The Standard Genetic Code Predominantly Assigns Uracil-Containing Codons To Amino Acids Enriched in Transmembrane Domains and Uracil-Free Codons To Amino Acids Enriched in Intrinsically Disordered Regions

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

Abstract

All organisms on Earth share a nearly identical genetic code, and the most typical one is called the standard genetic code. In previous research, based on the results of studies of possible inverse translation of the genetic code applied to various protein amino acid sequences, I proposed the idea that the genetic code uses local thymine density in the gene sequence to determine the presence of transmembrane domains (TMDs) or intrinsically disordered regions (IDRs) on proteins. However, I had not performed an analysis to determine how each specific codon-amino acid correspondence supported this hypothesis.

In this study, I examined the specific difference in the amino acid composition of TMDs and IDRs of different organisms by comparing the ratios between the average amino acid residue compositions of TMDs, IDRs, and the total sequence of each protein by organism. The results showed that the difference ratios between TMDs to total, IDRs to total for all 20 amino acids were almost inversely different between two regions and were well consistent across organisms. This consistency suggests that, regardless of species, TMDs and IDRs each have distinct characteristics in their amino acid composition. Furthermore, a comparison of these results with the codons corresponding to each amino acid in the genetic code revealed that the standard genetic code predominantly assigns uracil-containing codons to amino acids enriched in transmembrane domains and uracil-free codons to amino acids enriched in intrinsically disordered regions.

In my other recent study, I showed that TMD-rich and IDR-rich proteins are consistently two of the most statistically distinct domains/regions of amino acid composition of the proteome in any organism, and combined with the previous research finding that the genetic code has a structure in which TMDs and IDRs are encoded by gene sequences of each specific nucleotide composition, I concluded that this may explain why the standard genetic code is universal. The results of the current study show that the differentiation function of the genetic code is based on an elaborate simultaneous coordination of codon-amino acid correspondence. This finding supports the idea that the structure of the standard genetic code, which is influenced by the commonality of TMDs/IDRs, is unlikely to be a product of mere chance and at least has a purpose in differentiating these regions. This finding should provide a crucial insight into the undiscovered origins of the standard genetic code as the statistically largest piece of its puzzle. But at the same time, the piece must be quite small in the over-complexity of its overall mystery.

Keywords: genetic code, transmembrane domains, intrinsically disordered regions, purpose, origin

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

1. Background

All organisms on Earth share a nearly identical genetic code, the most typical of which is called the standard genetic code [1]. It is known that the codon-amino acid correspondences in the genetic code are not random [2]. For example, codons with uracil as the second letter consistently encode hydrophobic amino acids [2]. However, the meaning of such patterns has not been clearly explained. Although some explanations, such as error minimization and mutational robustness, suggest that hydrophobicity is maintained despite mutations in the first and third letters, the observed robustness is not at a "fully optimized" level and there is still room for further optimization [3], suggesting the need for a more comprehensive understanding of the design principles of the genetic code.

In previous research, based on the results of studying the range of possible gene nucleic acid compositions generated by inverse computation of the genetic code applied to various protein amino acid sequences, I proposed the idea that the genetic code uses local thymine density in gene sequences to determine the presence of transmembrane domains (TMDs) or intrinsically disordered regions (IDRs) in protein sequences [4]. However, I had not performed an analysis to determine how each specific codon-amino acid correspondence supported this hypothesis.

In this study, I examined the specific characteristics of the amino acid compositions of TMDs and IDRs of different organisms by comparing the ratios between the average amino acid residue compositions of TMDs, IDRs, and the total sequences of each protein by organism proteome. In addition, I compared these results with the codons corresponding to each amino acid in the genetic code to investigate how the standard genetic code is structured.

2. Materials and Methods

For this study, the 'reference proteome' dataset published by the European Bioinformatics Institute [5] was used. The original dataset contained 1,023,125 protein entries from 79 species representing all three domains of life. From this dataset, 4,121 protein entries were excluded due to discrepancies with the UniProtKB database or due to uncertain or unusual information in their sequences or annotations [6], resulting in a total of 1,019,004 proteins used for the analysis.

First, the amino acid residue composition of each protein sequence was calculated from the complete sequence or only the TMD or IDR sequences of each protein. This was done using the amino acid sequence from the 'reference proteomes' provided in FASTA format, together with TMD and IDR annotation information from the UniProtKB database [5, 6]. Each amino acid residue composition was calculated by first counting the target amino acid residues for each sequence, and then dividing the number of target amino acids by the total number of total residues. As a result, each amino acid composition took a value between 0 and 1, and their sum for each sequence was 1.

Second, the average compositions of TMDs, IDRs, and total protein were calculated for each of the 79 species. The natural log ratios of TMDs to total and IDRs to total were calculated and designated as "abundance in TMDs (Ln)" and "abundance in IDRs (Ln)", respectively.

Third, using the abundance values of TMDs and IDRs, each target amino acid was plotted on a scatter plot to visualize the general trends of its abundance in TMDs and IDRs across species. In addition, the mean values for each amino acid of all 79 species were calculated to obtain representative values across species.

Fourth, to examine the correlation between each amino acid and the corresponding abundance values in TMDs and IDRs, these values are listed in a table in order of their difference, with cells colored according to each value.

Fifth and finally, by incorporating the TMD/IDR difference values into the genetic code table, I analyzed how the codon-amino acid correspondence is structured.

Microsoft® Excel for Mac v16.80 (Microsoft Corporation, Redmond, WA, USA) was used for computational analysis, including table generation. JMP® 17.2.0 (SAS Institute Inc., Chicago, IL, USA) was used for statistical analysis and to generate graphs and figures.

3. Results

3.1. Target Proteins

Table 1 lists the number of target proteins for each organism.

Table 1: The number of target proteins by the organisms

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	4406	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	43666
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	49311
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

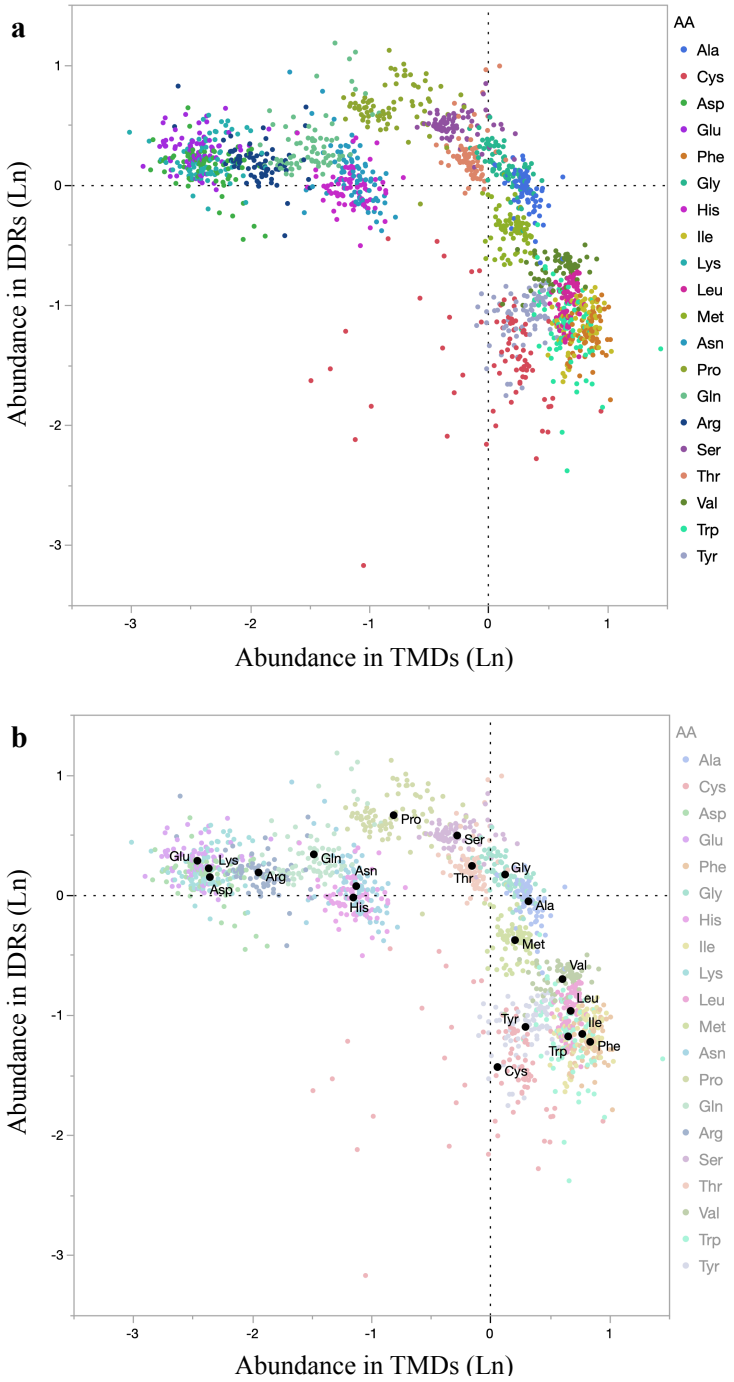
Organisms in the 'reference proteomes' dataset are listed. Each color bar in the rightmost column indicates its number by its length and is colored according to the domain to which it belongs.

3.2. Abundance of Each Amino Acid in TMDs and IDRs

Figure 1a shows plots for each species by abundance in TMDs and abundance in IDRs for each amino acid. The plots differ between amino acids, but it can be seen that the same amino acids cluster relatively close together.

In **Figure 1b**, the plots for each amino acid are superimposed on the plots in Figure 1a, with the average of all target organisms for each amino acid. These averages are also shown in Table 2.

Figure 1: Abundance plot of each amino acid in TMDs and IDRs



a Each plot color indicates the individual amino acids from a total of 79 organisms, except for cystine (Cys), where there was one bacterium that does not have Cys in its IDRs, so Cys is from 78 organisms. **b** Black plots indicate the mean values of the amino acid listed.

3.3. Differences in the abundance of amino acids and their codon correspondence

Table 2 lists the abundances in TMDs and IDRs, along with their differences, in order of the value of the difference. The differences in the right column are color-coded for clarity: red indicates larger values with 0 as the boundary, while blue indicates smaller values. In addition, each abundance in the left two columns is highlighted in green for larger values for better visibility. As a result, these abundances show almost opposite trends in TMDs and IDRs.

Figure 2 visually represents the differences in abundance in TMDs and IDRs from Table 2 by superimposing these differences on the codon-amino acid relationships in the standard genetic code. The figure illustrates that amino acids corresponding to uracil-free codons, shown within the black boxes, are colored almost blue and are almost all composed of amino acids that are enriched in IDRs than in TMDs. Conversely, amino acids corresponding to uracil-containing codons, shown outside the black boxes, are mostly colored red and consist predominantly of amino acids that are enriched in TMDs than in IDRs.

In addition, all amino acids corresponding to uracil-free codons inside the black boxes in Figure 2 have smaller difference values and are all found inside the black box in Table 2. This finding highlights a significant relationship between codon composition and the difference in amino acid distribution between TMDs and IDRs.

Table 2: Abundances and Difference (Ln)

AA	Abundance in TMDs (Ln)	Abundance in IDRs (Ln)	Difference
Phe	0.839	-1.225	2.063
Ile	0.772	-1.159	1.931
Trp	0.654	-1.179	1.832
Leu	0.675	-0.967	1.641
Cys	0.060	-1.434	1.494
Tyr	0.295	-1.100	1.395
Val	0.606	-0.702	1.308
Met	0.208	-0.377	0.585
Ala	0.320	-0.054	0.374
Gly	0.124	0.169	-0.044
Thr	-0.154	0.241	-0.395
Ser	-0.278	0.494	-0.772
His	-1.150	-0.022	-1.128
Asn	-1.126	0.072	-1.198
Pro	-0.811	0.663	-1.474
Gln	-1.482	0.337	-1.818
Arg	-1.946	0.186	-2.132
Asp	-2.355	0.147	-2.502
Lys	-2.364	0.221	-2.584
Glu	-2.459	0.283	-2.743

The abundance of each amino acid in each region and their differences are shown. The differences within the black box indicate the amino acids corresponding to the uracil-free codons. Each cell is colored according to its value.

Figure 2: Abundance Difference on the Genetic Code Table

	U		A		G		C		
U	UUU	Phe	UAU	Tyr	UGU	Cys	UCU	Ser	U
	UUA	Leu	UAA	Stop	UGA	Stop	UCA	Ser	A
	UUG	Leu	UAG	Stop	UGG	Trp	UCG	Ser	G
	UUC	Phe	UAC	Tyr	UGC	Cys	UCC	Ser	C
A	AUU	Ile	AAU	Asn	AGU	Ser	ACU	Thr	U
	AUA	Ile	AAA	Lys	AGA	Arg	ACA	Thr	A
	AUG	Met	AAG	Lys	AGG	Arg	ACG	Thr	G
	AUC	Ile	AAC	Asn	AGC	Ser	ACC	Thr	C
G	GUU	Val	GAU	Asp	GGU	Gly	GCU	Ala	U
	GUA	Val	GAA	Glu	GGA	Gly	GCA	Ala	A
	GUG	Val	GAG	Glu	GGG	Gly	GCG	Ala	G
	GUC	Val	GAC	Asp	GGC	Gly	GCC	Ala	C
C	CUU	Leu	CAU	His	CGU	Arg	CCU	Pro	U
	CUA	Leu	CAA	Gln	CGA	Arg	CCA	Pro	A
	CUG	Leu	CAG	Gln	CGG	Arg	CCG	Pro	G
	CUC	Leu	CAC	His	CGC	Arg	CCC	Pro	C

Each amino acid difference is superimposed on the standard genetic code table. The colors of the cells are the same as in Table 2. Black boxes indicate uracil-free codons and their corresponding amino acids.

4. Discussions

The current and most accepted explanation for the origin of the genetic code is that it arose from and/or is fixed by the chemical constraints between codons and amino acids and/or their robustness to mutation [2]. However, not only have subsequent analyses reported deviations from the standard genetic code in mitochondria and certain species [7, 8], but some studies suggest that the current code is not ultimately an optimized form and that a more robust genetic code may be possible [3]. This suggests that current explanations may not adequately explain the structure of the genetic code.

On the other hand, in my recent study I showed that TMD-rich and IDR-rich proteins are consistently two of the most statistically distinct domains/regions in the amino acid composition of the proteome of any organism [9]. Combined with previous findings that the genetic code is structured such that TMDs and IDRs are encoded by gene sequences with specific and distinct nucleotide compositions [4], I had concluded that they might explain the universality of the standard genetic code [9].

Originally, both TMDs and IDRs are two of the domains/regions responsible for function on a protein, and their properties are thought to be generated primarily by their characteristic amino acid composition. For TMDs, a correlation between thymine on the gene and membrane proteins has already been reported [10, 11], but for IDRs, their correlation with the nucleic acid composition of the gene has never been reported except by me [4, 9, 12, 13]. Furthermore, there have been no reports of TMDs or IDRs being associated with the genetic code, and the first such report was made only in my own series of reports [4, 9, 12, 13]. This means that I am either the only one who is right or, conversely, the only one who is wrong.

Recognizing these gaps in our understanding, I undertook a thorough investigation of the relationship between the genetic code and protein domains. This included examining how specific nucleotide compositions in gene sequences might influence the formation and function of TMDs and IDRs in proteins.

Figure 2 shows that almost all amino acids corresponding to uracil-free codons are more abundant in IDRs. Conversely, most amino acids corresponding to uracil-containing codons are more abundant in TMDs. In addition, a closer look reveals that in the standard genetic code, while most amino acids corresponding to uracil-containing codons are those that are more abundant in TMDs, amino acids corresponding to uracil-free codons are completely consistent with those that have smaller differences between TMDs and IDRs [inside the black box in Table 2]. These results suggest that while gene sequences with high thymine (corresponding to uracil in the codons) do not always lead to an amino acid composition of TMDs, gene sequences with low thymine are structurally and always inclined to have an amino acid composition of IDRs. Therefore, there must be a design rule in the standard genetic code, but the rule explaining this structure was thought to be rather complex, not so simple. From the above, I concluded that the standard genetic code has an elaborate coordination structure of codon-amino acid correspondences for the differentiation of TMDs and IDRs. And such a sophisticated coordination is unlikely to have arisen by chance, leading to the plausible conclusion that the design of the genetic code has a clear purpose in differentiating TMDs and IDRs.

Of course, functional domains in the proteome are not limited to TMDs and IDRs. For example, proteins that bind to nucleic acids such as DNA and RNA are thought to require significant placement of basic amino acid residues, which are predominantly encoded by adenine-containing codons in the genetic code. Therefore, the differentiation of functional domains influenced by nucleic acid composition is not exclusive to TMDs and IDRs; similar synthetic correlations should exist for other domains. On the other hand, statistical analysis of the amino acid composition of the entire proteome using principal component analysis extracted proteins rich in TMDs and IDRs [9]. Since principal component analysis is a method to extract the most significant statistical features in order, the extraction of TMDs and IDRs on the first and second principal components suggests that their differentiation in the genetic code may be the most significant from a statistical point of view.

5. Conclusion

The results of the current study show that the differentiation function of the genetic code is based on an elaborate simultaneous coordination of codon-amino acid correspondence, supporting the idea that the structure of the standard genetic code has a purpose to differentiate functional protein regions such as TMDs or IDRs. This finding should provide a crucial insight into the undiscovered origins of the standard genetic code as the statistically largest piece of its puzzle. But at the same time, the piece must be quite small in the over-complexity of its overall mystery.

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. Koonin, E. V., & Novozhilov, A. S. (2008). Origin and evolution of the genetic code: The universal enigma. In IUBMB Life (Vol. 61, Issue 2, pp. 99–111). Wiley. https://doi.org/10.1002/iub.146

3. Wnętrzak, M., Błażej, P., Mackiewicz, D., & Mackiewicz, P. (2018). The optimality of the standard genetic code assessed by an eight-objective evolutionary algorithm. In BMC Evolutionary Biology (Vol. 18, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1186/s12862-018-1304-0

4. 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

5. "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

6. 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

7. 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

8. 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 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

9. Esumi, G. (2023). Statistical Extremes of Amino Acid Residue Composition of the Proteome Proteins Can Explain the Origin of the Universality of the Genetic Code. Jxiv. https://doi.org/10.51094/jxiv.575

10. Prilusky, J., & Bibi, E. (2009). Studying membrane proteins through the eyes of the genetic code revealed a strong uracil bias in their coding mRNAs. In Proceedings of the National Academy of Sciences (Vol. 106, Issue 16, pp. 6662–6666). Proceedings of the National Academy of Sciences. https://doi.org/10.1073/pnas.0902029106

11. 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

12. Esumi, G. (2023). The α-helical transmembrane domains and intrinsically disordered regions on the human proteins are coded for by the skews of their genes' nucleic acid composition with the "universal" assignment of the genetic code table. Jxiv. https://doi.org/10.51094/jxiv.247

13. 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