The standard genetic code is designed to generate transmembrane domains and intrinsically disordered regions as projections of the thymine density on the gene

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

Abstract

We know that the codon-amino acid correspondence in the genetic code is not random. However, there was no established theory as to whether this correspondence was designed for any purpose or function. In a previous report, I showed that the proteins with high amounts of transmembrane domains and the proteins with high amounts of intrinsically disordered regions correspond to the high and low TA (thymine adenine) skew of their gene, respectively, and I speculated that these reflect the purpose behind the design of the genetic code. However, since most protein genes use their synonymous codon selection to balance their GC (guanine cytosine) content, i.e., their TA content, I hypothesized that the amount of only one of these two nucleic acids, thymine or adenine, might actually originate the characteristics of the amino acid composition of these two functional domains/regions.

In this study, I examined the correspondence between these two functional domains/regions and the estimated composition of each nucleic acid of various protein genes from different organism proteomes by back-calculating the possible nucleic acid compositions of the gene from the amino acid residue composition of the protein.

The results showed that the proteins with high amounts of transmembrane domains and the proteins with high amounts of intrinsically disordered regions were indeed correlated with the higher and lower estimated thymine composition on the genes, respectively. Upon detailed analysis, the transmembrane domains correlated more strongly with the maximum estimated thymine composition and the intrinsically disordered regions correlated more strongly with the minimum estimated thymine composition.

Since the amino acid compositions of membrane proteins with higher thymine composition genes correspond to the maximum estimated thymine compositions, and the amino acid compositions of intrinsically disordered proteins with lower thymine composition genes correspond to the minimum estimated thymine compositions, it is more reasonable to assume that the characteristic amino acid compositions of the two domains/regions are both formed by the thymine densities of the genes, rather than these thymine density structures being formed by selective pressure on amino acid compositions. Thus, the functions of these two functional domains/regions are thought to arise as projections of the thymine densities of their properly preformed gene sequences.

The results shown in this study suggest that the standard genetic code has an optimized structure that allows for optimized translation and synthesis of the functional domains of proteins. I conclude that the current genetic code must have been selected for this functional advantage, and I propose this as the "optimized translation" theory that explains the origin of the genetic code.

Keywords: standard genetic code, thymine composition, transmembrane domains, intrinsically disordered regions, optimized translation

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

1. Background

We know that the codon-amino acid correspondence in the genetic code is not random. However, there was no established theory as to whether this correspondence was designed for any purpose or function. In a previous report, I showed that the proteins with high amounts of transmembrane domains (TMDs) and the proteins with high amounts of intrinsically disordered regions (IDRs) correspond to the high and low TA (thymine adenine) skew of their gene, respectively, and I speculated that these reflect the purpose behind the design of the genetic code [1]. However, since most protein genes use their synonymous codon selection to balance their GC (guanine cytosine) content, i.e., their TA content [2], I hypothesized that the amount of only one of these two nucleic acids, thymine or adenine, might actually originate the characteristics of the amino acid composition of these two functional domains/regions.

Therefore, in this study, I examined the correspondence between these two functional domains/regions and the estimated composition of each nucleic acid of various protein genes from different organism proteomes by back-calculating the possible nucleic acid compositions of the gene from the amino acid residue composition of the protein.

2. Materials and Methods

In this study, I used a protein dataset published on the Internet as a "reference proteome" consisting of more than one million protein entries [3]. From this total of protein entries, I selected the target proteins of this study under the conditions that they matched the amino acid sequence in the UniProt database, had no missing or exceptional alphabetic codes in the amino acid sequence data, and had no missing or exceptional descriptions of the residue information of TMDs or IDRs from the UniProt database.

First, the amino acid composition of each protein was calculated by counting the amino acid residues for each protein in the FASTA format file and dividing this number by the sum of all 20 amino acids in the protein. As a result, each amino acid composition of a protein ranged between 0 and 1, and their total sum was 1.

Second, from these amino acid compositions of each protein, the estimated nucleic acid compositions on the gene were back-calculated according to the standard genetic code. However, because most amino acids have several corresponding codons, called synonymous codons, in this back-calculation process, these estimates cannot be determined to unique values and inevitably have some ranges. Therefore, in this study, I calculated three typical estimates, maximum estimated composition, minimum estimated composition, and average estimated composition. "Maximum estimated composition" means the estimated value using only the codon with the highest percentage of each target nucleic acid among the synonymous codons. "Minimum estimated composition" means the estimated value using only the codon with the lowest percentage of each target nucleic acid among the synonymous codons. "Average estimated composition" means the estimated value when each synonymous codon is used in the same proportion for each amino acid.

Third, the number of residues in the amino acid sequence of TMD and IDR were calculated for each protein, and then the TMD fraction and IDR fraction of each protein were calculated. Although there are two types of TMDs, alpha-helical and beta-barrel, but because only alpha-helical TMDs were found to have a correlation with the TA skew of their genes in the previous study [1], "TMD" in this study refers only to alpha-helical TMD unless otherwise noted.

Fourth, for each of the TMD and IDR fractions of the target protein, I set two boundaries: the first boundary was whether the TMD and IDR were present, and the second boundary was whether the fraction of the protein was in the upper 10th percentile of the total.

Fifth, for each of the two boundaries of TMD fraction and IDR fraction set in the fourth, whether each typical estimated nucleic acid composition calculated in the second could be a statistical estimator was evaluated using AUC values calculated from ROC curves.

Sixth, all typical estimated nucleic acid compositions back-calculated from the TMD and IDR sequences extracted from each protein and those of the remaining protein sequences other than each TMD and IDR were calculated and compared in their distributions.

In this study, I used Microsoft® Excel for Mac v16.77.1 (Microsoft Corporation, Redmond, WA, USA) to generate compositions and other calculation results. I also used JMP® 17.2.0 (SAS Institute Inc., Chicago, IL, USA) to generate ROC curves, AUC values, graphs, and figures.

3. Results

3.1. Target Proteins

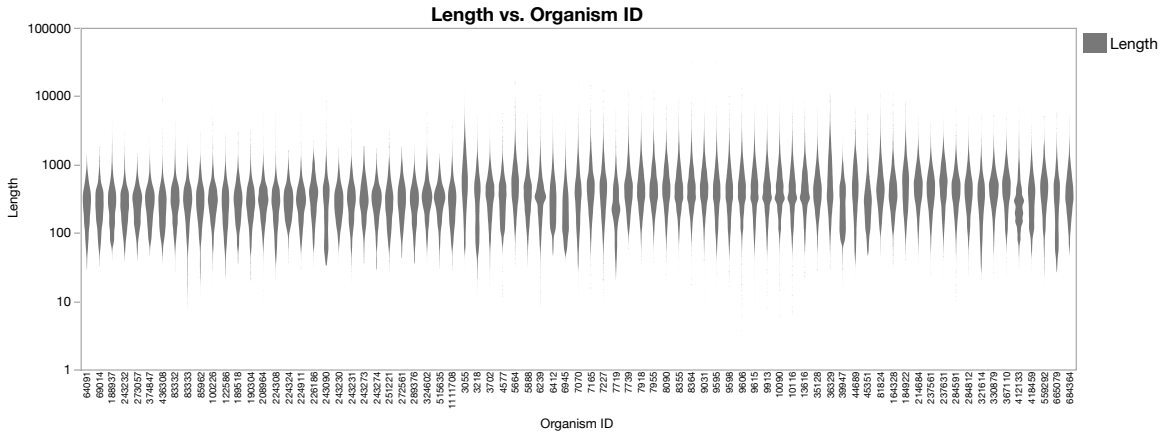
The "reference proteome" used in this study is a protein entry list consisting of 1,023,125 proteins from 79 species across all 3 domains [3]. I used the amino acid sequence FASTA data files attached to this list. Among all protein entries, the amino acid sequence was missing for 120 entries in the mouse proteome in the attached files, and the sequences of seven proteins were inconsistent with the UniProt database in the human proteome [4], so these 127 proteins were excluded. Among the remaining 1,022,998 proteins, I also excluded 3,994 proteins that contained indeterminate amino acid residues or unusual amino acid residues other than the typical 20 amino acids, or proteins with an incomplete description for their TMDs or IDRs, resulting in 1,019,004 proteins as targets for analysis. The number of target proteins from all 79 species is shown in Table 1, and the distribution of their protein length (number of amino acid residues) is shown for each organism in Figure 1. The size of each proteome varied by species and domain, but there were no extreme differences in protein length distribution among all organisms (Table 1, Figure 1).

Table 1: Table of target protein numbers for all 79 species

Taxonomy ID	Domain	Organism Name	Listed Proteins	Target Proteins
64091	Archaea	Halobacterium salinarum	2423	2423
69014	Archaea	Thermococcus kodakarensis	2301	2301
188937	Archaea	Methanosarcina acetivorans	4468	4456
243232	Archaea	Methanocaldococcus jannaschii	1787	1774
273057	Archaea	Saccharolobus solfataricus	2937	2936
374847	Archaea	Korarchaeum cryptofilum	1602	1602
436308	Archaea	Nitrosopumilus maritimus	1795	1795
83332	Bacteria	Mycobacterium tuberculosis	3995	3995
83333	Bacteria	Escherichia coli	4403	4393
85962	Bacteria	Helicobacter pylori	1554	1543
100226	Bacteria	Streptomyces coelicolor	8035	8035
122586	Bacteria	Neisseria meningitidis serogroup B	2001	2001
189518	Bacteria	Leptospira interrogans serogroup icterohaemorrhagiae serovar Lai	3676	3676
190304	Bacteria	Fusobacterium nucleatum subsp. nucleatum	2046	2046
208964	Bacteria	Pseudomonas aeruginosa	5564	5563
224308	Bacteria	Bacillus subtilis	4260	4259
224324	Bacteria	Aquifex aeolicus	1553	1550
224911	Bacteria	Bradyrhizobium diazoefficiens	8253	8253
226186	Bacteria	Bacteroides thetaiotaomicron	4782	4782
243090	Bacteria	Rhodopirellula baltica	7271	7271
243230	Bacteria	Deinococcus radiodurans	3084	3060
243231	Bacteria	Geobacter sulfurreducens	3402	3393
243273	Bacteria	Mycoplasma genitalium	483	483
243274	Bacteria	Thermotoga maritima	1852	1851
251221	Bacteria	Gloeobacter violaceus	4405	4406
272561	Bacteria	Chlamydia trachomatis	895	895
289376	Bacteria	Thermodesulfovibrio yellowstonii	1982	1977
324602	Bacteria	Chloroflexus aurantiacus	3850	3849
515635	Bacteria	Dictyoglomus turgidum	1743	1743
1111708	Bacteria	Synechocystis sp.	3507	3506
3055	Eukaryota	Chlamydomonas reinhardtii	17614	17602
3218	Eukaryota	Physcomitrium patens	31359	31287
3702	Eukaryota	Arabidopsis thaliana	27481	27476
4577	Eukaryota	Zea mays	39225	39198
5664	Eukaryota	Leishmania major	8038	8036
5888	Eukaryota	Paramecium tetraurelia	39461	39256
6239	Eukaryota	Caenorhabditis elegans	19827	19826
6412	Eukaryota	Helobdella robusta	23328	23294
6945	Eukaryota	Ixodes scapularis	20496	20461
7070	Eukaryota	Tribolium castaneum	16568	16552
7165	Eukaryota	Anopheles gambiae	13016	12989
7227	Eukaryota	Drosophila melanogaster	13821	13594
7719	Eukaryota	Ciona intestinalis	16680	16614
7739	Eukaryota	Branchiostoma floridae	26627	26421
7918	Eukaryota	Lepisosteus oculatus	18321	17988
7955	Eukaryota	Danio rerio	26249	26094
8090	Eukaryota	Oryzias latipes	23617	23614
8355	Eukaryota	Xenopus laevis	35860	35595
8364	Eukaryota	Xenopus tropicalis	22229	22104
9031	Eukaryota	Gallus gallus	18369	18337
9595	Eukaryota	Gorilla gorilla gorilla	21783	21493
9598	Eukaryota	Pan troglodytes	23051	22963
9606	Eukaryota	Homo sapiens	20586	20486
9615	Eukaryota	Canis lupus familiaris	20972	20935
9913	Eukaryota	Bos taurus	23841	23798
10090	Eukaryota	Mus musculus	21957	21680
10116	Eukaryota	Rattus norvegicus	22870	22816
13616	Eukaryota	Monodelphis domestica	21223	21084
35128	Eukaryota	Thalassiosira pseudonana	11717	11717
36329	Eukaryota	Plasmodium falciparum	5372	5368
39947	Eukaryota	Oryza sativa subsp. japonica	43672	43656
44689	Eukaryota	Dictyostelium discoideum	12726	12713
45351	Eukaryota	Nematostella vectensis	24427	24322
81824	Eukaryota	Monosiga brevicollis	9188	9177
164328	Eukaryota	Phytophthora ramorum	15349	15284
184922	Eukaryota	Giardia intestinalis	4900	4900
214684	Eukaryota	Cryptococcus neoformans var. neoformans serotype D	6604	6597
237561	Eukaryota	Candida albicans	6035	5984
237631	Eukaryota	Ustilago maydis	6788	6788
284591	Eukaryota	Yarrowia lipolytica	6449	6449
284812	Eukaryota	Schizosaccharomyces pombe	5122	5122
321614	Eukaryota	Phaeosphaeria nodorum	15998	15998
330879	Eukaryota	Aspergillus fumigatus	9647	9647
367110	Eukaryota	Neurospora crassa	9759	9759
412133	Eukaryota	Trichomonas vaginalis	50190	49814
418459	Eukaryota	Puccinia graminis f. sp. tritici	15688	15688
559292	Eukaryota	Saccharomyces cerevisiae	6060	6059
665079	Eukaryota	Sclerotinia sclerotiorum	14445	14445
684364	Eukaryota	Batrachochytrium dendrobatidis	8610	8610
Total			1023125	1019004

The number of target proteins for all 79 species is shown, where the length of each colored bar indicates the number of target proteins in the species proteome. Each bar is colored according to the domains to which it belongs. Archaea; blue, Bacteria; red, Eukaryotes; green.

Figure 1: Distributions of protein length (number of amino acid residues) for each organism

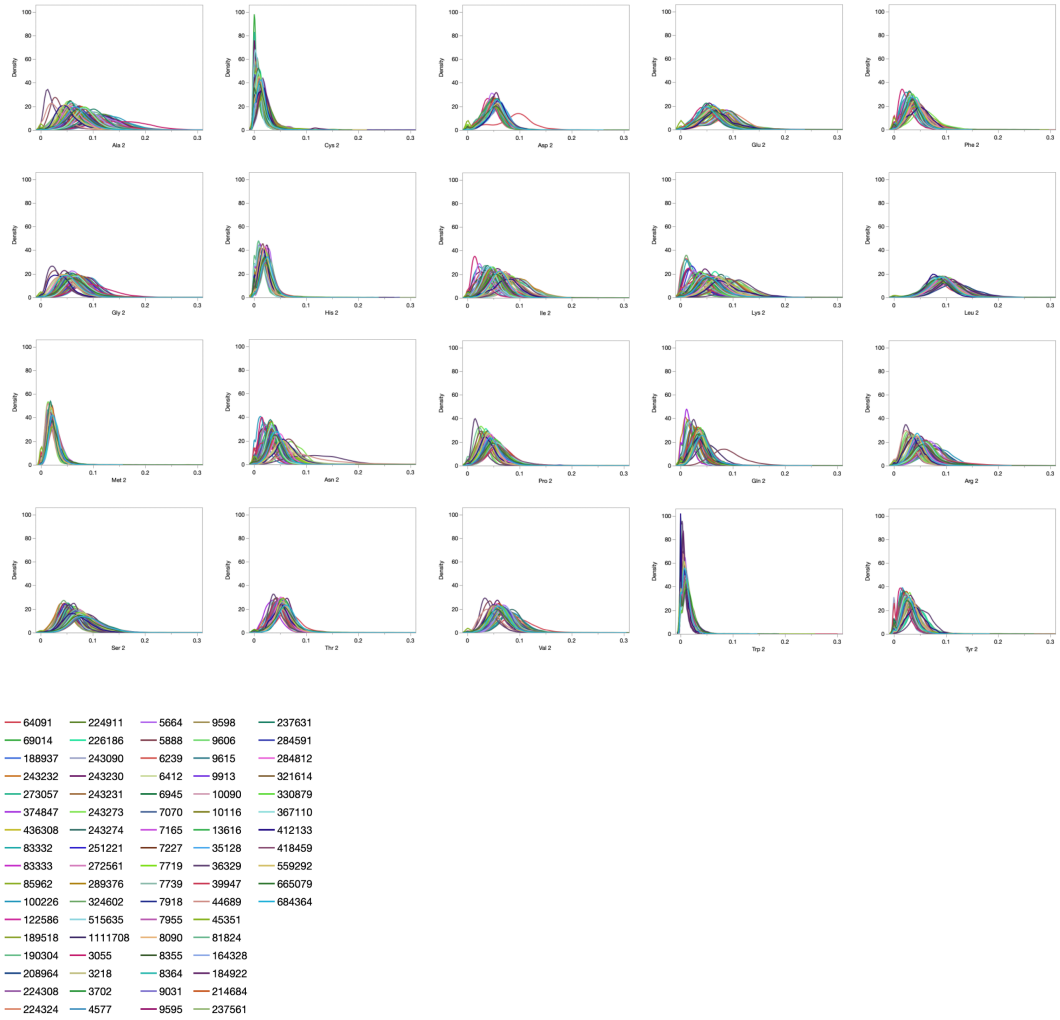


The protein length distributions are plotted by species (organism). The vertical axis indicating protein length is set to logarithmic. There do not appear to be extreme differences in protein length distributions across domains among all organisms.

3.2. Amino Acid Compositions

The distributions of the calculated amino acid compositions of the target proteins by organism are shown in Figure 2. These distributions were normalized so that the area under the curve was 1 and plotted for each of the total 20 amino acids for all organisms. Each distribution from each organism uniformly follows binomial distributions, as I have shown in a previous paper [5]. Further analysis showed that these peaks were most influenced by the GC content of the organism's genes (data not shown). I also found several distributions that were bimodal rather than unimodal, with the smaller of these peaks being formed by clusters of proteins with high fractions of TMDs (data also not shown).

Figure 2: Amino acid composition distributions of target proteins by organism



The distributions of the amino acid compositions by organism are shown. These distributions were normalized so that the area under the curve was 1 and plotted for each of the total 20 amino acids. Each distribution from each organism uniformly follows binomial distributions, as I have shown in a previous paper [5].

3.3. Reverse translation of the genetic code

For the backward calculation of the genetic code, the codons corresponding to each amino acid were enumerated, and the maximum, minimum, and average composition of each nucleic acid in these codons corresponding to each amino acid were calculated and shown in the table (Table 2). Then, these typical nucleic acid compositions of each corresponding synonymous codon set were used as coefficients corresponding to the composition of each amino acid. Specifically, the estimated nucleic acid compositions were calculated by summing the product of each amino acid composition and the corresponding coefficient.

Table 2: All 20 amino acids, their corresponding codons and their three typical nucleic acid composition estimates

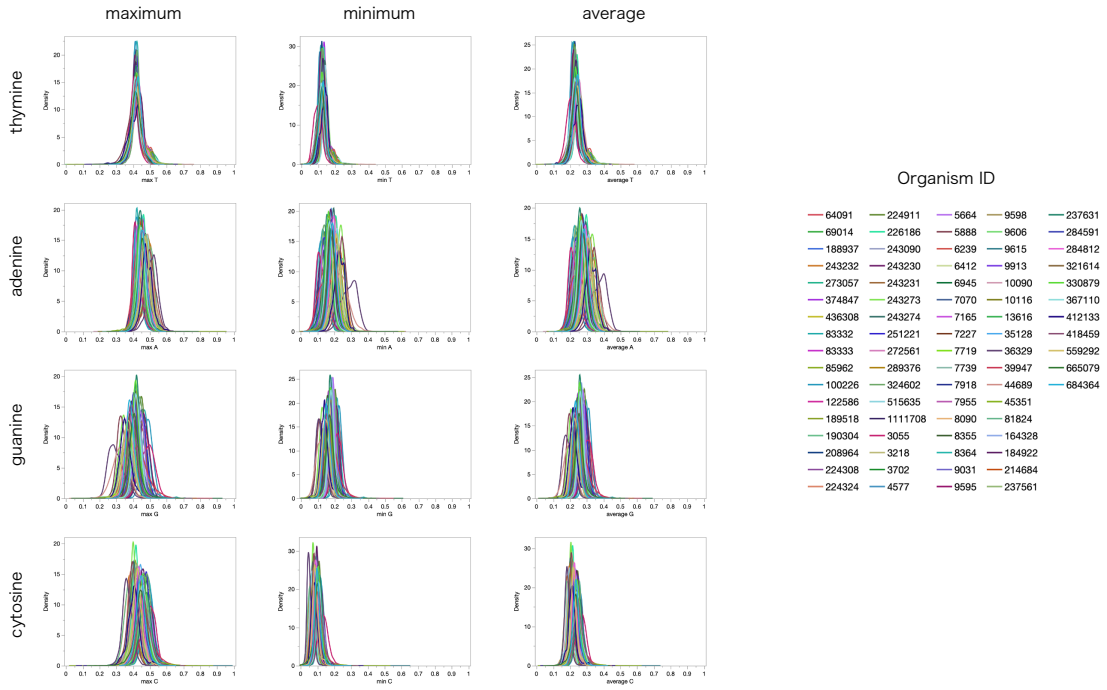
Amno Acid	Corresponding Codons						max U	min U	ave U	max A	min A	ave A	max G	min G	ave G	max C	min C	ave C
Ala	GCU	GCA	GCC	GCG			1 / 3	0 / 3	1/(4*3)	1 / 3	0 / 3	1/(4*3)	2 / 3	1 / 3	5/(4*3)	2 / 3	1 / 3	5/(4*3)
Cys	UGU	UGC					2 / 3	1 / 3	3/(2*3)	0 / 3	0 / 3	0/(2*3)	1 / 3	1 / 3	2/(2*3)	1 / 3	0 / 3	1/(2*3)
Asp	GAU	GAC					1 / 3	0 / 3	1/(2*3)	1 / 3	1 / 3	2/(2*3)	1 / 3	1 / 3	2/(2*3)	1 / 3	0 / 3	1/(2*3)
Glu	GAA	GAG					0 / 3	0 / 3	0/(2*3)	2 / 3	1 / 3	3/(2*3)	2 / 3	1 / 3	3/(2*3)	0 / 3	0 / 3	0/(2*3)
Phe	UUU	UUC					3 / 3	2 / 3	5/(2*3)	0 / 3	0 / 3	0/(2*3)	0 / 3	0 / 3	0/(2*3)	1 / 3	0 / 3	1/(2*3)
Gly	GGU	GGA	GGC	GGG			1 / 3	0 / 3	1/(4*3)	1 / 3	0 / 3	1/(4*3)	3 / 3	2 / 3	9/(4*3)	1 / 3	0 / 3	1/(4*3)
His	CAU	CAC					1 / 3	0 / 3	1/(2*3)	1 / 3	1 / 3	2/(2*3)	0 / 3	0 / 3	0/(2*3)	2 / 3	1 / 3	3/(2*3)
Ile	AUU	AUA	AUC				2 / 3	1 / 3	4/(3*3)	2 / 3	1 / 3	4/(3*3)	0 / 3	0 / 3	0/(3*3)	1 / 3	0 / 3	1/(3*3)
Lys	AAA	AAG					0 / 3	0 / 3	0/(2*3)	3 / 3	2 / 3	5/(2*3)	1 / 3	0 / 3	1/(2*3)	0 / 3	0 / 3	0/(2*3)
Leu	CUU	CUA	CUC	CUG	UUA	UUG	2 / 3	1 / 3	9/(6*3)	1 / 3	0 / 3	2/(6*3)	1 / 3	0 / 3	2/(6*3)	2 / 3	0 / 3	5/(6*3)
Met	AUG						1 / 3	1 / 3	1/(1*3)	1 / 3	1 / 3	1/(1*3)	1 / 3	1 / 3	1/(1*3)	0 / 3	0 / 3	0/(1*3)
Asn	AAU	AAC					1 / 3	0 / 3	1/(2*3)	2 / 3	2 / 3	4/(2*3)	0 / 3	0 / 3	0/(2*3)	1 / 3	0 / 3	1/(2*3)
Pro	CCU	CCA	CCC	CCG			1 / 3	0 / 3	1/(4*3)	1 / 3	0 / 3	1/(4*3)	1 / 3	0 / 3	1/(4*3)	3 / 3	2 / 3	9/(4*3)
Gln	CAA	CAG					0 / 3	0 / 3	0/(2*3)	2 / 3	1 / 3	3/(2*3)	1 / 3	0 / 3	1/(2*3)	1 / 3	1 / 3	2/(2*3)
Arg	CGU	CGA	CGC	CGG	AGA	AGG	1 / 3	0 / 3	1/(6*3)	2 / 3	0 / 3	4/(6*3)	2 / 3	1 / 3	8/(6*3)	2 / 3	0 / 3	5/(6*3)
Ser	UCU	UCA	UCC	UCG	AGU	AGC	2 / 3	0 / 3	6/(6*3)	1 / 3	0 / 3	3/(6*3)	1 / 3	0 / 3	3/(6*3)	2 / 3	0 / 3	6/(6*3)
Thr	ACU	ACA	ACC	ACG			1 / 3	0 / 3	1/(4*3)	2 / 3	1 / 3	5/(4*3)	1 / 3	0 / 3	1/(4*3)	2 / 3	1 / 3	5/(4*3)
Val	GUU	GUA	GUC	GUG			2 / 3	1 / 3	5/(4*3)	1 / 3	0 / 3	1/(4*3)	2 / 3	1 / 3	5/(4*3)	1 / 3	0 / 3	1/(4*3)
Trp	UGG						1 / 3	1 / 3	1/(1*3)	0 / 3	0 / 3	0/(1*3)	2 / 3	2 / 3	2/(1*3)	0 / 3	0 / 3	0/(1*3)
Tyr	UAU	UAC					2 / 3	1 / 3	3/(2*3)	1 / 3	1 / 3	2/(2*3)	0 / 3	0 / 3	0/(2*3)	1 / 3	0 / 3	1/(2*3)

All 20 amino acids, their corresponding codons, and their three typical nucleic acid composition estimates are shown. The amino acids are arranged according to their corresponding alphabetic symbols. In the table, "max U" indicates the maximum uracil composition among the corresponding codons of the amino acid. Uracil on the genetic code corresponds to thymine on the gene. These compositions have been used as the amino acid composition coefficient for reverse translation in the genetic code.

3.4. Estimated nucleic acid compositions

The distributions of the estimated nucleic acid compositions of the target proteins by organism are shown in the figure (Figure 3). For the other three nucleic acids except thymine, the distribution of their estimated compositions varied from organism to organism, but only the distribution of the estimated composition of thymine was very close across organisms for all typical estimates of maximum, minimum, and average composition.

Figure 3: Distribution of each estimated nucleic acid composition by organism



The distributions of the estimated nucleic acid compositions of the target proteins by organism are shown. These distributions were normalized so that the area under the curve was 1 and plotted for all 79 organisms. Only the distribution of estimated thymine composition was very close across organisms for all typical estimates of maximum, minimum, and average composition.

3.5. TMD and IDR fractions

All target proteins were divided into three groups by the two boundaries described in the method, and the area corresponding to the number of proteins for each organism was displayed (Figure 4a, 4b). Red: those without TMDs or IDRs, green: those with TMDs or IDRs but whose fraction is less than the upper 10th percentile, blue: those with TMDs or IDRs and whose fraction is in the upper 10th percentile. The proteomes of archaea and bacteria contained fewer proteins with IDRs; conversely, IDRs were relatively abundant in the proteomes of eukaryotes.

Figure 4a: Proportions of each TMD fraction-binned group of proteins displayed by organism

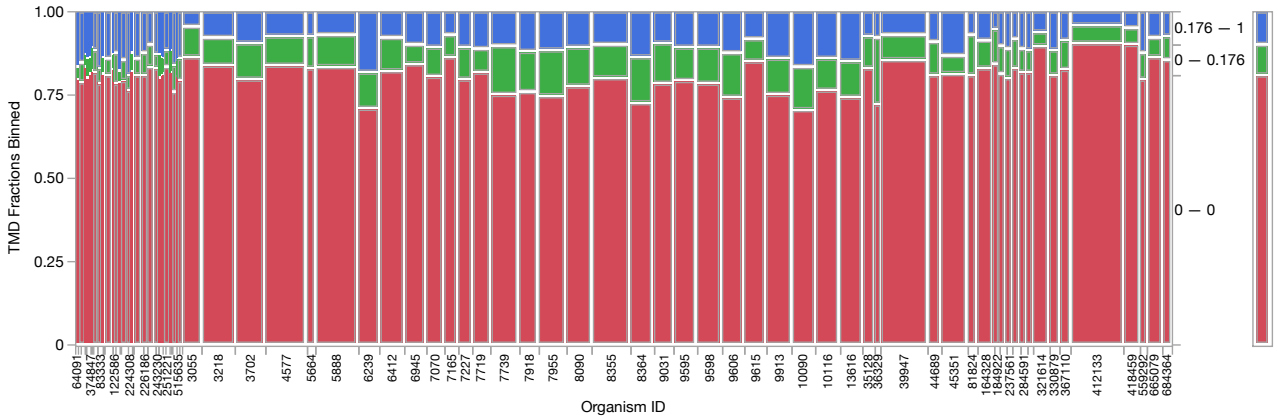
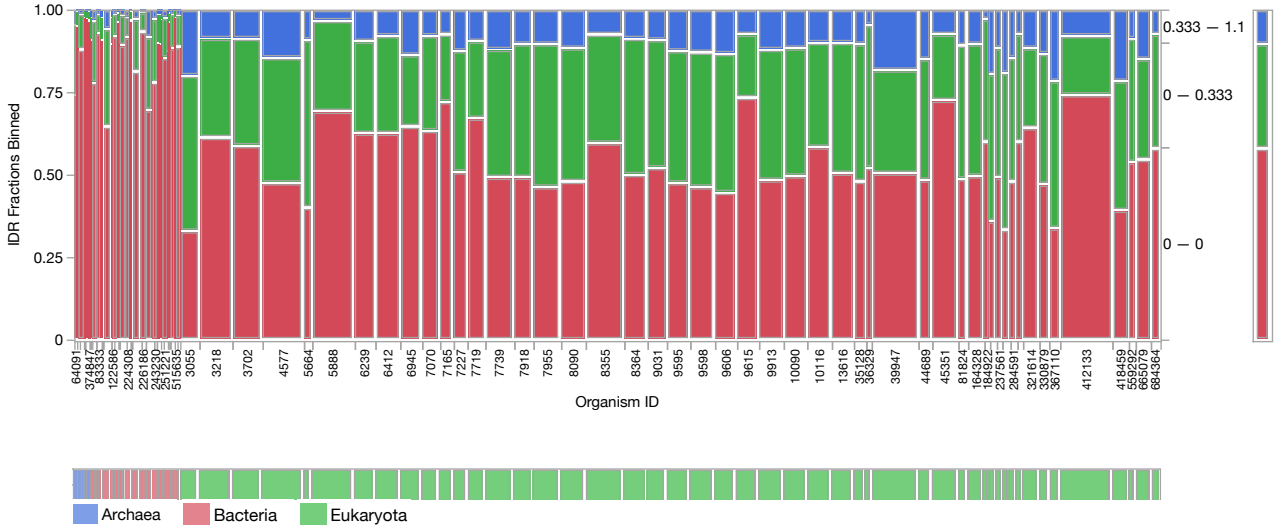


Figure 4b: Proportions of each IDR fraction-binned group of proteins displayed by organism



All target proteins were divided into three groups by the two boundaries described in the method and displayed in the area corresponding to the number of proteins for each organism. The organisms are listed from left to right: Archaea, Bacteria, and Eukaryotes. Red: those without TMDs or IDRs, green: those with TMDs or IDRs but whose fractions are less than the upper 10th percentile, blue: those with TMDs or IDRs and whose fractions are in the upper 10th percentile. The upper 10th percentile cutoffs were 0.176 for TMD fractions and 0.333 for IDR fractions. The proteomes of archaea and bacteria contained fewer proteins with IDRs; conversely, IDRs were relatively abundant in the proteomes of eukaryotes.

3.4. AUCs by ROC curves

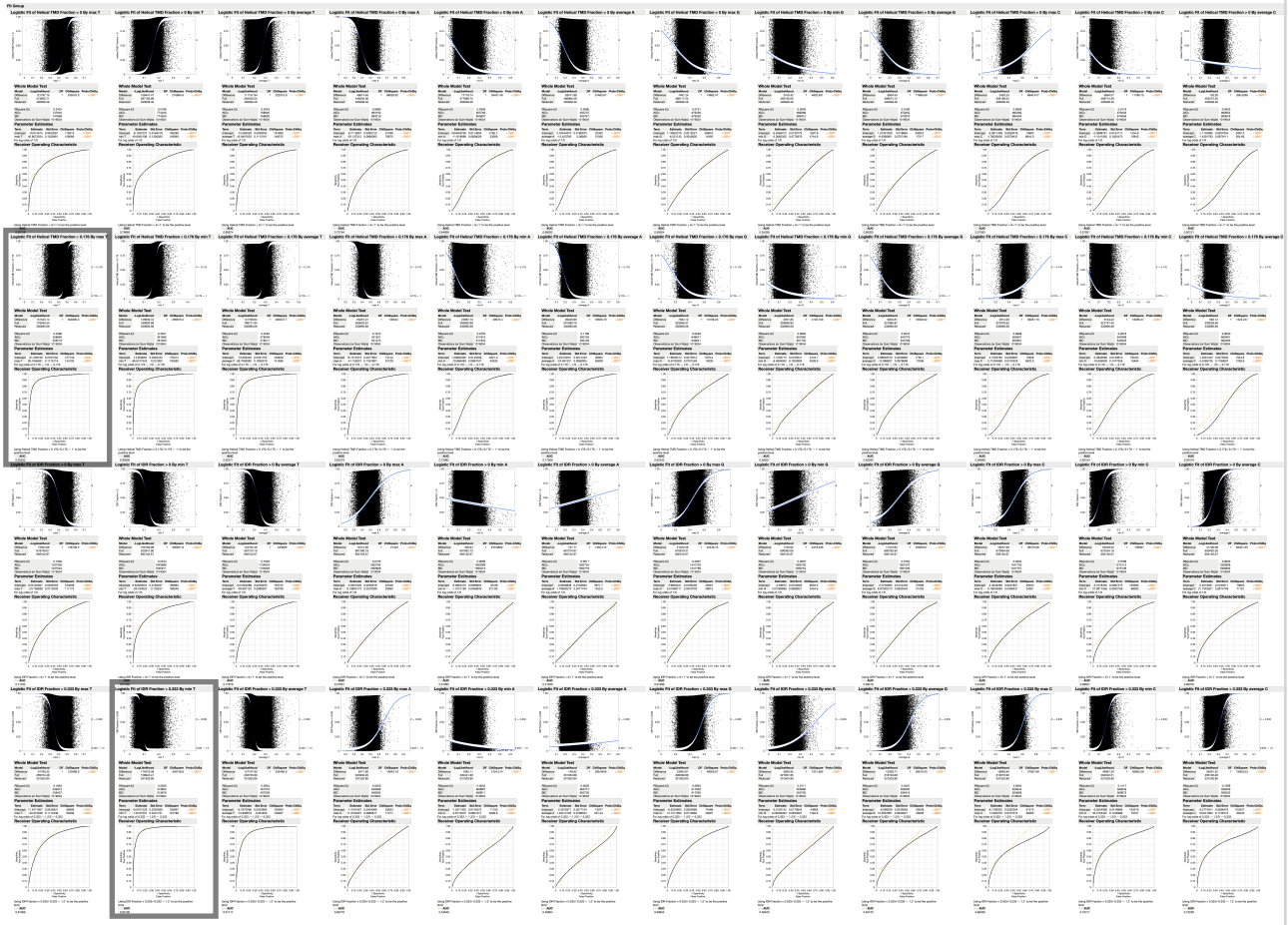
For each of the two boundaries of the TMD fraction and the IDR fraction of all target proteins, the AUC calculated from the ROC curve was used to evaluate whether the currently calculated typical estimated nucleic acid composition could be a statistical estimator for that boundary. The results showed that the maximum AUC for the TMD fractions was between their upper 10th percentile and the maximum estimated thymine composition (max T), while the maximum AUC for the IDR fraction was between the upper 10th percentile and the minimum estimated thymine composition (min T) (Table 3, Figure 5). Of the four nucleic acids, thymine appears to be the best estimator of proteins with both high TMD and high IDR amounts.

I calculated and evaluated several TA skew estimates using the current estimated nucleic acid compositions, but all AUCs were lower than the AUC of thymine alone. (data not shown).

Table 3: The AUCs in the ROC curve of each nucleic acid composition estimate and each cutoff of the TMD and IDR fractions

	max T	min T	ave T	max A	min A	ave A	max G	min G	ave G	max C	min C	ave C
TMD > 0	0.827	0.799	0.809	0.728	0.649	0.683	0.593	0.55	0.6	0.576	0.579	0.501
TMD ≧ 0.176	0.935	0.928	0.923	0.834	0.73	0.776	0.624	0.564	0.633	0.596	0.632	0.524
IDR > 0	0.714	0.82	0.777	0.58	0.52	0.519	0.62	0.527	0.584	0.61	0.686	0.664
IDR ≧ 0.333	0.811	0.951	0.911	0.608	0.546	0.508	0.697	0.56	0.641	0.664	0.754	0.722

Figure 5: The ROC curves of each nucleic acid composition estimate and each cutoff of the TMD and IDR fractions



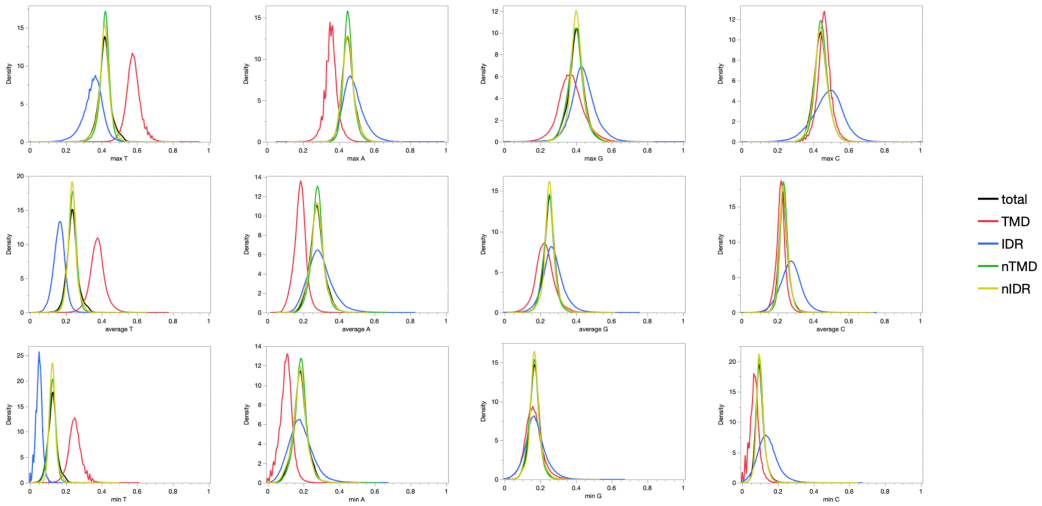
All AUCs and ROC curves examined in this study are shown in the table and figure. The highest AUCs of the TMD and IDR fractions are both marked with squares. Of the four nucleic acids, thymine appears to be the best estimator of proteins with both high TMD and high IDR amounts.

3.5. Distributions of estimated nucleic acid compositions of TMDs and IDRs

For all target proteins, the distributions of each estimated nucleotide composition were compared between the TMD, the non-TMD portion on that protein, the IDR on each protein, the non-IDR portion on that protein, and the total target proteins (Figure 6). The results show that the estimated compositions of the total and non-TMD and non-IDR portions were nearly identical for all estimated nucleotide compositions. On the other hand, the estimated thymine compositions of the TMDs were uniformly higher and those of the IDRs were uniformly lower compared to those of the other portions. In addition, the estimated adenine compositions of the TMDs were lower than those of the total, and the estimated cytosine compositions of the IDRs were higher and more widely distributed than those of the total, but these distribution differences to the total composition are smaller than those of thymine.

In the next figures, these distributions are shown separately for each of the 79 total target organisms by TMD and IDR (Figures 7a and 7b). The characteristics of the distributions shown in Figure 6 were observed uniformly across all organisms in Figure 7, indicating that these characteristics are universal and do not differ between organisms, particularly in the estimates of thymine composition.

Figure 6: The distributions of the estimated nucleic acid composition of the sequences for the TMDs and the IDRs.



The distributions of each estimated nucleic acid composition of the sequences of TMDs, non-TMDs of these TMD proteins, IDRs, non-IDRs of these IDR proteins, and total proteins are shown.

Total; those of total target proteins, TMD; those of TMD sequence, nTMD; those of non-TMD sequence of TMD proteins, IDR; those of IDR sequence, non-IDR; those of non-IDR sequence of IDR proteins. These distributions were normalized so that the area under the curve was 1.

Figure 7a: The distributions of the estimated nucleic acid composition of the sequences for the TMD and non-TMD regions by organism

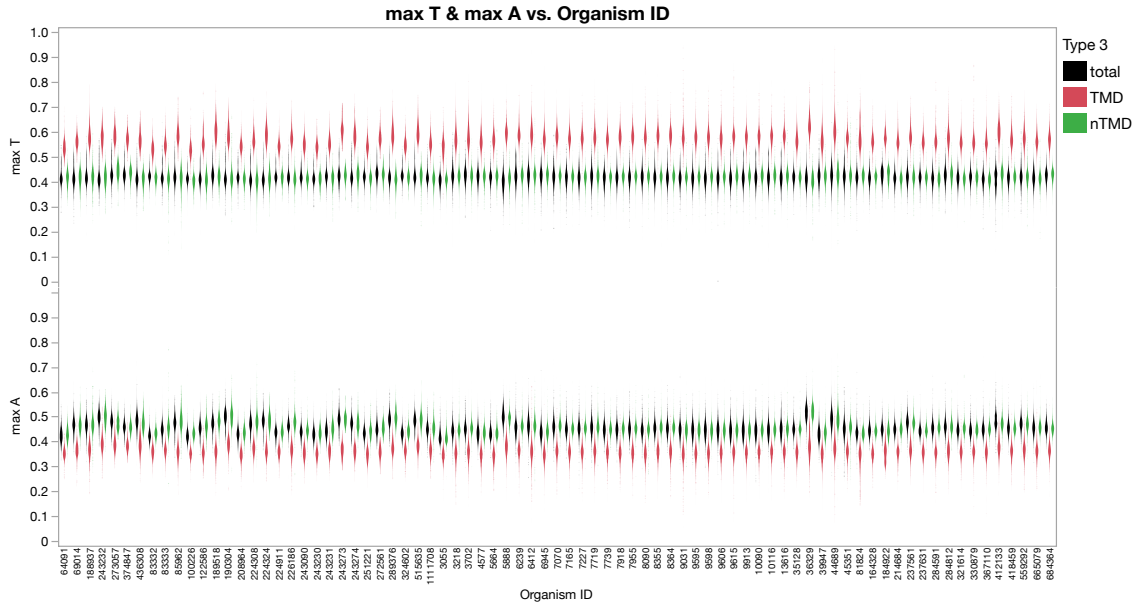
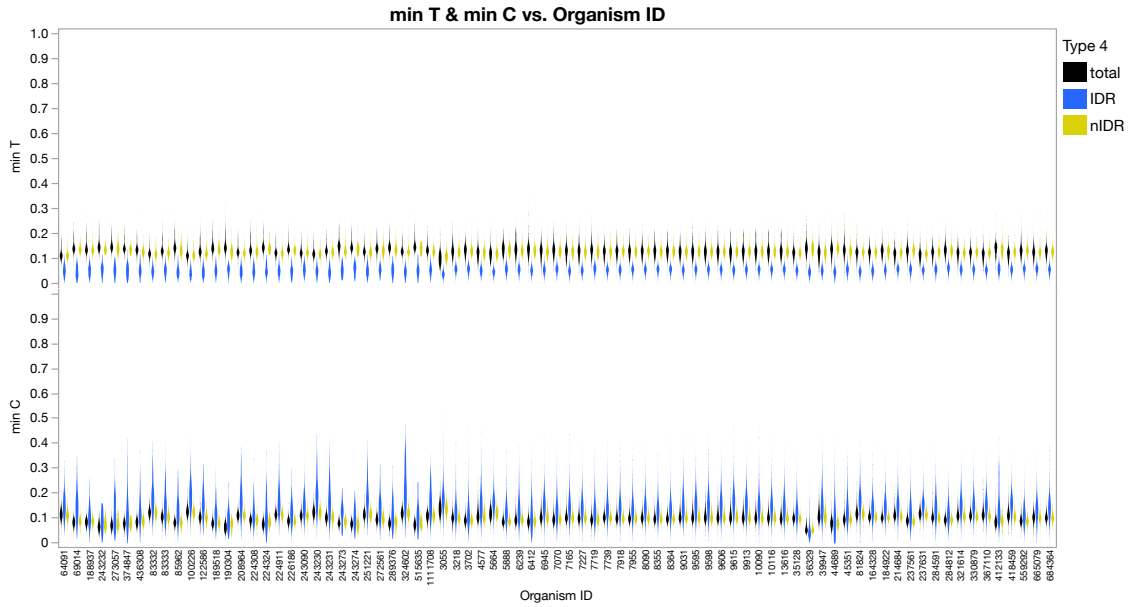


Figure 7b: The distributions of the estimated nucleic acid composition of the sequences for the IDR and non-IDR regions by organism



The distributions of Figure 6 are shown in detail for the TMDs and IDRs of all 79 organisms. Figure 7a shows those of the maximum estimated thymine composition (max T) and those of the maximum estimated adenine composition (max A). These two estimates of nucleotide composition were found to differ in TMDs from the total in Figure 6. Similarly, Figure 7b shows the distributions of the minimum estimated thymine composition (min T) and those of the minimum estimated adenine composition (min A).

4.Discussion

In a previous report, I showed that the proteins with high amounts of TMDs and the proteins with high amounts of IDRs correspond to the high and low TA (thymine adenine) skew of their gene, respectively [1]. However, there were papers showing that the amount of thymine in the gene correlated with membrane proteins and that the thymine-adenine difference in the gene correlated with the transmembrane domain [6, 7]. But I was the first to report that the TA skew correlated with the TMDs and also with the IDRs [1]. It is then natural to assume that the translation of the genetic code is behind the influence of the nucleic acid composition of genes on the amino acid compositional function of proteins. However, in studying the implementation of the genetic code, it has been quite complicated and not easy to analyze this by comparing the nucleic acid composition of actual genes with the amino acid composition of actual protein residues, because of the inevitable inclusion of ambiguity due to the presence of synonymous codons. I believe this is the reason why there have been few reports on the function of the genetic code itself, and no consistent explanation has been provided.

In this study, to avoid the ambiguity problem described above, the analyses were performed using estimated nucleic acid compositions calculated backwards from the amino acid compositions of the protein amino acid sequences, rather than the nucleic acid compositions of the actual genes. However, and not surprisingly, the synonymous codon problem does not disappear even when the estimated nucleic acid composition is used. Therefore, in this study, I used three representative values of the estimated nucleic acid composition: the maximum value, the minimum value, and the value when all synonymous codons are used equally (called the average value). If the selection of synonymous codons were random, this method might not be effective. However, I have reported in a previous report that there are certain directions and rules in the selection of synonymous codons [2], so I thought it was effective to use representative values for the analysis.

The results of this study showed that the estimated thymine compositions, calculated backwards from the amino acid residue composition of the protein, are good statistical estimators of the proportions of TMDs and IDRs on the protein. In the present results, thymine alone was a better statistical estimator than the TA skew calculated from both thymine and adenine. This may be due to the fact that the interference of synonymous codon selection is removed, allowing us to see the more essential behavior.

In the detailed analysis, among the estimates, TMD was found to be strongly correlated with the maximum estimated thymine composition (max T), while IDR was found to be strongly correlated with the minimum estimated thymine composition (min T). I will discuss the background in the following paragraph.

> When amino acid composition evolves by mutation, it has generally been assumed that selective pressure acts on amino acid composition as a phenotype. However, in a previous paper, I reported that amino acid sequences in genomes, whether protein-coding or not, are highly structured in terms of nucleic acid composition [8]. Therefore, it is possible to question whether domains are composed of originally structured, biased nucleic acid sequences, or whether the nucleic acid composition of domain sequences is biased as a result of biased amino acid composition. If selection pressure acts on amino acid composition, and a TMD that requires a large number of thymines to encode emerges as a result of competition with the existing balanced nucleic acid composition, then its amino acid composition is likely to be linked to the minimum estimated thymine composition. In contrast, the present analysis showed that the amino acid composition of the thymine-rich TMDs tended to correlate with the maximum estimated thymine composition, whereas the thymine-poor IDRs correlated with the minimum estimated thymine composition. This result suggests that the amino acid composition features of the TMD are in the form pushed up by the nucleotide composition of the gene, whereas the amino acid composition features of the IDR are in the form pushed down by the nucleotide composition of the gene. Thus, it was suggested that the amino acid composition of TMDs and IDRs is largely determined by variations in the density of thymine originally present in the genome, rather than by selective pressures on amino acid composition. Furthermore, in Figure 3, the distribution of estimated thymine compositions for all species was restricted to a very narrow range, presumably because

genes use their thymine density to generate protein function, and inadvertent variations in thymine density may not be tolerated.

If the discussion in the previous paragraph is correct, then the ability to construct a functional amino acid composition on a protein from the biased nucleic acid composition on the gene by the function of the genetic code itself would be very important for the stable production of functional proteins. I will explain how I think in the following paragraph.

A protein consists of a sequence of 20 different amino acids, and the composition of the amino acids that make up its domain is 20 dimensions of rather high-dimensional information. Consequently, if evolution occurs by direct mutation of amino acids, the composition of the domains is very likely to be sparse and divergent due to the curse of high dimensionality. However, it is thought that life, through the genetic code, has realized a mechanism to generate amino acid sequences of high-dimensional information from the relatively low-dimensional nucleic acid composition information of genes consisting of only four types of nucleic acids, thus preventing the amino acid composition of protein functional domains such as TMD and IDR from diverging. If such an advantage really exists in the current genetic code, it must be the reason why the "universal genetic code" is shared by all living organisms.

From this perspective, the standard genetic code can be assumed to be optimized for the generation and translation of functional domains of proteins. Since this is a new explanation of the function of the genetic code, different from the existing frozen theory and the existing error minimization theory, I thought it might be better to call it the "optimized translation theory".

More than 50 years have passed since the genetic code was first analyzed, and although there have been various discussions about the origin and universality of the genetic code's formation, there has been no settled theory, and the debate continues to this day. The first explanation of the universality of the genetic code was given by Click and was called the frozen-accident scenario, the frozen theory [9]. More recently, the biased assignment of the genetic code has been explained by error minimization, i.e., robustness to mutation [10, 11]. Some other studies now suggest that the genetic code arose from morphological constraints on the anticodon of tRNAs. However, no explanation has yet been offered that is considered conclusive [12, 13].

If the optimized translation theory proposed here is correct, then the genetic code would have the ability to convert nucleic acid compositional bias into functional amino acid compositional bias. On the other hand, such a genetic code would seem to be designed for error minimization because of the corresponding nature of such a genetic code. Furthermore, if the generation of TMDs and IDRs is supported by the genetic code itself, it probably indicates that both TMDs and IDRs are the most essential elements for the formation and maintenance of life.

The present analyses showed that the amino acid compositions of the TMD and IDR were strongly correlated with the higher and lower estimated thymine composition in the gene, respectively, by the functional structure of the genetic code. However, in detail, the amino acid composition of the TMD correlated with the maximum estimated composition and that of the IDR correlated with the minimum estimated thymine composition. Therefore, for better visualization, in Figure 8 the proteins were plotted according to the maximum, minimum and average thymine composition estimates for each organism, and these plots were colored according to the proportions of TMDs and IDRs in each protein. The TMD fractions were better aligned in rows by maximum estimated thymine composition (max T), and the IDR fractions were better aligned by minimum estimated thymine composition (min T). In addition, the plot of IDR fractions suggests that the lower limits of thymine density in the genome of archaea and bacteria are higher than those of eukaryotes, which seem to inhibit rich IDR coding. Conversely, the expansion of nucleic acid composition diversion across the genome in eukaryotes may have enabled the achievement of a highly functional proteome that contains a large amount of IDRs. This finding is consistent with my previous report on differences in genomic complexity between organisms [8].

Figure 8a: Estimated nucleic acid compositions of proteins and their TMD fractions for each organism

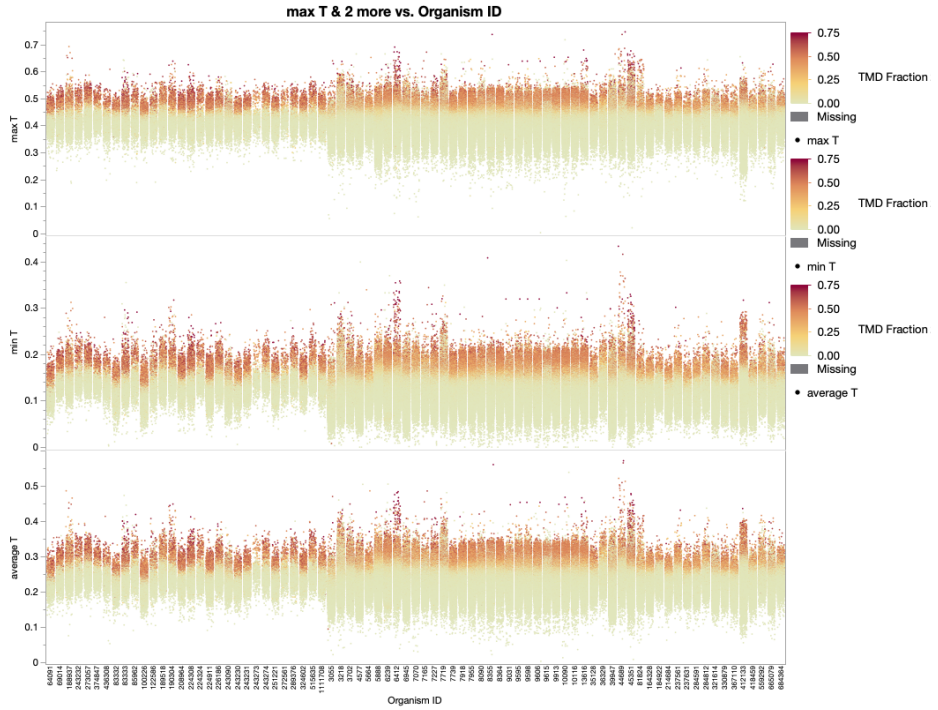
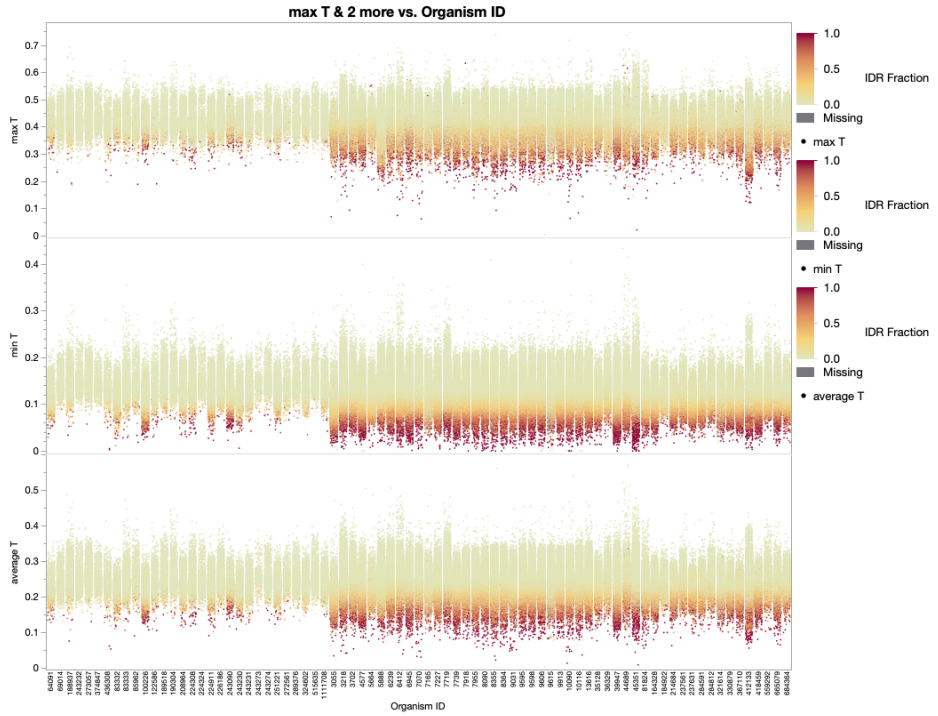


Figure 8b: Estimated nucleic acid compositions of proteins and their IDR fractions for each organism



The estimated nucleic acid compositions of the proteins and their TMD and IDR fractions are shown for each organism. The boundaries of the TMD fraction were best aligned at max T, and the boundaries of the IDR fraction were best aligned at min T. Max T; the maximum estimated thymine composition, min T; the minimum estimated thymine composition, average T: the average estimated thymine composition.

5. Conclusion

The results of the study using backward translation of the amino acid sequence of proteins from various organisms showed that the estimated thymine compositions of the TMD and IDR sequences are both different from those of other regions. Therefore, I concluded that these two functional domains can be generated from the thymine densities on the genes as projections onto the protein amino acid sequences. This suggests that the standard genetic code has a function of avoiding divergence of the higher dimensional amino acid composition of TMDs and IDRs by generating them from the lower dimensional thymine densities on the gene, and this advantage is assumed to be the reason why the standard genetic code is so common that it can be called universal. In this paper I propose this new explanation for the origin of the standard genetic code as "optimized translation theory".

6. References

1. 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
2. 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
3. "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
4. The UniProt Consortium. (2023) UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res. 51*:D523–D531. https://doi.org/10.1093/nar/gkac1052
5. 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
6. Vakirlis, N., Acar, O., Hsu, B., Castilho Coelho, N., Van Oss, S. B., Wacholder, A., Medetgul-Ernar, K., Bowman, R. W., II, Hines, C. P., Iannotta, J., Parikh, S. B., McLysaght, A., Camacho, C. J., O'Donnell, A. F., Ideker, T., & Carvunis, A.-R. (2020). De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. In *Nature Communications* (Vol. 11, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/s41467-020-14500-z
7. Efimov, V. M., Efimov, K. V., Kovaleva, V. Yu., & Matushkin, Yu. G. (2021). Principal Components of Genetic Sequences: Correlations and Significance. In *Mathematical Biology and Bioinformatics* (Vol. 16, Issue 2, pp. 299–316). Institute of Mathematical Problems of Biology of RAS (IMPB RAS). https://doi.org/10.17537/2021.16.299
8. 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
9. 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
10. Haig, D., & Hurst, L. D. (1991). A quantitative measure of error minimization in the genetic code. Journal of Molecular Evolution, 33(5), 412–417. https://doi.org/10.1007/BF02103132
11. Haig, D., & Hurst, L. D. (1991). A quantitative measure of error minimization in the genetic code. In *Journal of Molecular Evolution* (Vol. 33, Issue 5, pp. 412–417). Springer Science and Business Media LLC. https://doi.org/10.1007/bf02103132
12. 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
13. Tourancheau, A. B., Tsao, N., Klobutcher, L. A., Pearlman, R. E., & Adoutte, A. (1995). Genetic code deviations in the ciliates: evidence for multiple and independent events. The EMBO Journal, 14(13), 3262–3267. https://doi.org/10.1002/j.1460-2075.1995.tb07329.x