# Not Only Residue Amino Acid Composition but Also Gene Thymine–Adenine Balance Reflect Protein Hydropathy

**Esumi, Genshiro**

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

**Abstract**

Kyte and Doolittle's landmark study established the concept that a protein's hydropathy governs its conformation and membrane-spanning regions, and they also demonstrated that this hydropathy can be estimated by applying coefficients to the amino acid residue composition of the protein sequence. In contrast, the possibility of estimating protein hydropathy from the nucleotide composition of its gene sequence has rarely been explored. In my previous study, I showed that the balance of thymine and adenine in protein genes, termed "TA skew," correlates positively with the proportion of hydrophobic transmembrane domains (TMD) and negatively with that of hydrophilic intrinsically disordered regions (IDR). Therefore, I hypothesized that a gene's TA skew correlates with the hydropathy of its encoded protein sequence.

To test this hypothesis, I revisited the six example proteins examined in Kyte and Doolittle's original study to determine whether the TA skew of their gene sequences corresponds to their hydropathic indices and the documented structural features of their corresponding residue sequences. Furthermore, using sufficiently large protein datasets, I analyzed whether each gene's TA skew correlates with the GRAVY score (the average hydropathy of each entire protein) and with the proportions of two distinct protein domains (TMDs and IDRs).

Analysis of the proteins from that landmark study revealed strong correlations between TA skew, hydropathic indices, and their structural features. Moreover, in larger protein datasets, evident correlations between TA skew, the GRAVY score, and these representative protein domains were also observed. These findings reveal a previously unrecognized dimension of the correspondence between nucleotide composition and protein structures, suggesting the existence of an intricate function within the genetic code's codon–amino acid correspondence.

**Keywords:** Hydropathy, TA skew, Nucleotide composition, Genetic code, Chargaff's second parity rule, Optimized translation hypothesis.

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

# 1. Background

Anfinsen proposed the hypothesis that a protein's structure is determined solely by its amino acid sequence, which was later referred to as a "dogma" [1]. Building on this earlier concept, Kyte and Doolittle introduced the concept of **hydropathy** to explain a protein's conformational structure in terms of the hydrophilicity and hydrophobicity of its sequence [2]. Using several protein examples, they demonstrated that the hydropathy of an amino acid sequence can be estimated by applying coefficients to its amino acid composition, and that it indeed corresponds to the observed structural features of the protein [2]. Since then, this concept of hydropathy has significantly influenced the prediction of protein tertiary structures and remains a longstanding topic in modern contexts, including education.

However, predictions of hydropathy and other functional or structural aspects have so far relied exclusively on amino acid sequences and their residue compositions. Little consideration has been given to whether features of the corresponding gene sequences—such as nucleotide composition—could also determine protein characteristics.

In my previous report, I showed that the balance of thymine and adenine in gene sequences, termed **"TA skew,"** correlates positively with the proportion of transmembrane domains (TMD) and negatively with that of intrinsically disordered regions (IDR) [3]. TMDs are essential domains in membrane proteins, predominantly composed of hydrophobic amino acid residues that enable these proteins to traverse lipid bilayers. In contrast, IDRs are predominantly composed of highly hydrophilic amino acids and do not form a defined three-dimensional structure [4]. Considering that TA skew exhibits opposite correlations with TMDs and IDRs, and that these domains themselves also represent opposite ends of hydropathy, I hypothesized that a gene's TA skew correlates with the hydropathy of its encoded protein sequence.

To test this hypothesis, I first revisited the proteins analyzed in Kyte and Doolittle's original work to determine whether the TA skew of their gene sequences corresponds to their hydropathic indices and documented structural features. I then expanded the scope by using sufficiently large protein datasets, referencing the EMBL-EBI "Reference Proteomes" [5]. In these datasets, I examined whether each gene's TA skew correlates with the GRAVY score (the average hydropathy of entire proteins) and with the proportions of TMDs and IDRs. By integrating these findings, I aimed to clarify whether the balance of thymine and adenine in gene sequences can indeed reflect protein hydropathy and structural features.

# 2. Materials and Methods

### 2.1.1 Proteins from Kyte and Doolittle's work

In this study, I first analyzed the proteins examined in Kyte and Doolittle's landmark paper. That paper provided analytical results and structural information for six proteins: bovine chymotrypsinogen (CHYM) [6], dogfish lactate dehydrogenase (LDH) [7], erythrocyte glycophorin (GLYC) [8], rabbit cytochrome b5 (CB5R) [9], vesicular stomatitis virus glycoprotein (VSVG) [10], and bacteriorhodopsin (RHOD) [11]. Because no gene sequence information was included in that publication or its cited references, I obtained the corresponding amino acid and gene sequences from current public databases. To confirm consistency between the original information and the database-derived data, I performed side-by-side comparisons of their amino acid residue sequences for each pair.

### 2.1.2 Correlation Analysis on Example Proteins

In the first part of this study, I analyzed correlations among the hydropathic indices, structural features, and TA skew for the proteins illustrated in Kyte and Doolittle's landmark paper. Here, **TA skew** refers to the balance between thymine and adenine nucleotides in a gene sequence, defined as

$$\text{TA skew} = \frac{T - A}{T + A},$$

where, $T$, and $A$ denote the respective counts of thymine and adenine in the nucleotide sequence [3].

Before each correlation analysis, I overlaid the hydropathic index graphs published in the original paper with those generated from the modern database sequences used in this study to visually confirm their consistency. Next, for each moving window (referred to as a "Span" in the original paper) used to calculate the hydropathic index from the amino acid sequence, I extracted the corresponding gene sequence and computed its TA skew, generating TA skew plots. I then compared these TA skew plots with the original hydropathy plots and their associated structural features. Finally, to quantitatively assess the degree of correlation between the calculated hydropathic index values and the corresponding TA skew, I computed correlation coefficients.

### 2.2.1 Proteins from “Reference Proteomes” dataset

In the subsequent analysis, I used a dataset published as “Reference Proteomes” on the EMBL-EBI website [5]. The dataset I employed (release 2023_03) included a total of 1,023,125 amino acid sequences from 79 species spanning the three domains of life, along with the corresponding nucleotide sequences for these genes. However, within this dataset, there were numerous entries that clearly did not correspond to the amino acid sequence data when treated as coding sequences (CDS). It is likely that some mRNA or other non-CDS data were mixed into the dataset. Given the challenges of extracting CDS regions from each mRNA sequence under my current data-processing conditions, I decided to exclude any gene information that did not align directly with its corresponding amino acid sequence. Therefore, I cross-referenced the gene and protein sequences, removing any entries whose gene lengths did not match or that fell outside the known range of genetic code deviations [12]. This procedure ultimately yielded 857,750 proteins from 79 species across the three domains for analysis (Table 1).

### 2.2.2 TA Skew, GRAVY score, and TMD–IDR in the “Reference Proteomes”

In this analysis of the Reference Proteomes dataset, I used three values—**TA skew**, **GRAVY score**, and **TMD–IDR**. Here, I describe the calculation methods for each.

**GRAVY Score:**

The GRAVY score is calculated similarly to the hydropathic index, but the key difference is that the hydropathic index is derived from partial windows (or segments) of an amino acid sequence, whereas the GRAVY score is computed over the entire protein sequence based on its overall amino acid residue composition. In their original publication, Kyte and Doolittle indicated that this score reflects the distinctive features of a protein [2].

**TMD–IDR:**

Although the Reference Proteomes dataset does not directly provide structural information in the same manner as Kyte and Doolittle’s original paper, it is linked to UniProtKB protein entries. In this study, I calculated the proportions of transmembrane domain (TMD) residues and intrinsically disordered region (IDR) residues—relative to the total amino acid count of each protein—using the corresponding UniProtKB entries. Because TMD and IDR proportions are treated as independent variables, and these two regions are not assigned to

the same amino acid segment—and because they represent opposite extremes of hydropathy—I combined these two variables into a single measure, referred to as "TMD–IDR," defined as the proportion of TMD minus the proportion of IDR. In this scheme, a larger proportion of TMD drives the TMD–IDR value closer to +1, whereas a larger proportion of IDR shifts it toward −1. If neither TMD nor IDR is present, or their proportions are equal, the value is 0. Consequently, if a variable shows correlation with TMD–IDR, it can be considered correlated with a protein's structural characteristics.

**TA Skew:**

To calculate each gene's TA skew, I counted the number of thymine and adenine nucleotides within each protein's gene sequence and then computed TA skew using the same formula described in Section 2.1.2. However, because stop codons do not encode amino acids, I excluded them from this calculation in this analysis.

### 2.2.3 Correlation Analysis on the Larger Protein Dataset

After obtaining each protein's TA skew, GRAVY score, and TMD–IDR value according to the methods described above, I performed a combinatorial correlation analysis across the entire dataset to examine their interrelationships. Additionally, because eukaryotic proteins comprised the majority of the current dataset—and to assess whether results might vary with dataset composition—I also conducted the same analysis for each of the three domains of life, examining those outcomes separately.

## 2.3 Data Processing

All downloaded data were provided in FASTA format. All initial FASTA data handling—including verifying nucleotide and amino acid residue sequence matches for each protein, as well as calculating compositional values—was performed using Microsoft Excel (version 16.94, Microsoft 365) on macOS 15.3.1 (24D70). The fractions derived from UniProtKB annotations (TMDs and IDRs) were also calculated in Excel. Correlation coefficient calculations and plot generation were carried out in JMP Pro 18.1.2 (SAS Institute Inc., Cary, NC, USA). Finally, figures were prepared for publication using Microsoft PowerPoint (version 16.94, Microsoft 365) on macOS 15.3.1 (24D70).

# 3. Results on Six Example Proteins

In this section, I present side-by-side comparisons of each protein's hydropathic index plots (based on modern database sequence and gene information), the corresponding TA skew plots, and documented structural features, with reference to Kyte and Doolittle's original paper.

## 3.1 Bovine Chymotrypsinogen (CHYM)

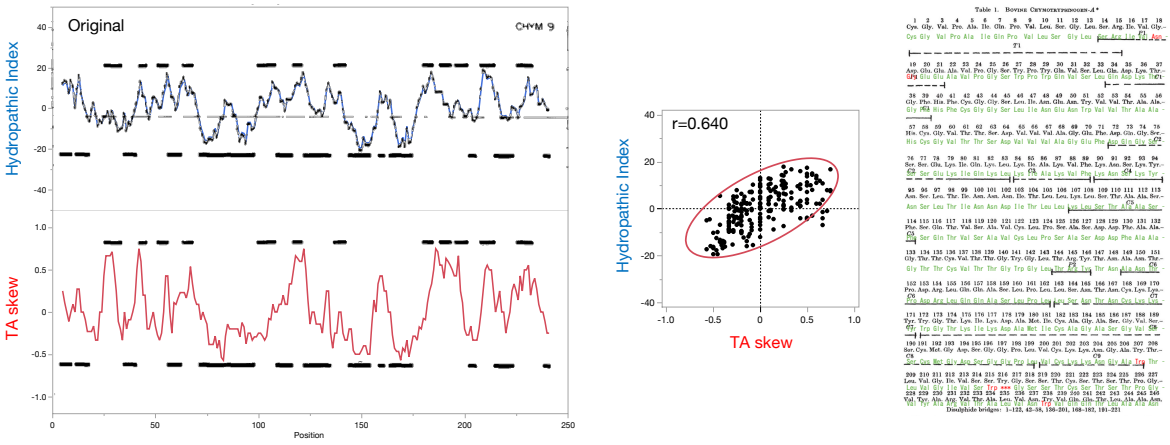



**Figure 1. Correlation in Bovine Chymotrypsinogen**

Figure 1 illustrates the results for bovine chymotrypsinogen [6,13]. In the upper-left portion, the black plot line represents the hydropathic indices calculated with a 9-amino-acid window from Kyte and Doolittle's seminal work, while the overlaid blue line shows the hydropathic index plot derived from current database sequences. Below these plots, the TA skew of the corresponding gene sequence is shown in red for each of those windows. In addition, both the hydropathic index and TA skew plots feature alternating horizontal lines drawn above or below the plot: lines above indicate portions of the protein structure that fold inward, whereas lines below indicate regions facing outward.

The middle figure is a correlation plot comparing the hydropathic index and TA skew (correlation coefficient r=0.640).

Finally, on the far right, the amino acid sequence documented in the original reference paper is shown alongside the corresponding data retrieved from the current database to verify the data's validity. Amino acids that match between these two sources are shown in green, while any discrepancies appear in red.

## 3.2 Dogfish Lactate Dehydrogenase (LDH)

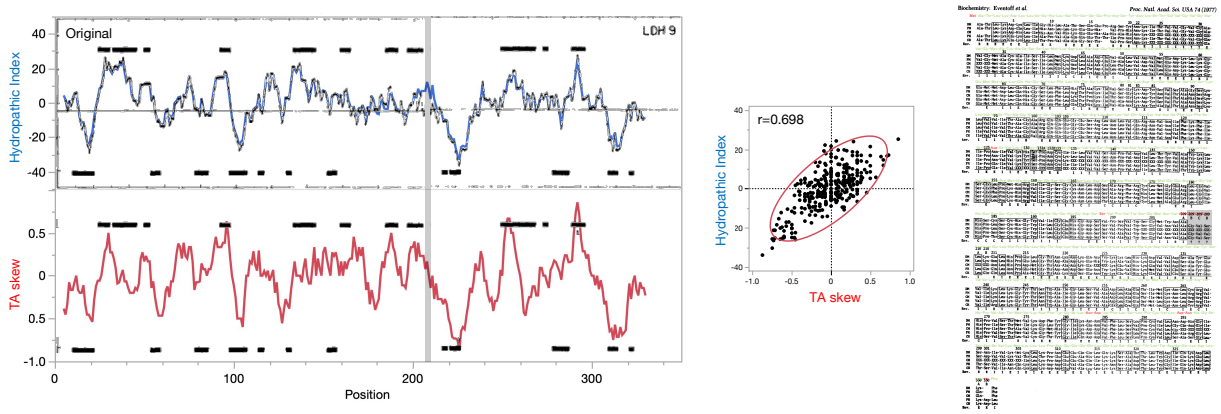



**Figure 2. Correlation in Dogfish Lactate Dehydrogenase**

Figure 2 shows the results for dogfish lactate dehydrogenase[7,14]. As in the previous figure, the black plot line in the upper-left portion represents the hydropathic indices calculated with a 9-amino-acid window from the original paper, and the overlaid blue line shows the hydropathic index plot generated from the current database sequences. Below these plots, the corresponding TA skew values are shown in red. As in Figure 1, lines above the plot indicate regions of the protein structure that fold inward, whereas lines below indicate outward-facing regions.

Next to these plots is a correlation diagram (correlation coefficient r=0.698) comparing the hydropathic index and TA skew. Farther to the right, the amino acid sequence from the original reference paper appears alongside the corresponding data retrieved from the current database. Amino acids that match between these two sources appear in green, while any discrepancies are shown in red. Because the current database data included three additional amino acid residues in the middle of the sequence, I highlighted that alignment in gray to indicate the shift, which is also shown on the left plot.

## 3.3 Erythrocyte Glycophorin (GLYC)

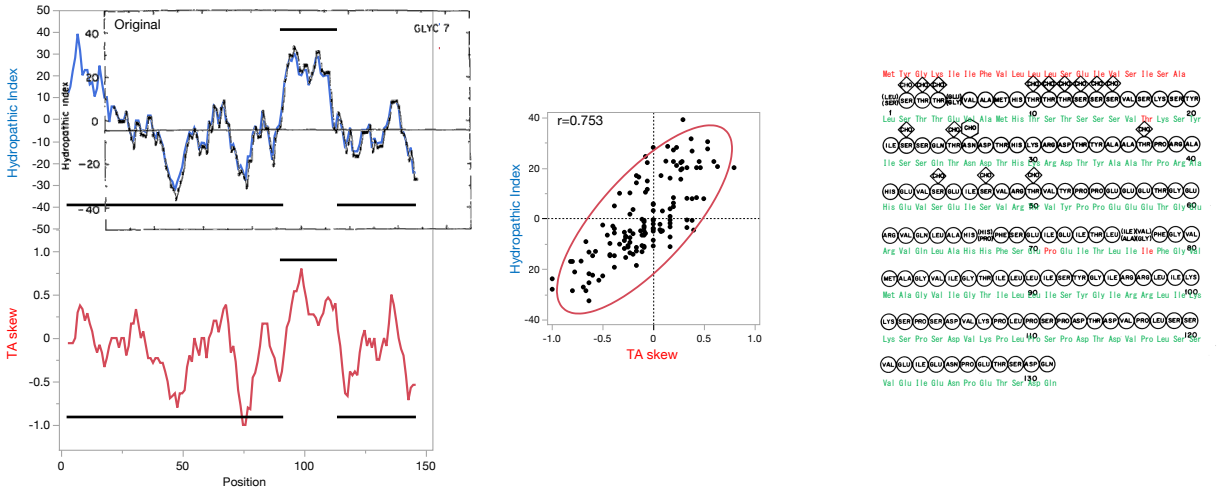



**Figure 3. Correlation in Erythrocyte Glycophorin**

Figure 3 shows the results for human erythrocyte glycophorin [8,15]. The format is the same as in Figures 1 and 2; however, in this figure, the window used to analyze the amino acid sequence is 7 amino acids instead of 9. In addition, the upper lines added to the left plot indicate membrane-spanning regions (transmembrane domains), while the lower lines indicate the remaining regions. Because the modern database includes an N-terminal sequence not present in the original paper's data, the original plot has been shifted and overlaid accordingly. The hydropathic index and TA skew exhibit a positive correlation, with a correlation coefficient of 0.753.

## 3.4 Rabbit Cytochrome b5 (CB5R)

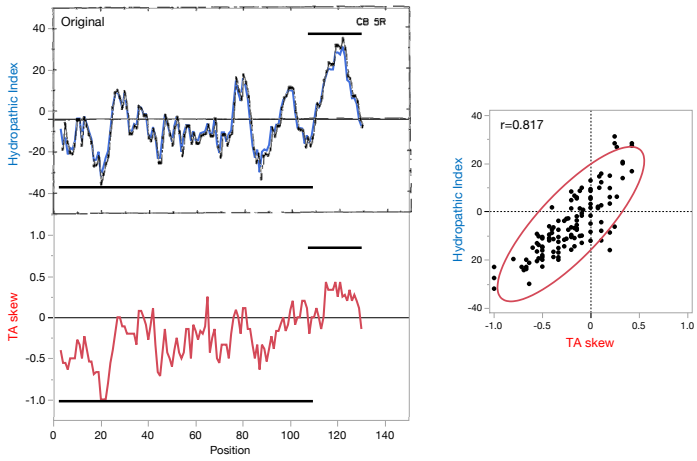



**Figure 4. Correlation in Rabbit Cytochrome b5**

Figure 4 shows the results for rabbit cytochrome b5 [9,16]. The format is the same as in Figure 3, so the upper lines on the left plot indicate membrane-spanning domains, while the lower lines denote other regions. The correlation coefficient here is 0.817. Because no sequence data were available in the reference, I could not perform an amino acid sequence alignment; however, the overlaid plots in the upper-left portion appear to match sufficiently well.

## 3.5 Vesicular Stomatitis Virus Glycoprotein (VSVG)

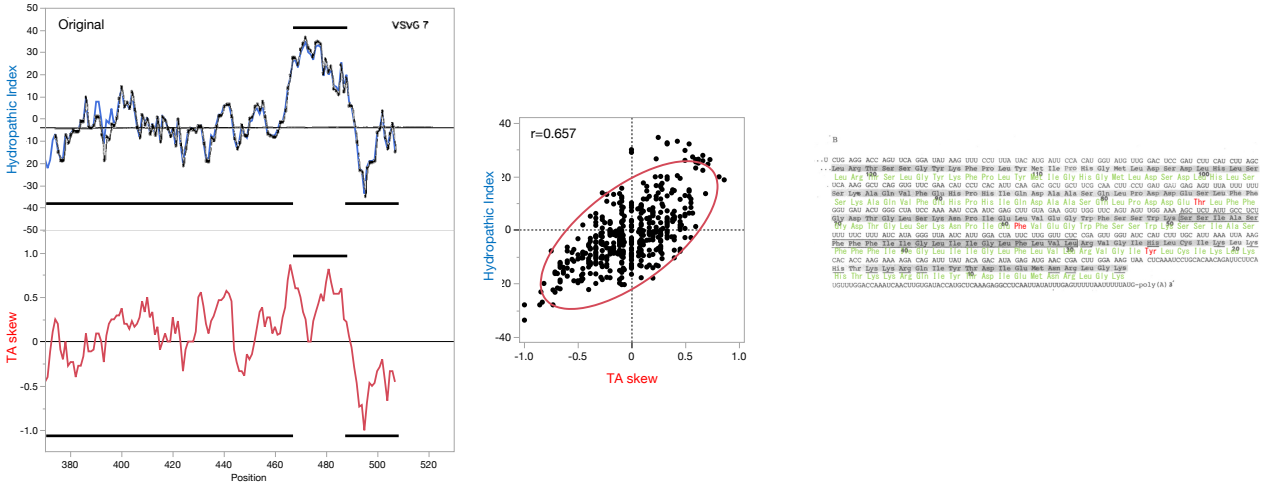



**Figure 5. Correlation in Vesicular Stomatitis Virus Glycoprotein**

Figure 5 shows the results for vesicular stomatitis virus glycoprotein [10,17]. The format is the same as in Figure 3, so the upper lines on the left plot indicate membrane-spanning domains, while the lower lines denote other regions. The correlation coefficient here is 0.657. Because the original reference only plotted the latter half of the protein, the right side shows a comparison focusing on that portion of the reference plot.

## 3.6 Bacteriorhodopsin (RHOD)

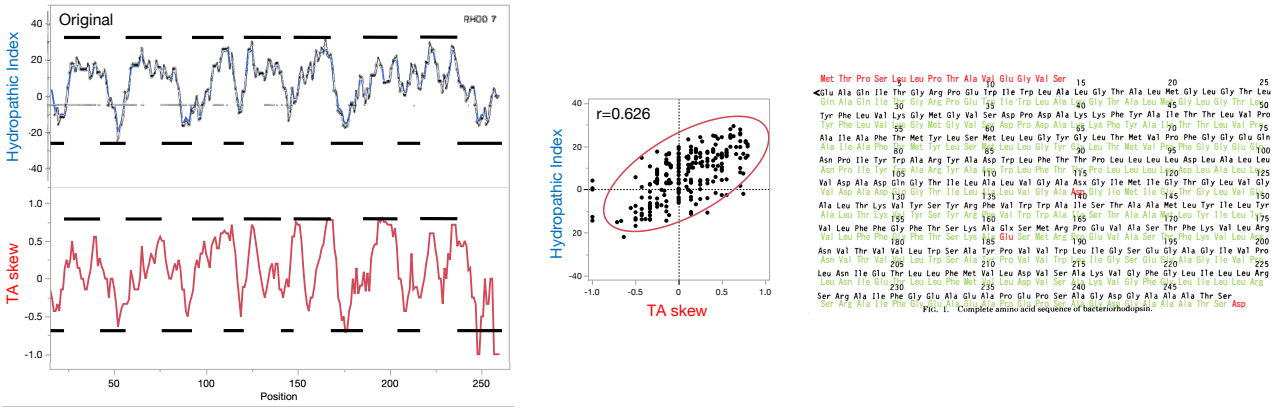



**Figure 6. Correlation in Bacteriorhodopsin**

Figure 6 shows the results for bacteriorhodopsin [11,18]. The format is the same as in Figure 3, so the upper lines in the left plot indicate membrane-spanning domains, while the lower lines denote other regions. The correlation coefficient here is 0.626. As shown on the right, the modern database data include additional sequences at both the N- and C-terminal regions. Consequently, for this analysis, I focused on the same sequence range as the original data and overlaid the plots for direct comparison.

# 4. Results on Larger Protein Datasets

In this section, I present the mutual correlations among the GRAVY score, TA skew, and TMD–IDR in larger protein datasets. First, I describe the composition of the dataset used in this analysis. Next, I show the overall correlations across the entire dataset—which includes more than 850,000 proteins—and finally, as an additional analysis, I provide the results of correlation analyses conducted separately for each domain of life (Archaea, Bacteria, and Eukaryota).

## 4.1 The Larger Dataset Used for This Analysis

Table 1 shows the species included in this study and the number of proteins analyzed for each species. The dataset, referred to as **"Reference Proteomes,"** contained amino acid sequence data for 1,023,125 proteins from 79 species in its 2023 release. However, as explained in the Materials and Methods section, some entries had gene information that did not match the corresponding amino acid sequence. By cross-referencing, I excluded any entries whose gene sequences did not align with their amino acid sequences, ultimately selecting 857,750 proteins for this analysis (see Table 1 at the end of this publication).

## 4.2 Mutual Correlations Among Indices in the Current Entire Dataset

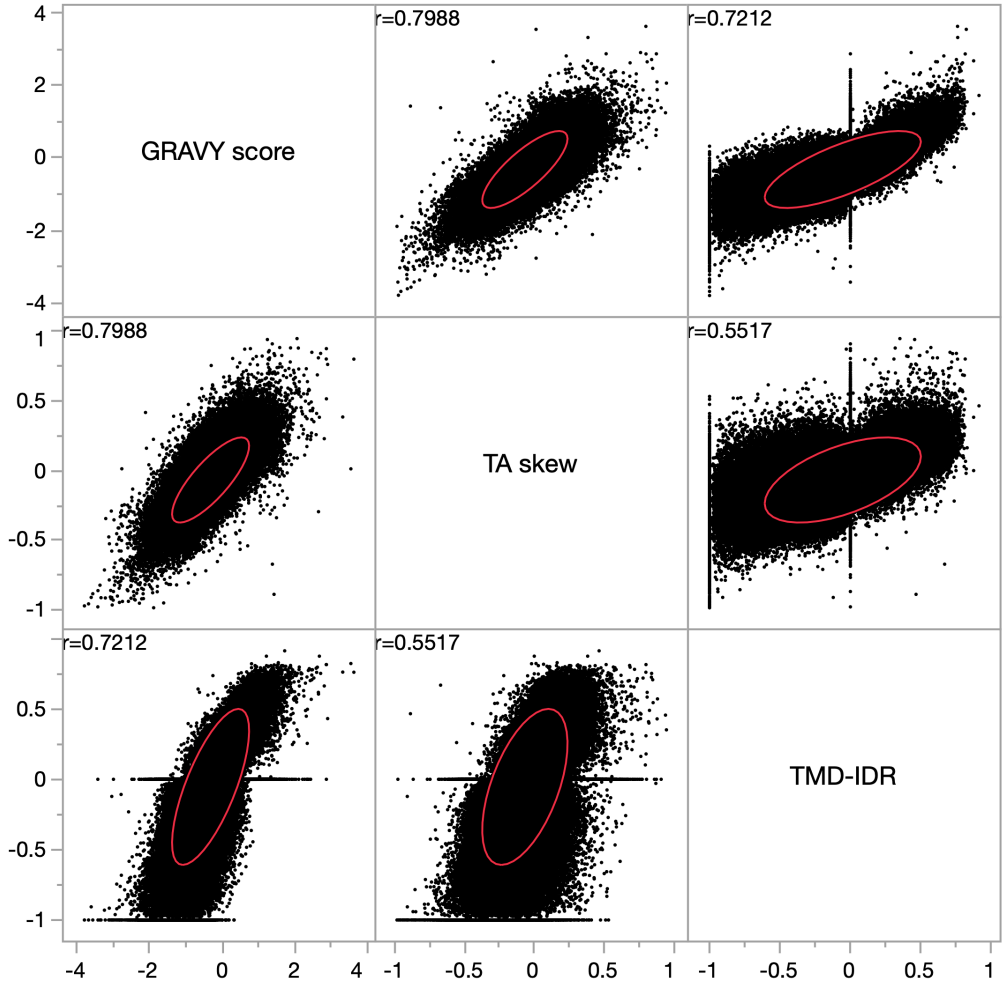
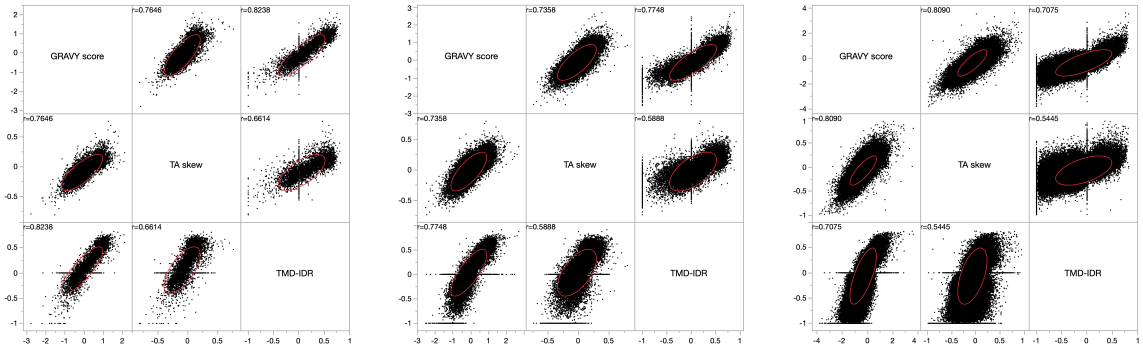



**Figure 7. Mutual Correlations Among Indices in the Current Entire Dataset**

Figure 7 shows the results of a mutual correlation analysis among the GRAVY score, TA skew, and TMD–IDR value in the current full Reference Proteomes dataset of 857,750 proteins. The correlation coefficient between GRAVY score and TA skew was 0.7988, between GRAVY score and TMD–IDR was 0.7212, and between TA skew and TMD–IDR was 0.5517.

### 4.2 Mutual Correlations Among Indices Analyzed by Domain




a. Archaea (n=16,985) b. Bacteria (n=81,603) c. Eukaryota (n=759,162)

**Figure 8. Mutual Correlations Among Indices Within Each Domain**

Figure 8(a–c) shows the results of analyzing the mutual correlations among these indices for each domain. The correlations observed in Figure 7 remained evident even when focusing on the smaller datasets from Archaea and Bacteria. Notably, a stronger correlation was found between TA skew and TMD–IDR in these domains, with correlation coefficients of 0.6614 in Archaea and 0.5888 in Bacteria.

## 5. Discussion

### 5.1 Anfinsen's Dogma and the Significance of Hydropathy-Based Predictions

From Anfinsen's "dogma," which asserts that a protein's three-dimensional structure is uniquely determined by its amino acid sequence, arose a line of research aimed at inferring conformational structures based on hydropathy. This approach has become a landmark in modern protein structure prediction, now culminating in machine learning–based large language models such as AlphaFold.

### 5.2 Overlooked Link Between Gene Sequence and Protein Structure

In contrast, very few studies have directly matched protein structure to the nucleotide sequence—or even the nucleotide composition—of its coding gene. I found only a single report from 2020 suggesting that gene sequences rich in thymine tend to encode membrane proteins (including those containing transmembrane domains) [19], and no other publications appeared to address this issue. Consequently, analyses examining whether

nucleotide composition might correlate with structural features of encoded proteins have largely been overlooked.

### 5.3 Motivation for Investigating TA Skew

Why might this possibility have been overlooked? Regarding amino acids, because each of the 20 amino acids that constitute proteins has distinct chemical properties, it is relatively straightforward to accept the concept that amino acid sequences shape conformational structure. However, while the genetic code uniquely maps nucleotide sequences to amino acids, synonymous codons introduce uncertainty in this relationship, making the correspondence less transparent. This likely explains why the idea that a gene's nucleotide composition could determine protein characteristics has not gained widespread acceptance.

So why and how did I choose to explore this issue? In my previous work, I calculated the amino acid compositions of an entire human exome (all proteins) alongside the nucleotide compositions of their corresponding genes, then performed principal component analyses (PCA) on both. The first through third principal components of the amino acid compositions were found to correlate with the first through third principal components of the nucleotide compositions, respectively—indicating that, statistically, a protein's amino acid composition originates from the nucleotide composition of its coding gene [3]. I also observed that the second principal component of the amino acid composition distinguished proteins rich in transmembrane domains (TMDs) from those rich in intrinsically disordered regions (IDRs) [3]. This same second principal component corresponded to the second principal component in the nucleotide composition analysis, representing the balance between thymine and adenine—namely, the TA skew. From these findings, I deduced that a gene's thymine–adenine balance correlates with the generation of TMDs and IDRs in proteins, leading to the hypothesis examined in this study.

### 5.4 Results of the Current Examination

In the former part of this study, I tested the above hypothesis by analyzing the six proteins documented in Kyte and Doolittle's paper. In the latter part, I expanded the scope to a larger dataset using the Reference Proteomes data.

Results from the former part showed that the newly obtained modern gene information generally matched well with that described in the original publication. In these data, the hydropathic index and TA skew of the six examined proteins were correlated ($r = 0.640$,

0.698, 0.753, 0.817, 0.657, and 0.626) (Figures 1–6, respectively), and these correlations coincided with structural features such as inward/outward folding and membrane-spanning domains. Notably, in Figure 1—focusing on bovine chymotrypsinogen—near the N-terminal (leftmost) region, characterized by inward folding (indicated by the upper black line), the hydropathic index is low while the TA skew is high. This suggests that TA skew may relate to conformational structures in a way not solely mediated by the hydropathic index, a particularly interesting observation.

Results from the latter part showed that the GRAVY score, which corresponds to the hydropathic index, correlates with TA skew in each gene and also with TMD–IDR—a measure reflecting the proportions of two protein domains (Figures 7, 8). In particular, the strong correlation between the GRAVY score and TA skew supports the conclusion that the correlations observed in the former part are not coincidental but persist across the entire dataset.

From these findings, I concluded that the TA skew of a protein gene correlates with both the protein's hydropathic index (and GRAVY score) and its conformational structures. The question, however, is whether these correlations with nucleotide compositions represent essential linkages or are merely reflections of other factors. This issue will be addressed in the following sections.

### 5.5 Can the Correlation Between TA Skew and Protein Domains Be Explained by the Genetic Code?

The correlation noted here—between TA skew, an index of nucleotide composition, and the proportions of two representative protein domains (TMD and IDR)—raises the question of whether it can be explained by the structure of the genetic code, i.e., the codon–amino acid correspondence. The genetic code has been studied extensively, and its non-randomness is well recognized. For instance, codons with U (T in the gene) in the second position consistently encode highly hydrophobic amino acids, a pattern frequently attributed to robustness against mutations [20]. However, in my comparisons of amino acid sequences across diverse exomes, I found that transmembrane domains are enriched in amino acids requiring thymine to be coded, whereas intrinsically disordered regions are enriched in amino acids that do not require thymine [21]. This suggests that the genetic code itself may be structured so that thymine-rich gene regions align with TMDs, while thymine-poor regions align with IDRs.

Nevertheless, synonymous codons add another layer of complexity. In my earlier work, I showed that synonymous codon usage is governed predominantly not by species per se, but by each gene's GC content [22]. Genes with higher GC content use synonymous codons richer in GC, whereas genes with lower GC content use synonymous codons lower in GC, ensuring a functionally stable amino acid composition despite variations in GC content among genes. Considering this mechanism, if thymine content is high but adenine is also high, the GC content decreases, and synonymous codon usage shifts accordingly—effectively "absorbing" the excess of thymine plus adenine. Consequently, rather than the absolute thymine level, it is ultimately the balance of thymine and adenine (i.e., TA skew) that correlates with TMD and IDR proportions.

### 5.6 Why TA Skew Determines Hydropathy and Structure

So far, we have shown that TA skew, an index of nucleotide composition, correlates with both the hydropathic index (calculated from amino acid composition) and structural features of proteins such as TMDs and IDRs. However, it remains entirely possible that this correlation is merely coincidental, devoid of deeper significance. If TA skew is simply the balance of thymine and adenine, why would it matter for the distribution of protein domains and hydropathy?

At this point, one previously puzzling phenomenon comes to mind: **Chargaff's second parity rule** [23]. In general, the pairing of thymine and adenine during DNA replication—often referred to as Chargaff's parity rule—stems from the empirical observation that each genome contains equal counts of thymine and adenine, as well as guanine and cytosine. Less widely known, however, is that Chargaff also reported a second empirical rule: within a single strand of the genome, if one considers a sufficiently long sequence, the amounts of thymine and adenine, and of guanine and cytosine, are "almost" the same. This observation later came to be called Chargaff's second parity rule. Subsequent analyses revealed that nearly all organisms' genomes follow this rule, whereas the genomes of their mitochondria and viruses, for reasons yet unclear, deviate from it. However, no satisfactory explanation has yet been provided for this phenomenon, leaving it shrouded in mystery [24].

Returning to the findings of this report: our investigation indicates that the balance of thymine and adenine within a gene correlates with the protein's hydropathy and its domain balance (TMD and IDR), mediated by the genetic code. Accordingly, each gene's TA skew in a genome sequence determines the protein's characteristics. At the same time, a gene's TA skew must depend on the balance of thymine and adenine in the genome that harbors it. If

the genome itself is constructed to maintain a balance of thymine and adenine, then each gene's TA skew would in turn maintain a stable balance in its encoded proteins. Viewed in this light, the previously unexplained Chargaff's second parity rule could be construed as a genomic mechanism that maintains a particular distribution of thymine and adenine—thereby regulating the hydropathic properties and the formation of related domains in encoded proteins. In other words, this insight offers a potential explanation for the longstanding mystery behind Chargaff's second parity rule.

Based on the observations and inferences presented here, it appears that the correlation of TA skew with the hydropathic index and structural features is not a mere coincidence, but rather a likely outcome of the inductive properties inherent in the genetic code—and it may even shed light on the enigma of Chargaff's second parity rule.

## 6. Conclusion

In this paper, by examining the amino acid compositions of various protein residue sequences and their corresponding genes, I demonstrated that **TA skew**—a nucleotide composition index reflecting the balance of thymine and adenine—significantly correlates with the hydropathic indices proposed by Kyte and Doolittle. Additionally, the analysis showed that TA skew correlates with structural features of proteins. While it remains debatable whether this correlation between TA skew and protein structures is inevitable or merely coincidental, based on the current discussion, I concluded that this phenomenon is not a chance occurrence but rather a manifestation of an intrinsic function arising from the structure of the genetic code.

# 7. Reference


1. **Anfinsen, C. B.** (1973). Principles that Govern the Folding of Protein Chains. *Science, 181*(4096), 223–230. https://doi.org/10.1126/science.181.4096.223

2. **Kyte, J., & Doolittle, R. F.** (1982). A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology, 157*(1), 105–132. https://doi.org/10.1016/0022-2836(82)90515-0

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

4. **Dyson, H. J., & Wright, P. E.** (2005). Intrinsically unstructured proteins and their functions. *Nature Reviews Molecular Cell Biology, 6*(3), 197–208. https://doi.org/10.1038/nrm1589

5. **EMBL-EBI.** (2023). Reference Proteomes (Release 2023_03) [Database]. Retrieved September 1, 2023, from https://www.ebi.ac.uk/reference_proteomes/

6. **Hartley, B. S.** (1964). Amino-Acid Sequence of Bovine Chymotrypsinogen-A. *Nature, 201*(4926), 1284–1287. https://doi.org/10.1038/2011284a0

7. **Eventoff, W., Rossmann, M. G., Taylor, S. S., Torff, H. J., Meyer, H., Keil, W., & Kiltz, H. H.** (1977). Structural adaptations of lactate dehydrogenase isozymes. *Proceedings of the National Academy of Sciences, 74*(7), 2677–2681. https://doi.org/10.1073/pnas.74.7.2677

8. **Tomita, M., & Marchesi, V. T.** (1975). Amino-acid sequence and oligosaccharide attachment sites of human erythrocyte glycophorin. *Proceedings of the National Academy of Sciences, 72*(8), 2964–2968. https://doi.org/10.1073/pnas.72.8.2964

9. **Strittmatter, P., Rogers, M. J., & Spatz, L.** (1972). The Binding of Cytochrome b5 to Liver Microsomes. *Journal of Biological Chemistry, 247*(22), 7188–7194. https://doi.org/10.1016/S0021-9258(19)44612-7

10. **Rose, J. K., Welch, W. J., Sefton, B. M., Esch, F. S., & Ling, N. C.** (1980). Vesicular stomatitis virus glycoprotein is anchored in the viral membrane by a hydrophobic domain near the COOH terminus. *Proceedings of the National Academy of Sciences, 77*(7), 3884–3888. https://doi.org/10.1073/pnas.77.7.3884

11. **Khorana, H. G., Gerber, G. E., Herlihy, W. C., Gray, C. P., Anderegg, R. J., Nihei, K., & Biemann, K.** (1979). Amino acid sequence of bacteriorhodopsin. *Proceedings of the National Academy of Sciences, 76*(10), 5046–5050. https://doi.org/10.1073/pnas.76.10.5046

12. **Hamashima, K., Kanai, A.** (2014). Unexpected tRNAs that do not consistently obey the universal genetic code (In Japanese). *Seikagaku. The Journal of Japanese Biochemical Society,* 86(4), 483-488. https://www.jbsoc.or.jp/seika/wp-content/uploads/2015/03/86-04-10.pdf

13. **National Center for Biotechnology Information.** (n.d.). *chymotrypsinogen A [Bos taurus]* (NCBI Reference Sequence: XP_003587247.4). Retrieved March 7, 2025, from https://www.ncbi.nlm.nih.gov/protein/XP_003587247.4/

14. **UniProt Consortium.** (n.d.). *L-lactate dehydrogenase A chain (Squalus acanthias)* (UniProt accession No. P00341). Retrieved March 7, 2025, from https://www.uniprot.org/uniprotkb/P00341/entry

15. **UniProt Consortium.** (n.d.). *Glycophorin-A (Homo sapiens)* (UniProt accession No. P02724). Retrieved March 7, 2025, from https://www.uniprot.org/uniprotkb/P02724/entry

16. **UniProt Consortium.** (n.d.). *Cytochrome b5 (Oryctolagus cuniculus)* (UniProt accession No. P00169). Retrieved March 7, 2025, from https://www.uniprot.org/uniprotkb/P00169/entry

17. **UniProt Consortium.** (n.d.). *Glycoprotein (Vesicular stomatitis Indiana virus)* (UniProt accession No. P04884). Retrieved March 7, 2025, from https://www.uniprot.org/uniprotkb/P04884/entry

18. **National Center for Biotechnology Information.** (n.d.). *bacteriorhodopsin [Halobacterium salinarum]* (NCBI Reference Sequence: WP_136361479.1). Retrieved March 7, 2025, from https://www.ncbi.nlm.nih.gov/protein/WP_136361479.1

19. **Vakirlis, N., Acar, O., Hsu, B., Castilho Coelho, N., van Oss, S. B., Wacholder, A., Medetgul-Ernar, K., Bowman, R. W., 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. *Nature Communications, 11*(1), 781. https://doi.org/10.1038/s41467-020-14500-z

20. **Freeland, S. J., & Hurst, L. D.** (1998). The Genetic Code Is One in a Million. *Journal of Molecular Evolution, 47*(3), 238–248. https://doi.org/10.1007/PL00006381

21. **Esumi, G.** (2024). 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 [Preprint]. Jxiv. https://doi.org/10.51094/jxiv.592

22. **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 [Preprint]. Jxiv. https://doi.org/10.51094/jxiv.561

23. **Rudner, R., Karkas, J. D., & Chargaff, E.** (1968). Separation of B. subtilis DNA into complementary strands. 3. Direct analysis. *Proceedings of the National Academy of Sciences, 60*(3), 921–922. https://doi.org/10.1073/pnas.60.3.921

24. **Forsdyke, D. R., & Mortimer, J. R.** (2000). Chargaff's legacy. *Gene, 261*(1), 127–137. https://doi.org/10.1016/S0378-1119(00)00472-8

# 8. Table

| No. | Domain | taxonomy_id | Organism | Listed | Matched |
|---|---|---|---|---|---|
| 1 | Archaea | 64091 | Halobacterium salinarum (strain ATCC 700922 / JCM 11081 / NRC-1) (Halobacterium halobium) | 2423 | 2332 |
| 2 | Archaea | 69014 | Thermococcus kodakarensis (strain ATCC BAA-918 / JCM 12380 / KOD1) (Pyrococcus kodakaraensis (strain KOD1)) | 2301 | 2299 |
| 3 | Archaea | 188937 | Methanosarcina acetivorans (strain ATCC 35395 / DSM 2834 / JCM 12185 / C2A) | 4468 | 4420 |
| 4 | Archaea | 243232 | Methanocaldococcus jannaschii (strain ATCC 43067 / DSM 2661 / JAL-1 / JCM 10045 / NBRC 100440) (Methanococcus jannaschii) | 1787 | 1667 |
| 5 | Archaea | 273057 | Saccharolobus solfataricus (strain ATCC 35092 / DSM 1617 / JCM 11322 / P2) (Sulfolobus solfataricus) | 2937 | 2871 |
| 6 | Archaea | 374847 | Korarchaeum cryptofilum (strain OPF8) | 1602 | 1601 |
| 7 | Archaea | 436308 | Nitrosopumilus maritimus (strain SCM1) | 1795 | 1795 |
| 8 | Bacteria | 83332 | Mycobacterium tuberculosis (strain ATCC 25618 / H37Rv) | 3995 | 3834 |
| 9 | Bacteria | 83333 | Escherichia coli (strain K12) | 4403 | 4327 |
| 10 | Bacteria | 85962 | Helicobacter pylori (strain ATCC 700392 / 26695) (Campylobacter pylori) | 1554 | 1503 |
| 11 | Bacteria | 100226 | Streptomyces coelicolor (strain ATCC BAA-471 / A3(2) / M145) | 8035 | 7980 |
| 12 | Bacteria | 122586 | Neisseria meningitidis serogroup B (strain MC58) | 2001 | 1972 |
| 13 | Bacteria | 189518 | Leptospira interrogans serogroup Icterohaemorrhagiae serovar Lai (strain 56601) | 3676 | 3645 |
| 14 | Bacteria | 190304 | Fusobacterium nucleatum subsp. nucleatum (strain ATCC 25586 / DSM 15643 / BCRC 10681 / CIP 101130 / JCM 8532 / KCTC 2640 / LMG 13131 / VPI 4355) | 2046 | 2022 |
| 15 | Bacteria | 208964 | Pseudomonas aeruginosa (strain ATCC 15692 / DSM 22644 / CIP 104116 / JCM 14847 / LMG 12228 / 1C / PRS 101 / PAO1) | 5564 | 5535 |
| 16 | Bacteria | 224308 | Bacillus subtilis (strain 168) | 4260 | 4213 |
| 17 | Bacteria | 224324 | Aquifex aeolicus (strain VF5) | 1553 | 1531 |
| 18 | Bacteria | 224911 | Bradyrhizobium diazoefficiens (strain JCM 10833 / BCRC 13528 / IAM 13628 / NBRC 14792 / USDA 110) | 8253 | 8192 |
| 19 | Bacteria | 226186 | Bacteroides thetaiotaomicron (strain ATCC 29148 / DSM 2079 / JCM 5827 / CCUG 10774 / NCTC 10582 / VPI-5482 / E50) | 4782 | 4768 |
| 20 | Bacteria | 243090 | Rhodopirellula baltica (strain DSM 10527 / NCIMB 13988 / SH1) | 7271 | 7194 |
| 21 | Bacteria | 243230 | Deinococcus radiodurans (strain ATCC 13939 / DSM 20539 / JCM 16871 / CCUG 27074 / LMG 4051 / NBRC 15346 / NCIMB 9279 / VKM B-1422 / R1) | 3084 | 2946 |
| 22 | Bacteria | 243231 | Geobacter sulfurreducens (strain ATCC 51573 / DSM 12127 / PCA) | 3402 | 3398 |
| 23 | Bacteria | 243273 | Mycoplasma genitalium (strain ATCC 33530 / DSM 19775 / NCTC 10195 / G37) (Mycoplasmoides genitalium) | 483 | 470 |
| 24 | Bacteria | 243274 | Thermotoga maritima (strain ATCC 43589 / DSM 3109 / JCM 10099 / NBRC 100826 / MSB8) | 1852 | 1819 |
| 25 | Bacteria | 251221 | Gloeobacter violaceus (strain ATCC 29082 / PCC 7421) | 4406 | 4385 |
| 26 | Bacteria | 272561 | Chlamydia trachomatis (strain D/UW-3/Cx) | 895 | 882 |
| 27 | Bacteria | 289376 | Thermodesulfovibrio yellowstonii (strain ATCC 51303 / DSM 11347 / YP87) | 1982 | 1982 |
| 28 | Bacteria | 324602 | Chloroflexus aurantiacus (strain ATCC 29366 / DSM 635 / J-10-fl) | 3850 | 3847 |
| 29 | Bacteria | 515635 | Dictyoglomus turgidum (strain DSM 6724 / Z-1310) | 1743 | 1743 |
| 30 | Bacteria | 1111708 | Synechocystis sp. (strain PCC 6803 / Kazusa) | 3507 | 3415 |
| 31 | Eukaryota | 3055 | Chlamydomonas reinhardtii (Chlamydomonas smithii) | 17614 | 17565 |
| 32 | Eukaryota | 3218 | Physcomitrium patens (Spreading-leaved earth moss) (Physcomitrella patens) | 31359 | 30772 |
| 33 | Eukaryota | 3702 | Arabidopsis thaliana (Mouse-ear cress) | 27481 | 26180 |
| 34 | Eukaryota | 4577 | Zea mays (Maize) | 39225 | 36849 |
| 35 | Eukaryota | 5664 | Leishmania major | 8038 | 8034 |
| 36 | Eukaryota | 5888 | Paramecium tetraurelia | 39461 | 39096 |
| 37 | Eukaryota | 6239 | Caenorhabditis elegans | 19827 | 18905 |
| 38 | Eukaryota | 6412 | Helobdella robusta (Californian leech) | 23328 | 20976 |
| 39 | Eukaryota | 6945 | Ixodes scapularis (Black-legged tick) (Deer tick) | 20496 | 13321 |
| 40 | Eukaryota | 7070 | Tribolium castaneum (Red flour beetle) | 16568 | 16416 |
| 41 | Eukaryota | 7165 | Anopheles gambiae (African malaria mosquito) | 13016 | 2323 |
| 42 | Eukaryota | 7227 | Drosophila melanogaster (Fruit fly) | 13821 | 13286 |
| 43 | Eukaryota | 7719 | Ciona intestinalis (Transparent sea squirt) (Ascidia intestinalis) | 16680 | 10168 |
| 44 | Eukaryota | 7739 | Branchiostoma floridae (Florida lancelet) (Amphioxus) | 26627 | 25416 |
| 45 | Eukaryota | 7918 | Lepisosteus oculatus (Spotted gar) | 18321 | 14018 |
| 46 | Eukaryota | 7955 | Danio rerio (Zebrafish) (Brachydanio rerio) | 26249 | 24384 |
| 47 | Eukaryota | 8090 | Oryzias latipes (Japanese rice fish) (Japanese killifish) | 23617 | 23183 |
| 48 | Eukaryota | 8355 | Xenopus laevis (African clawed frog) | 35660 | 34791 |
| 49 | Eukaryota | 8364 | Xenopus tropicalis (Western clawed frog) (Silurana tropicalis) | 22229 | 21477 |
| 50 | Eukaryota | 9031 | Gallus gallus (Chicken) | 18369 | 2457 |
| 51 | Eukaryota | 9595 | Gorilla gorilla gorilla (Western lowland gorilla) | 21783 | 21098 |
| 52 | Eukaryota | 9598 | Pan troglodytes (Chimpanzee) | 23051 | 22536 |
| 53 | Eukaryota | 9606 | Homo sapiens (Human) | 20586 | 1044 |
| 54 | Eukaryota | 9615 | Canis lupus familiaris (Dog) (Canis familiaris) | 20972 | 4638 |
| 55 | Eukaryota | 9913 | Bos taurus (Bovine) | 23841 | 19058 |
| 56 | Eukaryota | 10090 | Mus musculus (Mouse) | 21957 | 5078 |
| 57 | Eukaryota | 10116 | Rattus norvegicus (Rat) | 22870 | 10074 |
| 58 | Eukaryota | 13616 | Monodelphis domestica (Gray short-tailed opossum) | 21223 | 9042 |
| 59 | Eukaryota | 35128 | Thalassiosira pseudonana (Marine diatom) (Cyclotella nana) | 11717 | 9683 |
| 60 | Eukaryota | 36329 | Plasmodium falciparum (isolate 3D7) | 5372 | 5367 |
| 61 | Eukaryota | 39947 | Oryza sativa subsp. japonica (Rice) | 43672 | 41444 |
| 62 | Eukaryota | 44689 | Dictyostelium discoideum (Social amoeba) | 12726 | 12425 |
| 63 | Eukaryota | 45351 | Nematostella vectensis (Starlet sea anemone) | 24427 | 17020 |
| 64 | Eukaryota | 81824 | Monosiga brevicollis (Choanoflagellate) | 9188 | 8509 |
| 65 | Eukaryota | 164328 | Phytophthora ramorum (Sudden oak death agent) | 15349 | 13516 |
| 66 | Eukaryota | 184922 | Giardia intestinalis (strain ATCC 50803 / WB clone C6) (Giardia lamblia) | 4900 | 4897 |
| 67 | Eukaryota | 214684 | Cryptococcus neoformans var. neoformans serotype D (strain JEC21 / ATCC MYA-565) (Filobasidiella neoformans) | 6604 | 6515 |
| 68 | Eukaryota | 237561 | Candida albicans (strain SC5314 / ATCC MYA-2876) (Yeast) | 6035 | 5903 |
| 69 | Eukaryota | 237631 | Ustilago maydis (strain 521 / FGSC 9021) (Corn smut fungus) | 6788 | 6739 |
| 70 | Eukaryota | 284591 | Yarrowia lipolytica (strain CLIB 122 / E 150) (Yeast) (Candida lipolytica) | 6449 | 6431 |
| 71 | Eukaryota | 284812 | Schizosaccharomyces pombe (strain 972 / ATCC 24843) (Fission yeast) | 5122 | 5065 |
| 72 | Eukaryota | 321614 | Phaeosphaeria nodorum (strain SN15 / ATCC MYA-4574 / FGSC 10173) (Glume blotch fungus) (Parastagonospora nodorum) | 15998 | 15907 |
| 73 | Eukaryota | 330879 | Aspergillus fumigatus (strain ATCC MYA-4609 / CBS 101355 / FGSC A1100 / Af293) (Neosartorya fumigata) | 9647 | 9543 |
| 74 | Eukaryota | 367110 | Neurospora crassa (strain ATCC 24698 / 74-OR23-1A / CBS 708.71 / DSM 1257 / FGSC 987) | 9759 | 9697 |
| 75 | Eukaryota | 412133 | Trichomonas vaginalis (strain ATCC PRA-98 / G3) | 50190 | 44222 |
| 76 | Eukaryota | 418459 | Puccinia graminis f. sp. tritici (strain CRL 75-36-700-3 / race SCCL) (Black stem rust fungus) | 15688 | 15508 |
| 77 | Eukaryota | 559292 | Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast) | 6060 | 6039 |
| 78 | Eukaryota | 665079 | Sclerotinia sclerotiorum (strain ATCC 18683 / 1980 / Ss-1) (White mold) (Whetzelinia sclerotiorum) | 14445 | 14427 |
| 79 | Eukaryota | 684364 | Batrachochytrium dendrobatidis (strain JAM81 / FGSC 10211) (Frog chytrid fungus) | 8610 | 8120 |
| | | | | | |
| | | | SUM | 1023125 | 857750 |

**Table 1. The 79 Species Analyzed in This Study**

This table lists the 79 species included in the analysis, spanning the three domains of life (Archaea, Bacteria, Eukaryota). Columns indicate the taxonomic domain, taxonomy ID, organism name, and the number of proteins "Listed" versus "Matched." Here, "Listed" refers to the total proteins initially available in the reference proteome dataset, while "Matched" indicates the final count of proteins remaining after cross-referencing gene and protein sequences and excluding those that did not meet the selection criteria (see Section 2.2.1). The bottom row shows the summed totals across all 79 species.