The TA Skew of a Gene Primarily Determines the Type of Protein, Such as Membrane Protein or Intrinsically Disordered Protein

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

Abstract

Proteins differ in function and cellular localization according to their amino acid composition; however, there are few reports on how these characteristic compositions are organized.

Principal component analysis of the amino acid composition of all proteins in the human proteome and plotting of all proteins by their first and second principal components revealed that proteins with high fractions of α-helical transmembrane domains and those with high fractions of intrinsically disordered regions were located at each extreme of the plot. At the same time, each functional domain fraction corresponded primarily to the high and low TA skew of the gene nucleotide composition.

The codon corresponding to each amino acid in the genetic code consists of four nucleic acids, but the nucleic acid composition in the codons for each amino acid is initially skewed. Therefore, the amino acid composition of a protein is inevitably affected by the nucleic acid composition of its genes. However, there are few reports on the consequences of this effect. In contrast, the present study showed that the largest source of diversity in the amino acid composition of proteins is the diversity in the nucleic acid composition of their genes, such as TA skew and GC content. Furthermore, this study showed that TA skew plays an important role in determining protein properties. Finally, I conclude that both the TA skew of the gene and the skewed assignment of the genetic code work together to maintain the correct properties of proteins in the proteome.

Keywords: TA skew, nucleic acid composition, transmembrane domain, intrinsically disordered region, genetic code

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

1. Background

Each of the 20 amino acids that make up a protein has different chemical properties; as a result, proteins have different properties and functions depending on the amino acid composition of their amino acid residues. However, because this organization of amino acid composition in proteins is generally considered to be merely the result of random genetic mutations and the pressure of natural selection, little thought has been given to how it is actually organized.

In an earlier Japanese paper, I showed that the proportions of transmembrane domains and intrinsically disordered regions in proteins are positively and negatively correlated, respectively, with their gene's TA skew, an index of a gene's nucleic acid composition. And showed that TA skew may play a key role in the synthesis of functional proteins [1]. In this paper, I revisit previous results and discuss the significance of TA skew in protein synthesis.

2. Materials and methods

In a previous study, I performed principal component analysis on the amino acid composition of proteins in the Human Proteome Dataset and examined the relationships between variation in their amino acid composition, variation in their protein properties, and variation in the nucleic acid composition of their genes. However, a dataset containing all these data was not available in the existing public databases, so I generated a new dataset by matching the amino acid sequences and gene coding sequences from NCBI with the protein domain information from UniProt using the RefSeqID and the length (residue count) of each protein [2, 3]. This generated a dataset of 25,095 proteins for analysis [1]. Principal component analysis was then performed on the amino acid composition of all proteins in the human proteome dataset, and all these proteins were plotted using their first and second principal components. To examine the relationship between these protein properties and their plot placement, each plot on the sheet was colored according to the indices of nucleic acid composition and residue proportions (fractions) of α-helical transmembrane domains (α-TMD) and intrinsically disordered regions (IDR).

In the previous and current study, I define three indices: TA skew = (T-A)/(T+A), GC skew = (G-C)/(G+C), and GC content = (G+C)/(T+A+G+C), where each capital letter indicates the number of each nucleic acid in the sequence window. As a result, TA skew and GC skew take values between -1 and 1, and GC content takes values between 0 and 1.

And in this study, I used Microsoft® Excel for Mac v16.74 (Microsoft Corporation, Redmond, WA, USA) to generate amino acid compositions of proteins, nucleic acid composition indices of genes, and other calculation results. I also used JMP® 17.1.0 (SAS Institute Inc., Chicago, IL, USA) for statistical analyses and to generate graphs and figures.

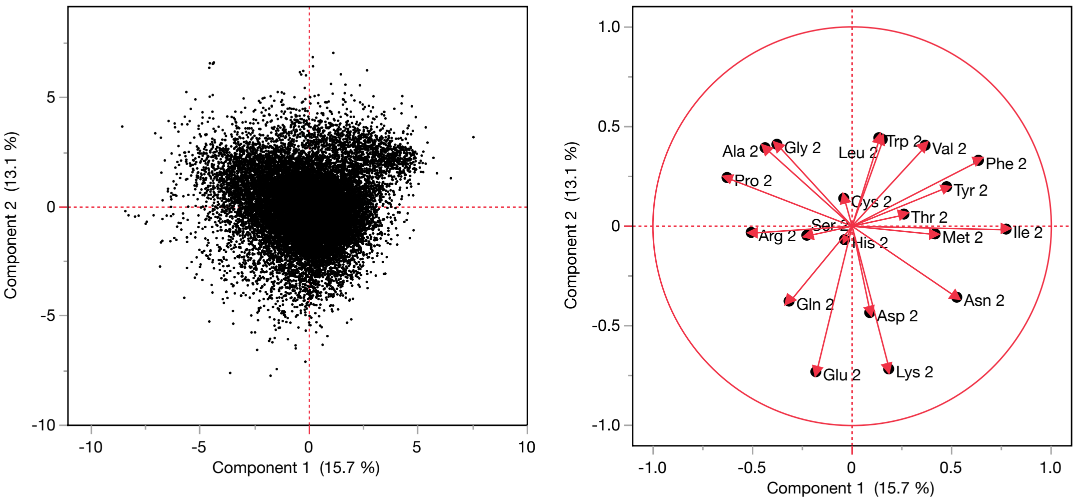
3. Results

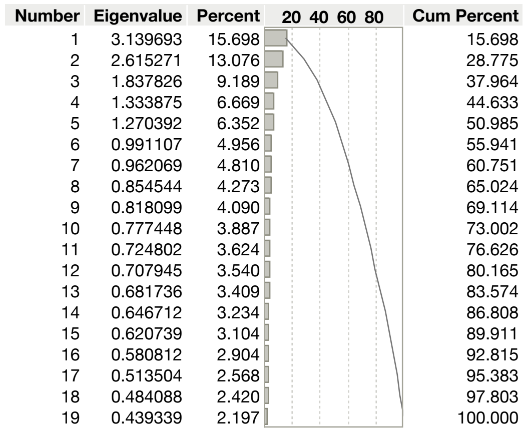
3.1. Principal component analysis

Figure 1 shows the results of principal component analysis of the amino acid composition of all proteins in the human proteome dataset. As a result, the contribution of the first principal component was 15.7% and that of the second principal component was 13.1%. The plots of the first and second principal components showed an overall triangular distribution and also appeared to be composed of two clusters, a large cluster (lower left) and a small cluster (upper right).

Figure 1

Principal component analysis plot of the amino acid composition of all proteins in the human proteome





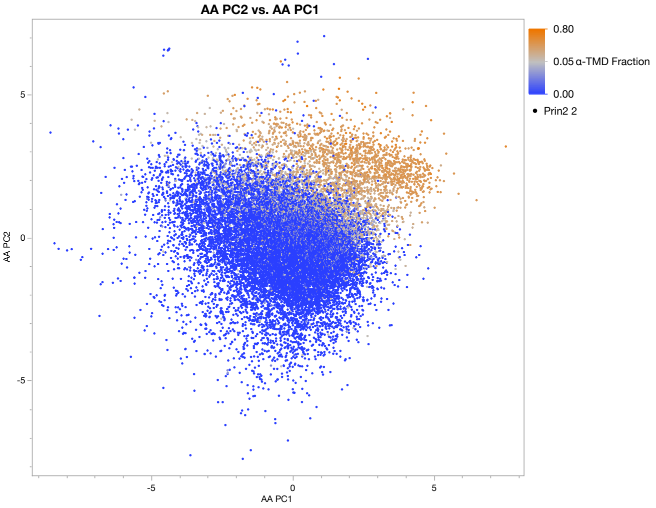
Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	3.139693	15.698		15.698
2	2.615271	13.076		28.775
3	1.837826	9.189		37.964
4	1.333875	6.669		44.633
5	1.270392	6.352		50.985
6	0.991107	4.956		55.941
7	0.962069	4.810		60.751
8	0.854544	4.273		65.024
9	0.818099	4.090		69.114
10	0.777448	3.887		73.002
11	0.724802	3.624		76.626
12	0.707945	3.540		80.165
13	0.681736	3.409		83.574
14	0.646712	3.234		86.808
15	0.620739	3.104		89.911
16	0.580812	2.904		92.815
17	0.513504	2.568		95.383
18	0.484088	2.420		97.803
19	0.439339	2.197		100.000

3.2. Transmembrane domains

Figure 2 shows the principal component plot of the amino acid composition of the proteins, colored according to the fractions of α-helical transmembrane domains on the proteins. This shows that the smaller cluster in the upper right is composed of membrane proteins with high proportions of transmembrane domains.

Figure 2

Principal component plot of the amino acid composition of the proteins, colored with the fractions of α-helical transmembrane domains

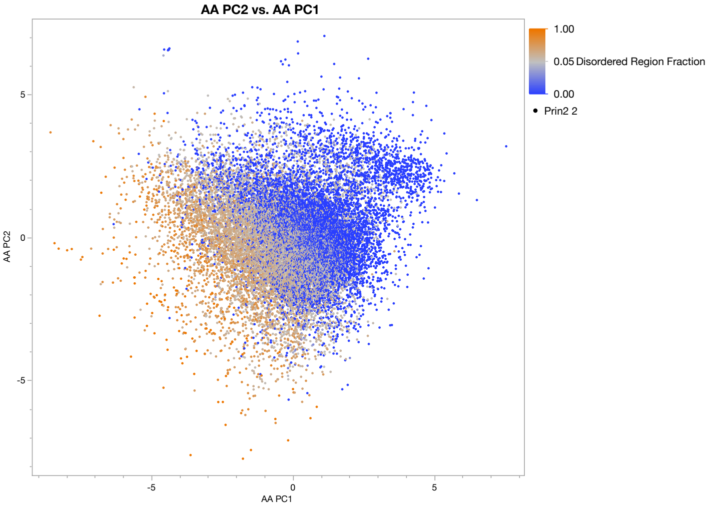


3.2. Intrinsically disordered regions

Figure 3 shows the principal component plot of protein amino acid composition colored with the fractions of intrinsically disordered regions on the proteins. It can be seen that proteins with high fractions of intrinsically disordered regions are plotted at the opposite extreme from those with higher proportions of α-helical transmembrane domains.

Figure 3

Principal component plot of protein amino acid composition colored with the fractions of intrinsically disordered regions

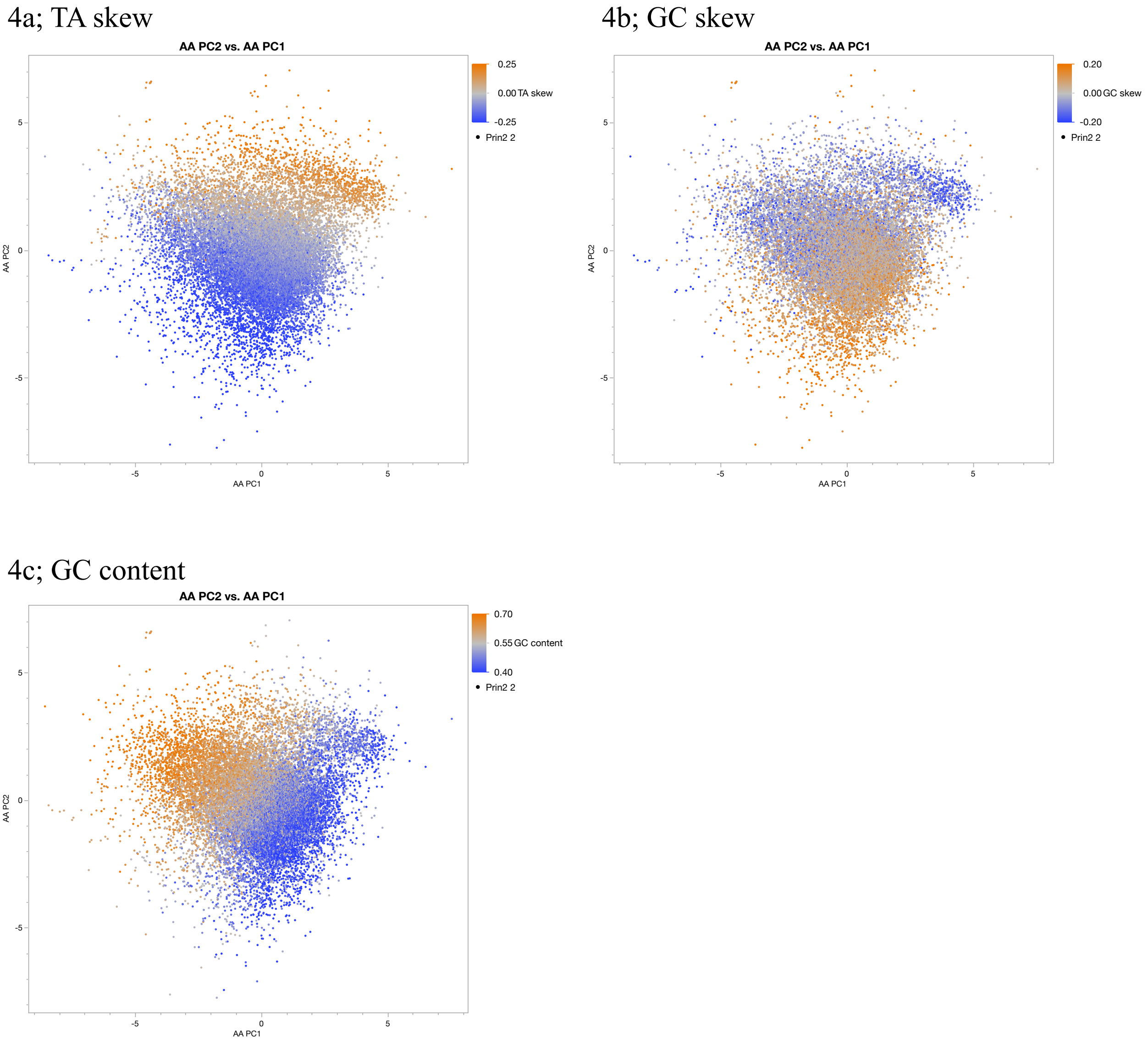


3.3. Nucleic acid compositions and their indices

Figure 4a, 4b, 4c show the principal component plot of protein amino acid composition colored by their TA skew, GC skew, and GC content, respectively. The coloring of TA skew and GC content were clearly separated on the plots, but the separation for GC skew was weaker compared to the other two. Moreover, the color separation of TA skew has the same direction as that of transmembrane domain and intrinsically disordered regions mentioned in the previous subsection [Figure 2, Figure 3].

Figure 4

Principal component plot of protein amino acid composition colored with each nucleic acid composition index (4a; TA skew, 4b; GC skew, 4c; GC content)



4a; TA skew

4b; GC skew

4c; GC content

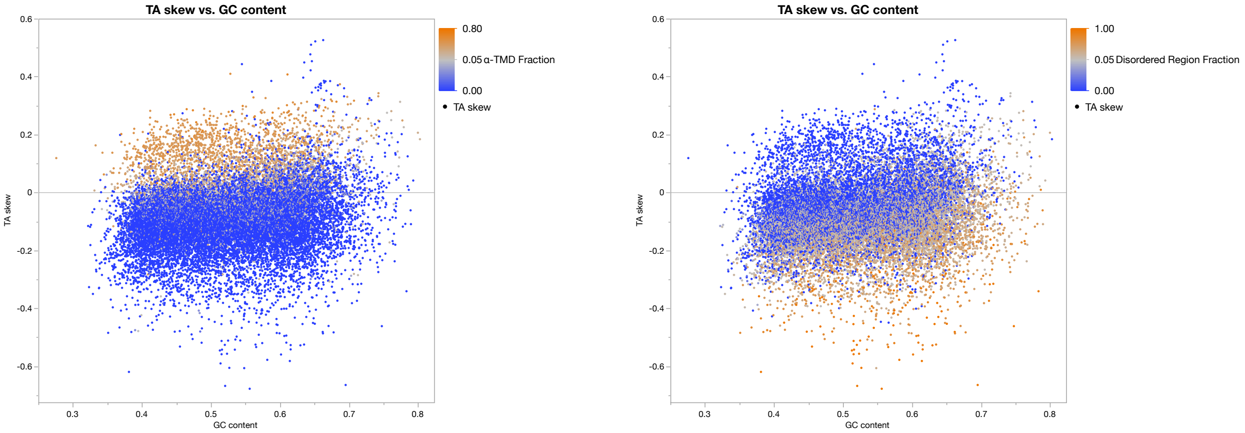
3.4. Relationships between these variations

Figure 5a, 5b show the plot of TA skew and GC content colored according to the fractions of α-helical transmembrane domains (α-TMD) and intrinsically disordered regions (IDR) on the proteins, respectively. The fraction of α-TMD is higher in the proteins with higher TA skew, and that of IDR is higher in the proteins with lower TA skew. (Figure 5a, 5b are essentially the same as those published in previous Japanese paper [1].)

Figure 6 shows the relative correlations between each nucleic acid index and two domain fractions. Among these correlations, that of TA skew to α-TMD and that of TA skew to IDR were the two largest (orange boxes).

Figure 5

Plots of TA skew and GC content, colored according to the fractions of protein domains (5a; α-TMD fraction, 5b; IDR fraction)

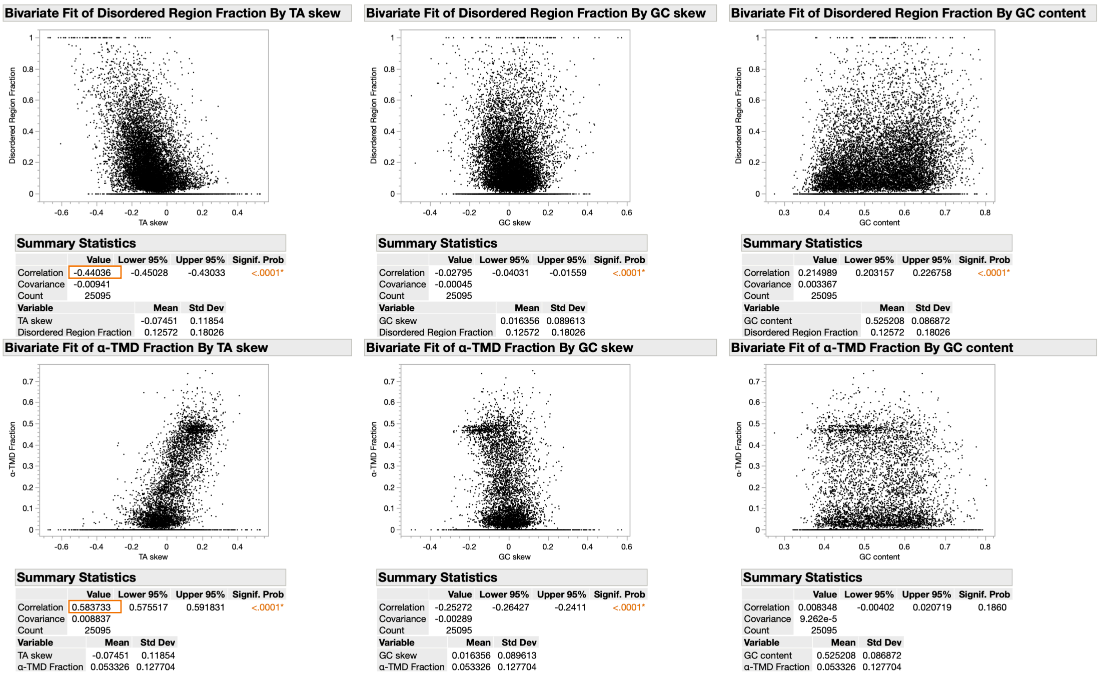


5a; α-TMD fraction

5b; IDR fraction

Figure 6

Correlations between each nucleic acid index and two domain proportions



Bivariate Fit of Disordered Region Fraction By TA skew

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	-0.44036	-0.45028	-0.43033	<.0001*
Covariance	-0.00941			
Count	25095			

Variable	Mean	Std Dev
TA skew	-0.07451	0.11854
Disordered Region Fraction	0.12572	0.18026

Bivariate Fit of Disordered Region Fraction By GC skew

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	-0.02795	-0.04031	-0.01559	<.0001*
Covariance	-0.00045			
Count	25095			

Variable	Mean	Std Dev
GC skew	0.016356	0.089613
Disordered Region Fraction	0.12572	0.18026

Bivariate Fit of Disordered Region Fraction By GC content

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.214989	0.203157	0.226758	<.0001*
Covariance	0.003367			
Count	25095			

Variable	Mean	Std Dev
GC content	0.525208	0.086872
Disordered Region Fraction	0.12572	0.18026

Bivariate Fit of α-TMD Fraction By TA skew

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.583733	0.575517	0.591831	<.0001*
Covariance	0.008837			
Count	25095			

Variable	Mean	Std Dev
TA skew	-0.07451	0.11854
α-TMD Fraction	0.053326	0.127704

Bivariate Fit of α-TMD Fraction By GC skew

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	-0.25272	-0.26427	-0.2411	<.0001*
Covariance	-0.00289			
Count	25095			

Variable	Mean	Std Dev
GC skew	0.016356	0.089613
α-TMD Fraction	0.053326	0.127704

Bivariate Fit of α-TMD Fraction By GC content

Summary Statistics

	Value	Lower 95%	Upper 95%	Signif. Prob
Correlation	0.008348	-0.00402	0.020719	0.1860
Covariance	9.262e-5			
Count	25095			

Variable	Mean	Std Dev
GC content	0.525208	0.086872
α-TMD Fraction	0.053326	0.127704

4. Discussions

The codon corresponding to each amino acid in the genetic code consists of four nucleic acids, but the nucleic acid composition in the codons for each amino acid is initially skewed. For example, all codons whose second letter in the codon is T (U) correspond to hydrophobic amino acids. Therefore, the amino acid composition of a protein is inevitably influenced by the nucleic acid composition of its genes. However, there are few reports on the consequences of this effect.

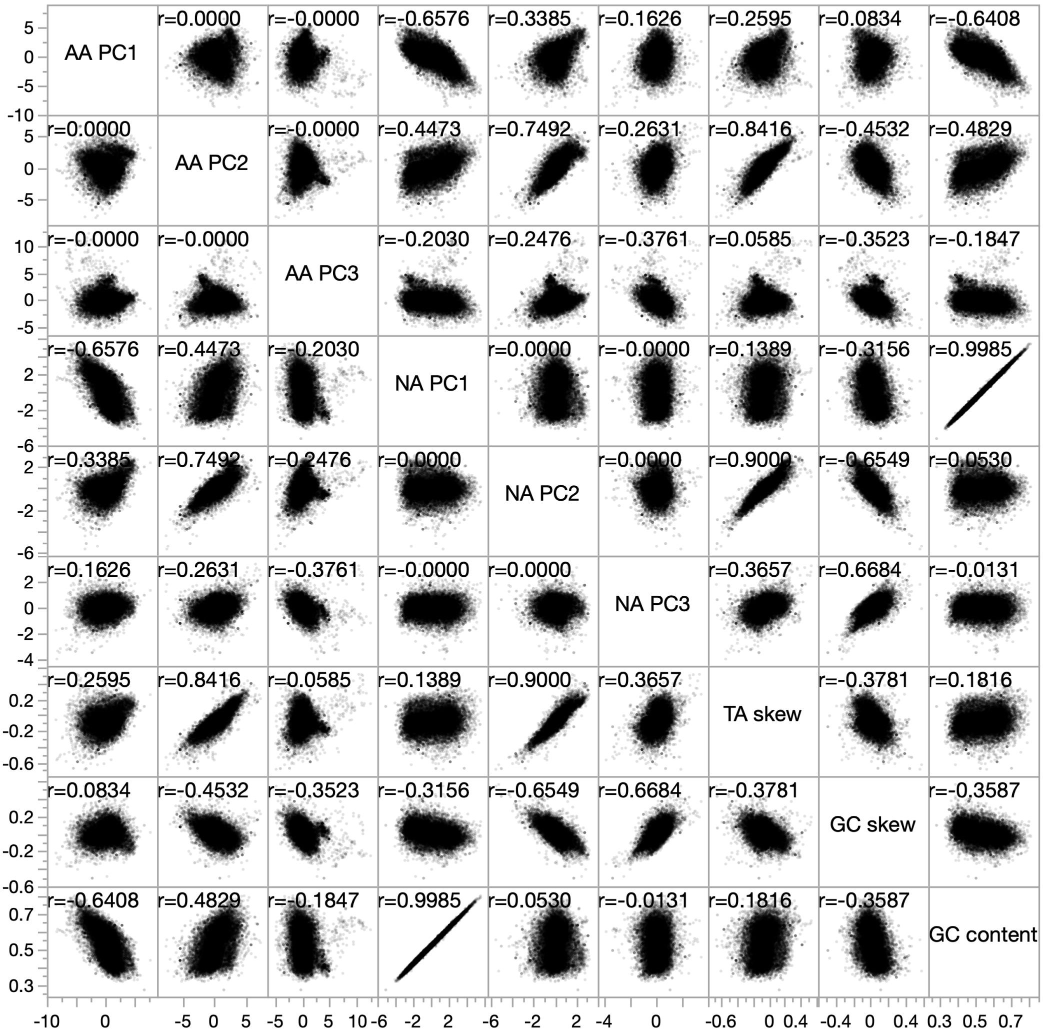
The transmembrane domain is a functional domain found in all membrane proteins, and in particular the α-helical transmembrane domain consists of a continuous sequence of hydrophobic amino acid residues. If this sequence were organized only by random genetic mutations, it would require patience and trial and error to create, just as it is not easy to line up certain markers in a game of cards. On the other hand, intrinsically disordered regions have attracted attention in recent years as domains on proteins that do not adopt a specific three-dimensional structure but change shape depending on the situation. It is known that this domain is characterized by being composed of relatively hydrophilic amino acids. Therefore, it has been speculated that it may not be easy for life to encode such a sequence with a skewed amino acid composition to produce a protein.

In this study, I showed that the first and second principal components of the amino acid composition of proteome proteins, the first and second largest diversity direction vectors, are primarily correlated with the nucleic acid composition indices of protein genes. Therefore, I concluded that the largest source of diversity in the amino acid composition of proteins is the nucleic acid composition of their genes. And since TA skew and not GC content correlates fractions of both α-TMD and IDR, TA skew should play a key role in determining these two domains. Since the codon correspondence in the genetic code is initially skewed, I then conclude that both the TA skew of the gene and the skewed assignment of the genetic code work together to maintain the correct proportions and properties of proteome proteins.

Figure 7 shows the correlations and coefficients of the first through third principal components of amino acid composition of each protein in the human proteome, the first through third principal components of nucleic acid composition of these genes, and the three indices of gene nucleic acid composition. The first principal component of amino acid composition correlates with the first principal component of nucleic acid composition and GC content, and the second principal component of amino acid composition correlates with the second principal component of nucleic acid composition and TA skew. From these results, I concluded that amino acid composition, nucleic acid composition, and nucleic acid indices are all correlated with each other.

Figure 7

Correlations and coefficients of the amino acid and nucleic acid composition principal components and three indices



In the present study, among the three indices of nucleotide composition, GC content and TA skew correlated primarily with the first and second principal components of amino acid composition, respectively. However, the current results indicate that only TA skew determines these protein domains. So why is TA skew alone a determinant of protein domains, while GC content is not? In a previous study, I reported that the codon usage bias in the genetic code is primarily used to compensate for the GC content of each gene [4]. In other words, the variation in the GC content of a gene is compensated by the redundancy of the genetic code. Therefore, the diversity of GC content was eliminated and the TA skew was used for domain determination.

In the present analysis, I analyzed not only the relationship between amino acid composition and protein properties, but also their relationship with the nucleotide composition of gene sequences. However, there was no publicly available database that would provide information on the correlation between the protein properties and gene sequences. The lack of such a database may be the reason why the corresponding relationship between protein properties and gene nucleotide composition had not yet been analyzed.

Finally, in this study the results of the analysis are presented only for the human proteome. However, although not presented here, the phenomenon of TA skew determining transmembrane domains has been confirmed in other organisms [5]. Since almost all organisms synthesize proteins using the same 20 amino acids and genetic code, the correspondence between nucleic acid composition and protein properties is likely to be similar in species from all three taxonomic domains. I believe this explains the origin of the universality of the universal genetic code.

5. Conclusion

In this paper, I showed that the TA skew of a gene primarily corresponds to the type of protein, such as membrane proteins or intrinsically disordered proteins. In addition, considering that the genetic code assignments are initially skewed, I concluded that both the TA skew of the gene and the correspondingly skewed genetic code work together to maintain the correct properties of proteome proteins.

6. References

1. 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
2. Genome dataset, “Homo sapiens (human) / Genome assembly T2T-CHM13v2.0”. NCBI website. https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_009914755.1/
3. Reference proteome, “Proteomes · Homo sapiens (Human)”. UniProt website. https://www.uniprot.org/proteomes/UP000005640
4. 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
5. Esumi, G. (2022). Synthesis assistance of transmembrane domains is a fundamental function of the genetic code table assignment. *Jxiv*. https://doi.org/10.51094/jxiv.139