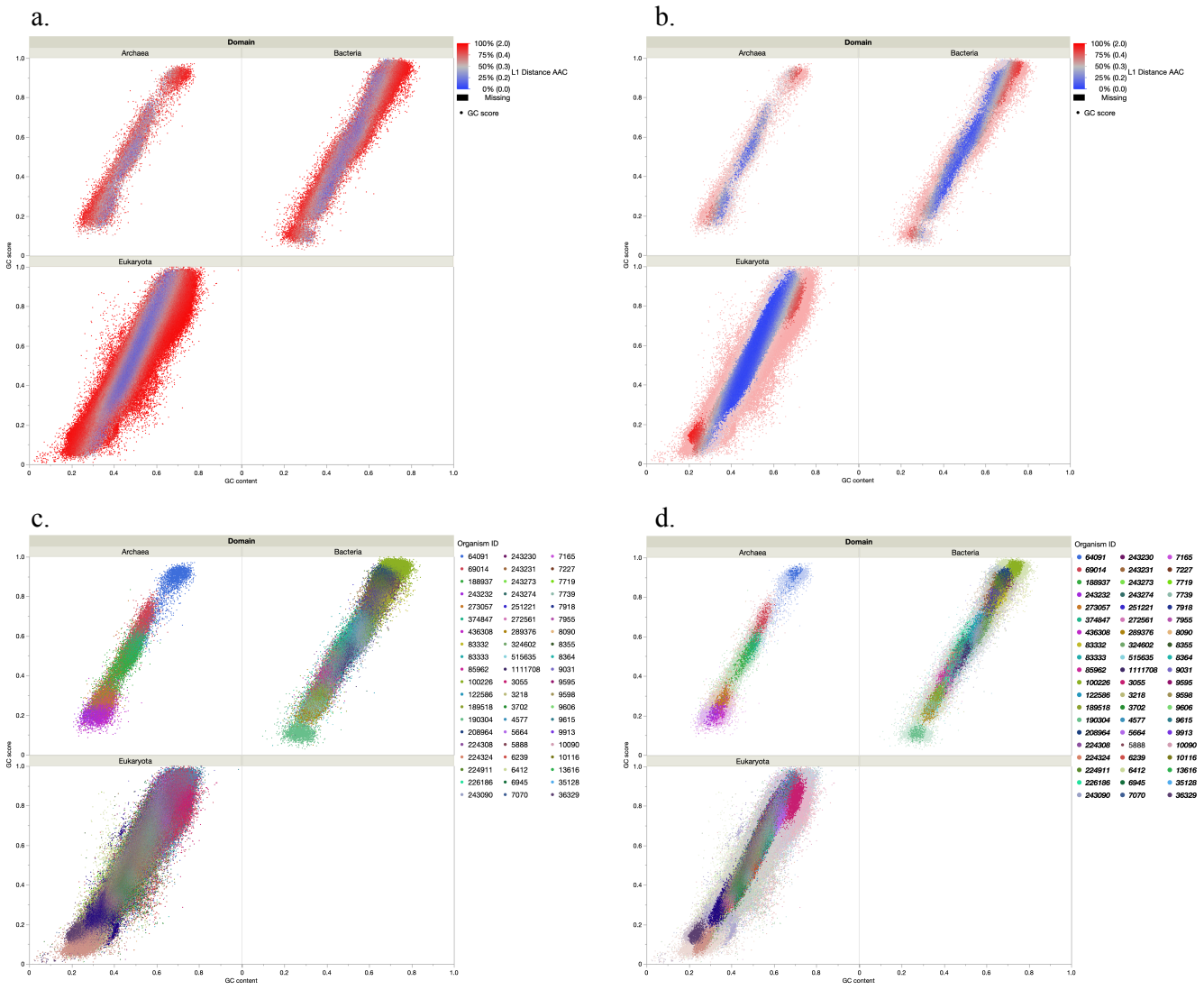
Then why does the GC content of the individual gene determine its codon usage selection? Some papers have already mentioned that the GC content of the third letters of codons is even higher in genes with high total GC content and even lower in genes with low total GC content, and these were usually explained by evolutionary pressure on their genomic GC content and its subsequent shift [1]. However, while this explanation explained the behavior of the third letter of the codon, it did not adequately explain the behavior of the first and second letters.

To provide a new explanation for the overall behavior of synonymous codon usage, I added an analysis of amino acid composition distances between proteins. Since in previous work the amino acid composition distributions of each organism's proteome appeared to be in a state of narrow convergence [7, 8], and since we know that most organisms share some highly conserved proteins across species boundaries, I speculated that the centers of amino acid composition of the proteomes of all organisms are relatively close to each other. And if this is true, then organisms must derive their amino acid composition from genes on their genomes with different ranges of GC content. Therefore, in this additional analysis, I examined where and how the proteins near the central amino acid composition of total protein and the proteins near the central amino acid composition of each organism's proteome would lie on a scatter plot of their GC score and their GC content.

For the additional analysis mentioned above, I used the L1 distance, called the Manhattan distance, as a metric to measure the distance of each protein's amino acid composition from both the central composition of all target proteins and the central composition of each species' proteome. First, I made a scatter plot of all protein genes by their GC score and their GC content, and then I displayed their distance from the central composition of all target proteins. And in the next sheet, I further highlighted only those proteins that were relatively close to the center of each organism's proteome. These plots of protein genes were also colored according to the species to which they belonged [Figure 4].

As a result of the current additional study, I found that protein genes close to the central amino acid composition of all proteins were not only distributed over a fairly wide range of GC content, but also that they were almost lined up in a narrow belt [Figure 4]. And I also found that proteins close to the center of each organism were also relatively close to the overall center [Figure 4]. From these results, I concluded that the strong correlation between GC score and GC content was because the GC score, an index of synonymous codon usage, is used to balance the GC content of their gene to synthesize interspecies relatively conserved amino acid compositions of proteome proteins from genes with a wide range of GC content.

Figure 4.



These figures are the scatter plots of the GC scores and GC contents of all target protein genes by the domains to which they belong. Top left (a) shows the plots colored by their distance from the central, i.e. average, amino acid composition of all proteins, and top right (b) shows the same plots but highlighting only proteins that are close to each central amino acid composition of each organism, with distances less than 0.2. Bottom left (c) and right (d) are the same as above, but colored according to the species to which they belong. All distributions of amino acid composition and all distributions of their distances are shown in supplementary figures [Supplementary Figure 1a.pdf, Supplementary Figure 1b.pdf, Supplementary Figure 2.pdf].

In Figure (a), protein genes close to the central amino acid composition of the total target proteins were not only distributed over a rather wide range of GC content, but also almost lined up in a narrow belt. Figure (b) shows that protein genes close to the individual centers of each organism were also relatively close to the overall center (colored almost in blue).

In figures (c) and (d), the plots of the protein genes of each organism showed a somewhat broad but linear correlation with their GC content. If the species were the determinant of their codon usage bias, their GC scores should be correlated with the species, not with their GC content. In the current result, I found that they correlated with their GC content, leading me to conclude that individual GC content itself is the primary determinant of their codon usage bias.

Deviations from the correlation of GC score with GC content were overwhelmingly observed in eukaryotes compared to archaea and bacteria, but the many deviating proteins were also found to contain a non-trivial amount of intrinsically disordered regions [Supplementary Figure 3.pdf]. In a previous report, I showed that the structure of the genetic code itself is designed to convert gene sequences with less thymine into intrinsically disordered regions [9]. Since the amount of thymine used to determine amino acid composition is thought to be controlled by the GC content and GC score of the gene, it seemed reasonable that deviations in GC score would be associated with intrinsically disordered regions. This is also thought to be consistent with a report that codons encoding intrinsically disordered regions often deviate from the codon usage bias of the organism [10].

Finally, there have been several reports that codon usage bias is determined by the GC content of each organism's genome [1, 11, 12, 13]. However, there has been no report that the GC content of "individual genes" is the most dominant determinant of codon usage bias, and this is the first report to state this.

5. Conclusion

From the results of this study, I concluded that synonymous codon selection in a protein gene primarily counterbalances its individual GC content to maintain a balanced amino acid composition of the proteome. The ability to generate proteins with functionally balanced amino acid composition from genes with different ranges of GC content is considered to be one of the basic functions achieved by the standard genetic code itself. And perhaps this explains one of the reasons why the genetic code has degenerated.

6. References

1. Hershberg, R., & Petrov, D. A. (2009). General Rules for Optimal Codon Choice. In M. W. Nachman (Ed.), PLoS Genetics (Vol. 5, Issue 7, p. e1000556). *Public Library of Science* (PLoS). https://doi.org/10.1371/journal.pgen.1000556
2. Dong, H., Nilsson, L., & Kurland, C. G. (1996). Co-variation of tRNA Abundance and Codon Usage inEscherichia coliat Different Growth Rates. In *Journal of Molecular Biology* (Vol. 260, Issue 5, pp. 649–663). Elsevier BV. https://doi.org/10.1006/jmbi.1996.0428
3. 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
4. "Quest for Orthologs" group. (2023) Reference proteomes - Primary proteome sets for the Quest For Orthologs, RELEASE 2023_03. https://www.ebi.ac.uk/reference_proteomes/ Accessed 1 Sep 2023
5. Bateman, A., Martin, M.-J., Orchard, S., Magrane, M., Ahmad, S., Alpi, E., Bowler-Barnett, E. H., Britto, R., Bye-A-Jee, H., Cukura, A., Denny, P., Dogan, T., Ebenezer, T., Fan, J., Garmiri, P., da Costa Gonzales, L. J., Hatton-Ellis, E., Hussein, A., … Zhang, J. (2022). UniProt: the Universal Protein Knowledgebase in 2023. In *Nucleic Acids Research* (Vol. 51, Issue D1, pp. D523–D531). Oxford University Press (OUP). https://doi.org/10.1093/nar/gkac1052
6. Hamashima, K., & Kanai, A. (2013). Alternative genetic code for amino acids and transfer RNA revisited. In *BioMolecular Concepts* (Vol. 4, Issue 3, pp. 309–318). Walter de Gruyter GmbH. https://doi.org/10.1515/bmc-2013-0002
7. 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
8. Esumi, G. (2023). The Distributions of Amino Acid Compositions of Proteins in an Organism's Proteome Uniformly Approximate Binomial Distributions. *Jxiv*. https://doi.org/10.51094/jxiv.408
9. Esumi, G. (2023). The standard genetic code is designed to generate transmembrane domains and intrinsically disordered regions as projections of the thymine density on the gene. *Jxiv*. https://doi.org/10.51094/jxiv.533
10. Homma, K., Noguchi, T., & Fukuchi, S. (2016). Codon usage is less optimized in eukaryotic gene segments encoding intrinsically disordered regions than in those encoding structural domains. In *Nucleic Acids Research* (Vol. 44, Issue 21, pp. 10051-10061). Oxford University Press (OUP). https://doi.org/10.1093/nar/gkw899
11. Masłowska-Górnicz, A., van den Bosch, M. R. M., Saccenti, E., & Suarez-Diez, M. (2022). A large-scale analysis of codon usage bias in 4868 bacterial genomes shows association of codon adaptation index with GC content, protein functional domains and bacterial phenotypes. In *Biochimica et Biophysica Acta* (BBA) - Gene Regulatory Mechanisms (Vol. 1865, Issue 6, p. 194826). Elsevier BV. https://doi.org/10.1016/j.bbagrm.2022.194826
12. Behura, S. K., & Severson, D. W. (2012). Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. In *Biological Reviews* (Vol. 88, Issue 1, pp. 49–61). Wiley. https://doi.org/10.1111/j.1469-185x.2012.00242.x
13. Parvathy, S. T., Udayasuriyan, V., & Bhadana, V. (2021). Codon usage bias. In *Molecular Biology Reports* (Vol. 49, Issue 1, pp. 539–565). Springer Science and Business Media LLC. https://doi.org/10.1007/s11033-021-06749-4