Statistical Extremes of Amino Acid Residue Composition of the Proteome Proteins Can Explain the Origin of the Universality of the Genetic Code

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

Abstract

Organisms have evolved and diverged from a common ancestor, and today there are many different species in many different environments. Because these organisms share a nearly identical genetic code, it is believed that all species have changed little in their genetic code from that of the ancestor over the course of evolution. However, the reasons for this universality, why almost all organisms have never changed their genetic code, are not well understood.

In the present study, principal component analyses of the amino acid residue composition of proteome proteins from different species revealed that proteins with high amounts of transmembrane domains (TMDs) and proteins with high amounts of intrinsically disordered regions (IDRs) almost universally occupy the two extremes of each proteome plot of their first and second principal components. These TMD- and IDR-rich proteins correlated not only with the amino acid composition of the proteins, but also with the nucleic acid composition of their corresponding genes.

In my previous report, I showed that the genetic code itself has a structure that can assist the generation of TMDs and IDRs by exploiting the partial biases of nucleic acid composition in gene sequences. With the current statistical analyses, I also showed that TMD- and IDR-rich proteins always occupy the statistical extremes of amino acid composition in the proteomes of different organisms. If TMDs and IDRs are always the two largest domains/regions with extreme amino acid composition in the proteome, and if the genetic code has a structure that helps synthesize TMDs and IDRs, then I can conclude that the structure of the current genetic code may have been chosen to meet the requirements of the typical amino acid composition of these functional domains. If this assumption is true, it would be reasonable to assume that such a genetic code has become universal.

This is a new explanation for the universality of the genetic code, and I call it "The Optimized Translation Theory". This theory should provide a partial explanation for the origin of the standard genetic code in terms of its functions.

Keywords: amino acid composition, transmembrane domain, intrinsically disordered region, genetic code, optimized translation theory, Chargaff’s second parity rule, GC content, TA skew, GC skew

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

1. Background

All living organisms synthesize proteins by translating the nucleic acid sequences of genes into the amino acid sequences of proteins according to the genetic code. This genetic code is known to be common to nearly all organisms and is called the standard genetic code (SGC) [1]. It was originally thought that the genetic code used by the common ancestors of existing organisms was frozen and carried over to the present [1], but subsequent analysis has shown that there are several organisms that deviate from the SGC [2]. However, even if there are deviations, it is believed that most organisms have not changed their genetic code over the course of evolution. But there is no standard explanation for why almost all organisms have never changed their genetic code, and why even in organisms that have changed their genetic code, the changes from the standard genetic code are so small [2, 3, 4].

In my previous analyses, I showed that the SGC have the structure to assist the synthesis of transmembrane domains (TMDs) and intrinsically disordered regions (IDRs) of proteins as a projection of the bias in the partial nucleic acid composition of gene sequences [5]. In another analysis, I showed that TMD-rich proteins and IDR-rich proteins appear at the extremes of the first and second principal component plots of the amino acid residue composition of protein sequences in the human proteome [6]. Hypothetically, if TMD- and IDR-rich proteins always appear at the extremes of their principal component plots in all organisms, then the genetic code itself could be configured to encode their extreme amino acid compositions.

Therefore, in the first part of this study, I performed principal component analyses of the amino acid residue composition of proteome proteins from the proteome datasets of various organisms published as "reference proteomes" [7] to examine where the TMD- and IDR-rich proteins would be located in each proteome by coloring each protein plot on their principal component plots.

In the second part of this study, I examined the correlations between the principal components of the amino acid composition of the proteins and the nucleic acid composition of their genes by coloring the PCA plots with their nucleic acid composition indices.

In the third and final part of this study, I examined the correlations between the nucleic acid composition of the genes and the TMD and IDR fractions of the proteins by examining the color distributions of TMD-rich and IDR-rich proteins on plots of their nucleic acid composition.

2. Materials and Methods

For this study, a list of proteins published by the European Bioinformatics Institute as "reference proteomes" [7] was used. The original dataset contained 1,023,125 protein entries from 79 species representing all three domains of life. For the amino acid composition analyses, 4,121 protein entries were excluded due to discrepancies with the UniProtKB database or uncertain or unusual information in their sequence or annotation, leaving 1,019,004 proteins for the target [5, 8]. And for the subsequent gene composition analysis, proteins with mismatches in nucleic acid sequence length or codon variations that did not match known genetic code variations were excluded, leaving 826,386 protein genes for examination [2, 9].

The amino acid residue composition of each protein subjected to principal component analysis was calculated using the accompanying data of the "reference proteomes" provided in FASTA format. The amino acid residue composition was calculated by first counting the amino acid residues for each protein and then dividing the number of amino acid residues by the total number of residues. As a result, each amino acid composition took a value between 0 and 1, and the sum for each protein was 1.

In the PCA of this study, the data distribution was first standardized and then analyzed by PCA. Although PCA determines the values of each principal component, its polarity is not absolute. Therefore, in the results of this analysis, for the sake of consistency in the listing, I chose these polarities as those in which the average of the principal component values of the proteins with the high amount of TMDs is higher than the average of the other proteins in the same proteome.

The TMD and IDR fractions of each protein were calculated using the annotation information in UniProtKB. The ratios of residues in the TMDs and IDRs to the total length of the amino acid residue sequence of each protein were calculated and set as the TMD fraction and IDR fraction, respectively [8].

In the PCA plots of the results of this paper, the proteins in the proteome of each organism were plotted by species, with the first and second principal components calculated in each principal component analysis performed for each species. In the first examination, these plots were colored by the TMD and IDR fractions in red and blue, respectively, and these two sheets were superimposed to create a single plot sheet containing both TMD and IDR information. In the second examination, these plots were colored by the nucleic acid composition indices of their genes, GC content, TA skew, and GC skew. Each index was calculated using the following equations.

$$\text{GC content} = \frac{G + C}{T + A + G + C}, \quad \text{TA skew} = \frac{T - A}{T + A}, \quad \text{GC skew} = \frac{G - C}{G + C}.$$

In these equations, the capital letters T, A, G, and C represent the number of thymine, adenine, guanine, and cytosine, respectively, in the nucleic acid sequence of each gene. And, unless otherwise noted, the nucleic acid sequences of genes in this paper have been set so that they do not contain stop codons that do not encode amino acids.

In the later nucleic acid composition plots of the results, the proteins in the proteome of each organism were again plotted by species, but with their combination of gene GC content and TA skew, and their GC skew and TA skew. In this investigation, these plots were again colored by the TMD and IDR fractions in red and blue, respectively, and these two sheets were again superimposed to create a single plot sheet containing both TMD and IDR information.

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.

3. Results

3.1. Principal Component Analyses of Each Proteome Proteins

The amino acid composition of each protein generated from the "reference proteomes" dataset [7] was subjected to principal component analysis for each organism. The number of registered proteins of each organism in the "reference proteomes" and the number of proteins subjected to amino acid composition analysis and those subjected to nucleic acid composition analysis in this study are listed in Table 1 for each organism [Table 1].

Figure 1 plots each protein by organism using the first and second principal components calculated by principal component analysis for each organism, and in this figure the proteins are color-coded according to the domain to which each organism belongs [Figure 1]. The contribution of the first principal component ranged from 11.5 to 19.7% (mean 14.4%), and the contribution of the second principal component ranged from 8.6 to 14.7% (mean 11.0%). (Data not shown)

Table 1. Number of target proteins in this study

Taxonomy ID	Domain	Organism Name	Listed Proteins	AAC Targets	NAC Targets
64091	Archaea	Halobacterium salinarum	2423	2423	2332
69014	Archaea	Thermococcus kodakarensis	2301	2301	2298
188937	Archaea	Methanosarcina acetivorans	4468	4456	4344
243232	Archaea	Methanocaldococcus jannaschii	1787	1774	1667
273057	Archaea	Saccharolobus solfataricus	2937	2936	2871
374847	Archaea	Korarchaeum cryptofilum	1602	1602	1601
436308	Archaea	Nitrosopumilus maritimus	1795	1795	1795
83332	Bacteria	Mycobacterium tuberculosis	3995	3995	3821
83333	Bacteria	Escherichia coli	4403	4393	4324
85962	Bacteria	Helicobacter pylori	1554	1543	1503
100226	Bacteria	Streptomyces coelicolor	8035	8035	7969
122586	Bacteria	Neisseria meningitidis serogroup B	2001	2001	1972
189518	Bacteria	Leptospira interrogans serogroup Icterohaemorrhagiae serovar Lai	3676	3676	3645
190304	Bacteria	Fusobacterium nucleatum subsp. nucleatum	2046	2046	2022
208964	Bacteria	Pseudomonas aeruginosa	5564	5563	5533
224308	Bacteria	Bacillus subtilis	4260	4259	4212
224324	Bacteria	Aquifex aeolicus	1553	1550	1530
224911	Bacteria	Bradyrhizobium diazoefficiens	8253	8253	8192
226186	Bacteria	Bacteroides thetaiotaomicron	4782	4782	4768
243090	Bacteria	Rhodopirellula baltica	7271	7271	7194
243230	Bacteria	Deinococcus radiodurans	3084	3060	2946
243231	Bacteria	Geobacter sulfurreducens	3402	3393	3387
243273	Bacteria	Mycoplasma genitalium	483	483	470
243274	Bacteria	Thermotoga maritima	1852	1851	1818
251221	Bacteria	Gloeobacter violaceus	4406	4406	4385
272561	Bacteria	Chlamydia trachomatis	895	895	882
289376	Bacteria	Thermodesulfovibrio yellowstonii	1982	1977	1977
324602	Bacteria	Chloroflexus aurantiacus	3850	3849	3846
515635	Bacteria	Dictyoglomus turgidum	1743	1743	1737
1111708	Bacteria	Synechocystis sp.	3507	3506	3415
3055	Eukaryota	Chlamydomonas reinhardtii	17614	17602	17545
3218	Eukaryota	Physcomitrium patens	31359	31287	29632
3702	Eukaryota	Arabidopsis thaliana	27481	27476	26146
4577	Eukaryota	Zea mays	39225	39198	34382
5664	Eukaryota	Leishmania major	8038	8036	8032
5888	Eukaryota	Paramecium tetraurelia	39461	39456	38989
6239	Eukaryota	Caenorhabditis elegans	19827	19826	18900
6412	Eukaryota	Helobdella robusta	23328	23294	19395
6945	Eukaryota	Ixodes scapularis	20496	20461	11497
7070	Eukaryota	Tribolium castaneum	16568	16552	16367
7165	Eukaryota	Anopheles gambiae	13016	12989	2200
7227	Eukaryota	Drosophila melanogaster	13821	13594	13037
7719	Eukaryota	Ciona intestinalis	16680	16614	8353
7739	Eukaryota	Branchiostoma floridae	26627	26421	25272
7918	Eukaryota	Lepisosteus oculatus	18321	17988	10552
7955	Eukaryota	Danio rerio	26249	26094	24156
8090	Eukaryota	Oryzias latipes	23617	23614	21188
8355	Eukaryota	Xenopus laevis	35860	35595	34538
8364	Eukaryota	Xenopus tropicalis	22229	22104	21315
9031	Eukaryota	Gallus gallus	18369	18337	2078
9595	Eukaryota	Gorilla gorilla gorilla	21783	21493	19340
9598	Eukaryota	Pan troglodytes	23051	22963	21351
9606	Eukaryota	Homo sapiens	20586	20486	1026
9615	Eukaryota	Canis lupus familiaris	20972	20935	3978
9913	Eukaryota	Bos taurus	23841	23798	17948
10090	Eukaryota	Mus musculus	21957	21680	5025
10116	Eukaryota	Rattus norvegicus	22870	22816	9606
13616	Eukaryota	Monodelphis domestica	21223	21084	8242
35128	Eukaryota	Thalassiosira pseudonana	11717	11717	9246
36329	Eukaryota	Plasmodium falciparum	5372	5368	5361
39947	Eukaryota	Oryza sativa subsp. japonica	43672	43586	32746
44689	Eukaryota	Dictyostelium discoideum	12726	12713	12386
45351	Eukaryota	Nematostella vectensis	24427	24322	13316
81824	Eukaryota	Monosiga brevicollis	9188	9177	8313
164328	Eukaryota	Phytophthora ramorum	15349	15284	13304
184922	Eukaryota	Giardia intestinalis	4900	4900	4896
214684	Eukaryota	Cryptococcus neoformans var. neoformans serotype D	6604	6597	6515
237561	Eukaryota	Candida albicans	6035	5984	5900
237631	Eukaryota	Ustilago maydis	6788	6788	6724
284591	Eukaryota	Yarrowia lipolytica	6449	6449	6431
284812	Eukaryota	Schizosaccharomyces pombe	5122	5122	5063
321614	Eukaryota	Phaeosphaeria nodorum	15998	15998	15907
330879	Eukaryota	Aspergillus fumigatus	9647	9647	9537
367110	Eukaryota	Neurospora crassa	9759	9759	9697
412133	Eukaryota	Trichomonas vaginalis	50190	49311	43946
418459	Eukaryota	Puccinia graminis f. sp. tritici	15688	15688	15488
559292	Eukaryota	Saccharomyces cerevisiae	6060	6059	6033
665079	Eukaryota	Sclerotinia sclerotiorum	14445	14445	14421
684364	Eukaryota	Batrachochytrium dendrobatidis	8610	8610	7910
Total			1023125	1019004	826386

Table 1 shows the number of registered proteins for each organism in the 'reference proteomes' dataset, along with the number of proteins analyzed for amino acid composition and nucleic acid composition in this study, these are in 'AAC Targets' and 'NAC Targets' respectively. The color of the data bar corresponds to the domain to which it belongs.

Figure 1. Principal component plots of the proteome proteins of each organism

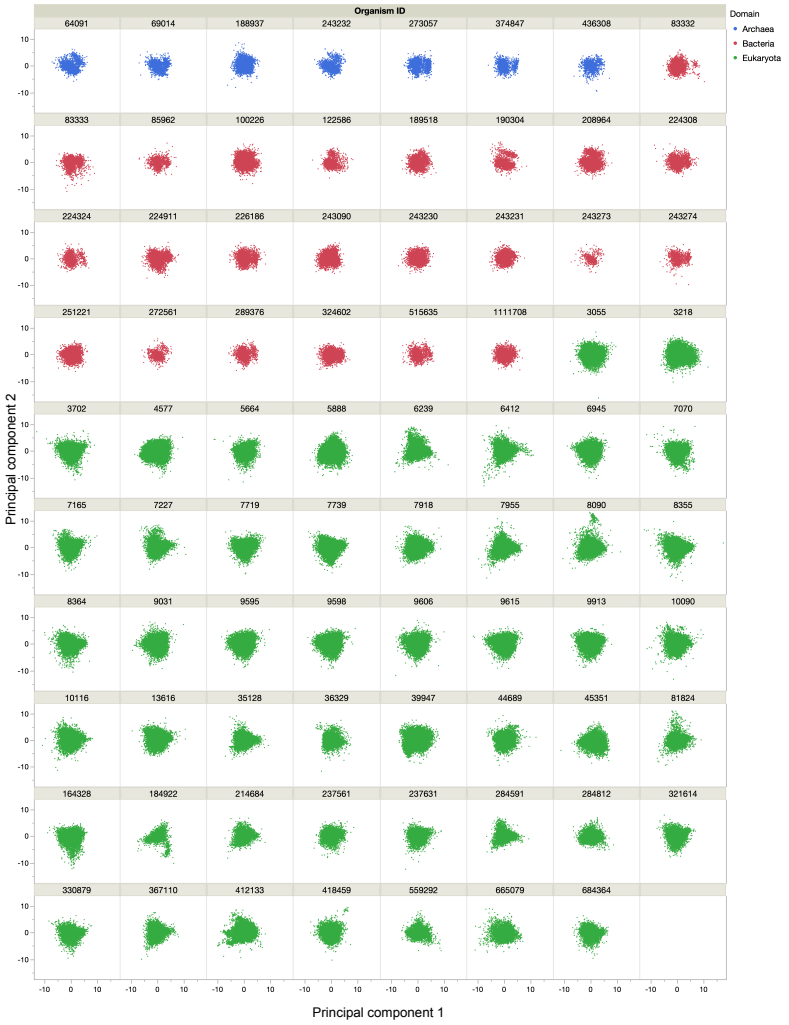


Figure 1 shows scatter plots of each protein by organism using the first and second principal components calculated by principal component analysis for each organism, and in this figure the color of each plot bar corresponds to the domain of the organisms represented. The horizontal axis corresponds to the first principal component and the vertical axis corresponds to the second principal component. Blue represents archaea, red represents bacteria, and green represents eukaryotes.

3.2. TMD and IDR Fractions on Principal Component Plots

Figure 2 plots each protein by organism using the first and second principal components calculated by principal component analysis for each organism, and in this figure the proteins are doubly color-coded, red and blue, according to their TMD and IDR fractions, respectively. Proteins with high amounts of TMDs and proteins with high amounts of IDRs were found to almost universally and oppositely occupy the two extremes of the first and second principal components of each proteome plot [Figure 2]. In some organisms (e.g. organism ID 83332) where these region-rich proteins were not located at either end, the TMD- and IDR-rich proteins were located at both ends of their third principal component. (Data not shown)

Figure 2. PCA plots of the proteome proteins of each organism, colored according to their TMD and IDR fractions

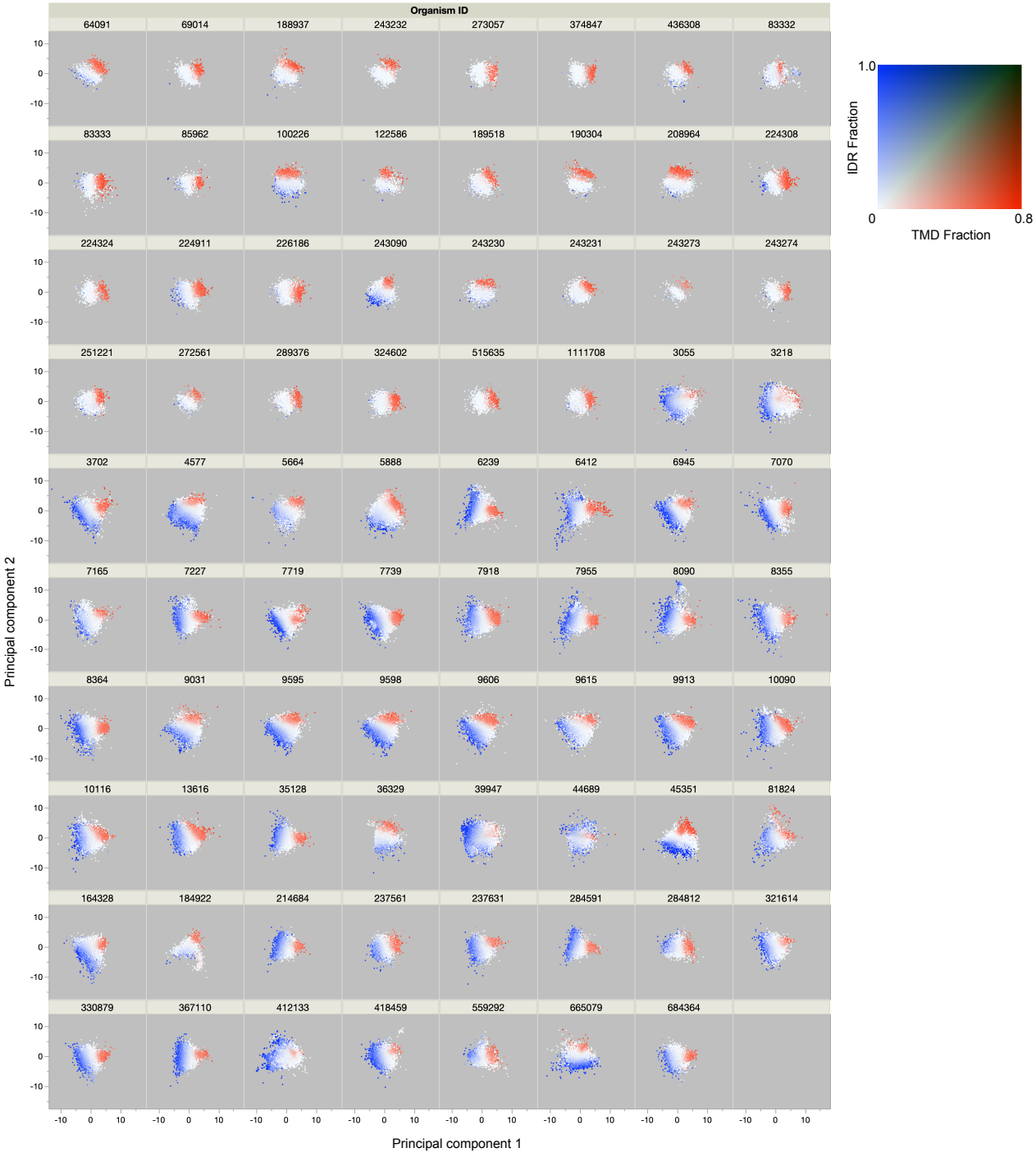


Figure 2 plots each protein by organism using the first and second principal components calculated by principal component analysis for each organism, and in this figure the proteins are doubly color-coded, red and blue, according to their TMD and IDR fractions, respectively. Proteins with high amounts of TMDs and proteins with high amounts of IDRs were found to almost universally and oppositely occupy the two extremes of their first and second principal components of each proteome plot. At the same time, this figure also shows that very few, if any, proteins are rich in both TMD and IDR.

3.3. Gene Nucleic Acid Indices on Principal Component Plots

Figures 3a, b, c plot each protein by organism using the first and second principal components calculated by principal component analysis for each organism, and in this figure the proteins are color-coded according to their gene nucleic acid indices, GC content, TA skew, and GC skew. The color distributions of these PCA plots are not random at all, but all show gradations, indicating that they are all correlated with their gene nucleic acid composition [Figures 3].

In these plots, while all three indices of the four nucleic acid compositions correlated with the PCA plots, the contrast of the gradients for GC skew was weaker than those for TA skew or GC content, even though the plot color of GC skew was on the same color scale as that of TA skew.

Figure 3. PCA Plots of the Proteome Proteins of Each Organism, Colored According to Their Gene Nucleic Acid Indices

Figure 3a. Colored by their GC content.

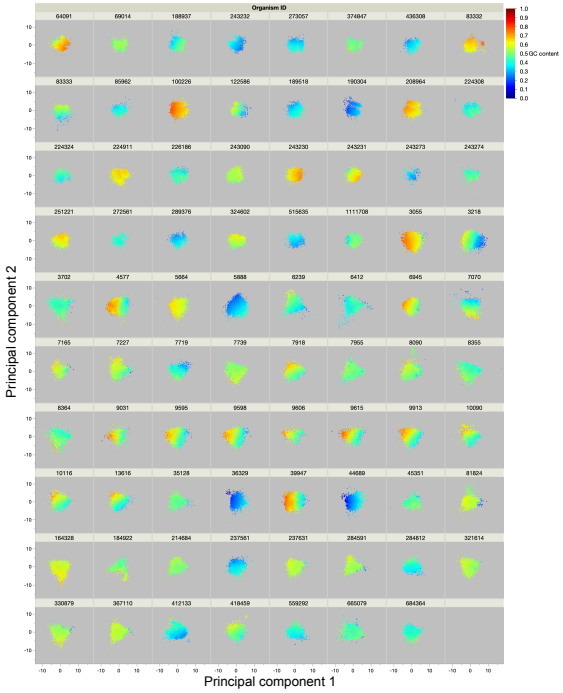


Figure 3b. Colored by their TA skew.

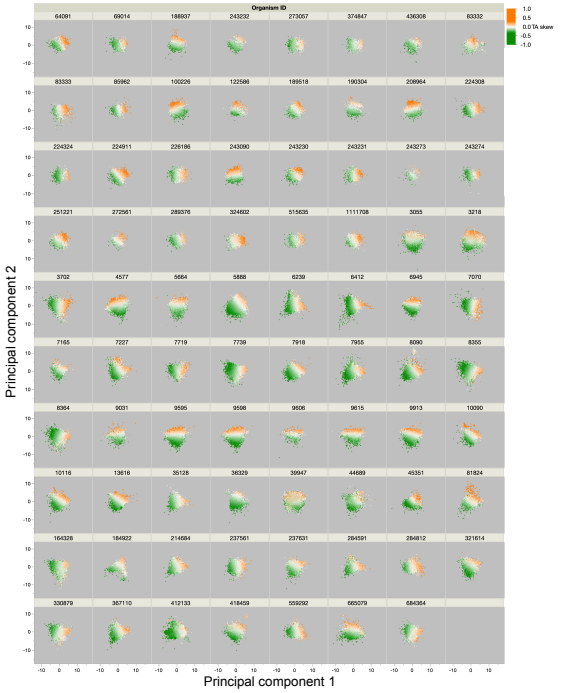
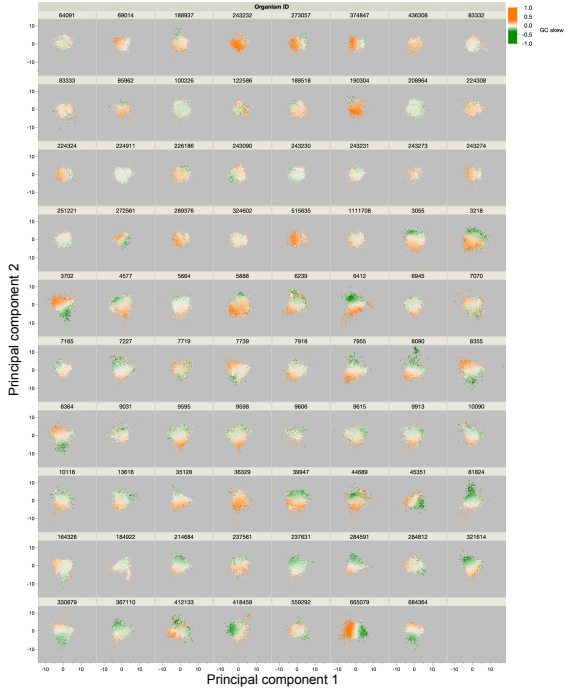


Figure 3c. Colored by their GC skew.



Figures 3a, b, c plot each protein by organism using the first and second principal components calculated by principal component analysis for each organism.

In this figure the proteins are color-coded according to their gene nucleic acid indices, GC content, TA skew, and GC skew. The color distributions of these PCA plots are not random at all, but all show gradations, indicating that they are all correlated with their gene nucleic acid composition.

In these plots, while all three indices of the four nucleic acid compositions correlated with the PCA plots, the contrast of the gradients for GC skew was weaker than those for TA skew or GC content, even though the plot color of GC skew was on the same color scale as that of TA skew.

3.4. TMD and IDR Fractions on Gene Nucleic Acid Index Plots

Figures 4a, b show each protein plotted by organism with its gene nucleic acid composition indices: 4a is a plot with GC content and TA skew, and 4b is a plot with GC skew and TA skew. And in these figures, the proteins are doubly color-coded according to their TMD and IDR fractions, as in Figure 2 [Figures 4]. In Figure 4a, TMD-rich proteins uniformly occupy the region of high TA skew, and IDR-rich proteins uniformly occupy the region of primarily lower TA skew and partially higher GC content. In Figure 4b, GC skew does not seem to correlate with TMD- and IDR-rich proteins compared to TA skew and GC content.

Figure 4. PCA Plots of the Proteome Proteins of Each Organism, Colored According to Their Gene Nucleic Acid Indices

Figure 4a. Plots with GC content and TA skew

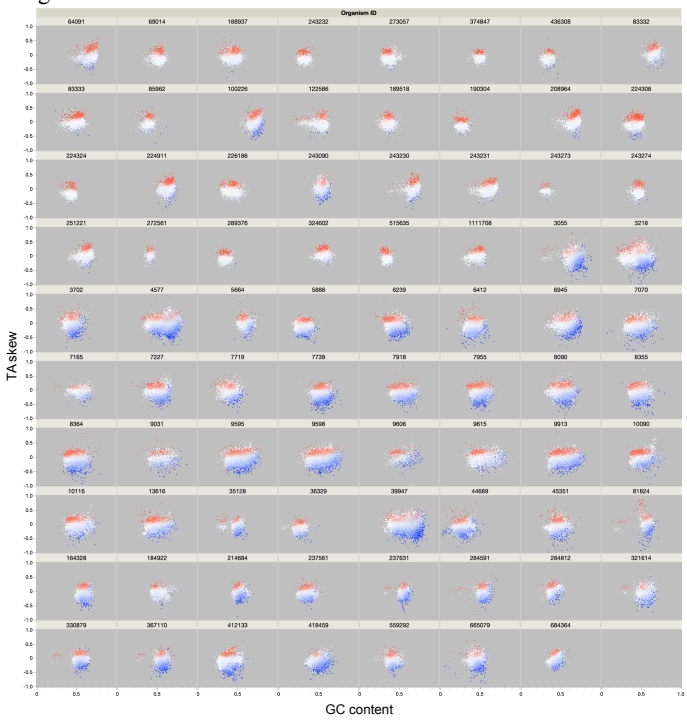
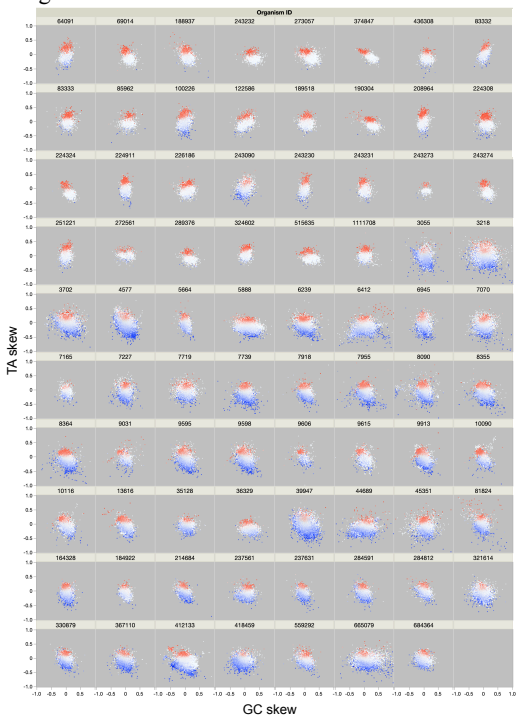
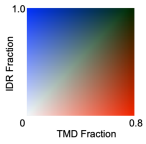


Figure 4b. Plots with GC skew and TA skew





Figures 4a, b show each protein plotted by organism with its gene nucleic acid composition indices: 4a is a plot with GC content and TA skew, and 4b is a plot with GC skew and TA skew. And in these figures, the proteins are doubly color-coded according to their TMD and IDR fractions, as in Figure 2.
In Figure 4a, TMD-rich proteins uniformly occupy the region of high TA skew, and IDR-rich proteins uniformly occupy the region of primarily lower TA skew and partially higher GC content.
In Figure 4b, GC skew does not seem to correlate with TMD- and IDR-rich proteins compared to TA skew and GC content.

4. Discussions

The deciphering of the genetic code began with Nirenberg's seminal discovery that poly-U codes for poly-phenylalanine in 1961 [10]. Subsequent analysis revealed that various organisms share a common genetic code, leading to Crick's "The Frozen Accident Theory" in 1968 [1]. Despite more than 50 years of debate, there is still no definitive explanation for the structure or universality of the genetic code, including why the codon UUU codes for phenylalanine [3, 4].

In my previous report, I showed that the genetic code itself has a structure that can assist the generation of TMDs and IDRs by exploiting the partial biases of nucleic acid composition in gene sequences [5]. From a simple analysis of the genetic code itself, it appears that the genetic code is designed to encode TMDs by dense thymine sequences in the gene and IDRs in regions of sparse thymine sequence [5]. However, upon detailed analysis, most genes uniformly adjust their codon usage bias according to their GC content, counterbalancing the influence of GC content on amino acid composition [9]. As a result, the behavior of TMDs and IDRs has been linked to their TA skew rather than to the density of thymine in the gene sequence, as shown in Figures 4a and 4b.

With the current statistical analyses of this study, I have shown that TMD- and IDR-rich proteins always occupy the statistical extremes of amino acid composition in the proteomes of different organisms. If TMDs and IDRs are always the two largest domains/regions with extreme amino acid composition in the proteome, and if the genetic code has a structure that helps synthesize TMDs and IDRs, then I can conclude that the structure of the current genetic code may have been chosen to meet the requirements of the typical amino acid composition of these functional domains. If this assumption is true, it would be reasonable to assume that such a genetic code has become universal.

Previous research on the origin of the genetic code has focused primarily on the structure of the code itself or the structures of tRNA itself. However, there is a lack of studies that analyze the genetic code from the perspective of the synthesized proteins or the overall structure of the proteome. My report is the first to suggest that the structure of the genetic code could be designed to assist the generation of domains with biased amino acid compositions in the proteins of the proteome.

Chargaff's Second Parity Rule (CSPR) is an empirical rule that states that in a genome sequence longer than a certain length, the number of thymine (T) and adenine (A) will be approximately equal, and the number of guanine (G) and cytosine (C) will also be approximately equal, although there is no confirmed requirement by the DNA itself or by the chromosomes themselves. It's known that this rule applies to most organisms with genomes of double-stranded DNA, and this has been a mystery in biology [11]. On the other hand, the results of the present study suggest that organisms could control the amount of their TMDs and IDRs in their proteomes by controlling the TA skew distributions in their genome sequences, and thus the balance between T and A in genome sequences must be crucial for maintaining optimal TMD and IDR fractions in genes. Looking at it the other way around, it's possible that CSPR results from genome sequences being controlled in coordination with the genetic code to regulate the fractions of TMDs and IDRs in the proteome by each organism controlling the distribution of TA skew in their genomes. This finding is consistent with previous reports that the entire genome is highly structured in terms of nucleic acid composition indices such as TA skew, GC skew, and GC content [12]. Therefore, it has been suggested that this CSPR may operate in conjunction with the current genetic code.

It is already known that the genetic code in mitochondrial genomes deviates from the standard code [13] and that mitochondrial DNA exhibits deviations from the CSPR [14]. However, the reasons for these deviations were not clear. In free-living organisms, the balanced generation of TMDs and IDRs appears to be a critical requirement, suggesting that significant deviations from the standard genetic code and the CSPR are typically not tolerated. In contrast, mitochondria have evolved to a state of complete intracellular parasitism. I hypothesize that the evolutionary selective pressure to encode functional proteins with such TMDs and IDRs has decreased in mitochondria, potentially leading to deviations in the genetic code and CSPR in these organelles.

My findings provide a new perspective on the structural design of the genetic code and link it to the functional requirements of proteomes. This insight may contribute to a better understanding of the evolutionary constraints and adaptability of the genetic code in different biological contexts.

Finally, and in addition, I propose an answer to the long-standing question of why UUU encodes phenylalanine. As I have shown in this study, the genetic code is thought to have a structure that converts dense thymine gene sequences into TMDs and scarce thymine sequences into IDRs. Thus, the UUU corresponding to the triple thymine on the gene must encode amino acid residues that are common in the TMDs and rare in the IDRs. And phenylalanine is a hydrophobic amino acid residue that is abundant in the TMDs and rare in the IDRs. Therefore, I assumed that UUU codes for phenylalanine because it is abundant in the TMDs and rare in the IDRs. I believe this is a possible answer to Nirenberg's historical question.

5. Conclusion

In this report, I have shown that the TMDs and IDRs always occupy their statistical extremes in the amino acid residue composition distribution of proteome proteins in all organisms. And together with previous reports that the genetic code has a function that assists its encoding of the TMDs and IDRs, I have also shown that the genetic code could become universal because these two domains/regions are universally required in all organisms.

This is a new explanation for the universality of the genetic code, and I call it "The Optimized Translation Theory". This theory should also provide a new perspective on the origin of the standard genetic code in terms of its functions.

6. References

1. Crick, F. H. C. (1968). The origin of the genetic code. In Journal of Molecular Biology (Vol. 38, Issue 3, pp. 367–379). Elsevier BV. https://doi.org/10.1016/0022-2836(68)90392-6

2. 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

3. Kun, Á., & Radványi, Á. (2018). The evolution of the genetic code: Impasses and challenges. In Biosystems (Vol. 164, pp. 217–225). Elsevier BV. https://doi.org/10.1016/j.biosystems.2017.10.006

4. Seki, M. (2023). On the origin of the genetic code. In Genes & Genetic Systems (Vol. 98, Issue 1, pp. 9–24). Genetics Society of Japan. https://doi.org/10.1266/ggs.22-00085

5. 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

6. Esumi, G. (2023). The TA Skew of a Gene Primarily Determines the Type of Protein, Such as Membrane Protein or Intrinsically Disordered Protein. Jxiv. https://doi.org/10.51094/jxiv.446

7. "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

8. 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 Accessed 1 Sep 2023

9. Esumi, G. (2023). 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. Jxiv. https://doi.org/10.51094/jxiv.561

10. Nirenberg, M. W., & Matthaei, J. H. (1961). The dependence of cell-free protein synthesis in E. coli upon naturally occurring or synthetic polyribonucleotides. In Proceedings of the National Academy of Sciences (Vol. 47, Issue 10, pp. 1588–1602). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.47.10.1588

11. Fariselli, P., Taccioli, C., Pagani, L., & Maritan, A. (2020). DNA sequence symmetries from randomness: the origin of the Chargaff's second parity rule. In Briefings in Bioinformatics (Vol. 22, Issue 2, pp. 2172–2181). Oxford University Press (OUP). https://doi.org/10.1093/bib/bbaa041

12. Esumi, G. (2023). 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. Jxiv. https://doi.org/10.51094/jxiv.436

13. Osawa, S., Ohama, T., Jukes, T. H., & Watanabe, K. (1989). Evolution of the mitochondrial genetic code I. Origin of AGR serine and stop codons in metazoan mitochondria. In Journal of Molecular Evolution (Vol. 29, Issue 3, pp. 202–207). Springer Science and Business Media LLC. https://doi.org/10.1007/bf02100203

14. Nikolaou, C., & Almirantis, Y. (2006). Deviations from Chargaff's second parity rule in organellar DNA. In Gene (Vol. 381, pp. 34–41). Elsevier BV. https://doi.org/10.1016/j.gene.2006.06.010