The Gene's GC Content Is the Greatest Source of Inter-Species Differences in Protein Amino Acid Composition

Esumi, Genshiro

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

Abstract

Organisms synthesize proteins based on sequences of 20 amino acids specified by their genes, and protein function is determined by these amino acid sequences and compositions. Previous studies in Bacteria have shown that an organism's genomic GC content is a key determinant of the amino acid composition of its proteins. However, a more generalized behavior that includes organisms from other domains of life has remained unclear.

In this study, I performed principal component analysis (PCA) on the amino acid compositions of approximately 1.5 million proteins from 81 species spanning all three domains of life and examined how their principal component scores varied among species. The results revealed that, while the first principal component exhibited considerable variation among species, the variation in all other principal components was significantly limited.

To investigate this further, I developed a function to back-calculate the GC content of a gene from its amino acid composition under the assumption of equal usage of synonymous codons. I then compared the estimated GC content derived from this reverse transformation with the first principal component from the PCA, observing a correlation coefficient of 0.98, which indicates an almost perfect match. Because the first principal component of amino acid composition was essentially the only component that showed substantial interspecies variation, and its values strongly correlated with the back-calculated GC content, I conclude that the greatest source of diversity in protein amino acid composition lies in the gene's GC content, which is substantially governed by the organism's genomic GC content.

Keywords: Amino acid composition, GC content, TA skew, Species Difference, Diversity

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

Background

Organisms synthesize proteins based on sequences of 20 amino acids specified by their genes, and, according to Anfinsen's dogma, protein function is determined by these amino acid sequences and, to a substantial extent, by their compositions. Meanwhile, previous studies in Bacteria have shown that amino acid composition varies among species, and that these species-specific differences in amino acid composition correlate with each species' genomic GC content [1]. However, whether this relationship applies more generally across different species and domains of life has remained unclear.

Given this background, in the present study I analyzed statistical information on the amino acid compositions of roughly 1.5 million proteins from 81 organisms, spanning all three domains of life, using publicly available "reference proteomes." My aim was to investigate the source of interspecies differences in protein amino acid composition and to determine whether GC content can be regarded as the primary underlying factor across all life forms.

Subjects and Methods

Reference Proteomes and Amino Acid Compositions

For the present analysis of various organisms' exomic proteins, I used the "reference proteomes" dataset published by EMBL-EBI [2]. This dataset spans the three domains of life (Archaea, Bacteria, and Eukaryotes) and includes amino acid sequences from 1,547,370 proteins across a total of 81 different species. In this study, I analyzed each protein sequence by counting the number of each amino acid residue and then dividing by the total number of residues in that protein to obtain its amino acid composition (which sums to 1).

Principal Component Analysis of Protein Amino Acid Compositions Across All Domains

Next, I performed principal component analysis (PCA) on the amino acid compositions of the 1,547,370 proteins calculated above. Because the number of exomic protein entries varied among the 81 species, I assigned each protein a weight inversely proportional to the total number of proteins in its species. This approach ensures that species with fewer exomic proteins are not disproportionately underrepresented in the analysis. Amino acid composition encompasses 20 amino acids, yielding 19 degrees of freedom, so principal components up to the 19th component were extracted.

Distribution of Principal Component Scores by Species

From this calculation, principal component (PC) scores from the 1st through the 19th components were assigned to each of the 1,547,370 proteins. Using these scores, I examined the distributions for each species. Specifically, I employed the “Compare Densities” feature in the “Bivariate” analysis of JMP 18 to overlay distribution graphs for different species and compared their variability. I regarded those components that showed large variations among species as reflecting the principal differences in amino acid composition across organisms.

Back-Calculation of Estimated GC Content from Amino Acid Composition

The amino acid composition of a protein is nearly uniquely determined by the base sequence of its gene and by the genetic code that translates nucleotide codons into amino acids. However, when considering the reverse conversion, multiple codons (synonymous codons) can correspond to a single amino acid, meaning the back-conversion is not strictly unique. In this study, to obtain a simplified estimate, I assumed that each set of synonymous codons for a particular amino acid is used in equal proportions. Under this assumption, I created a function to back-calculate the GC content of a gene. Note that the GC content was defined as follows:

$$\text{GC Content} = \frac{G + C}{T + A + G + C}$$

Where G, C, T and A are the counts of each base in the amino acid–coding region of the gene. Using this approach, I calculated the back-calculated (estimated) GC content for every protein and examined its correlation with the PC scores from the 1st through the 19th components. A high correlation would indicate that the corresponding principal component is closely linked to the gene’s GC content.

Back-Calculation of Estimated TA Skew and GC Skew

A gene is composed of a sequence of four bases, and because the total amount of each base is fixed, there are effectively three degrees of freedom in base composition. Accordingly, in addition to GC content, two other independent variables can be defined. In line with previous studies, this report uses TA skew and GC skew [3]. As with the back-calculation function for GC content, these indices were also back-calculated under the assumption of uniform usage of synonymous codons. I developed these back-calculation functions under the above conditions. Note that the TA skew and GC skew were defined as follows:

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

Data Processing

All data processing and table creation in this study were performed using Microsoft Excel (Microsoft Corp., Redmond, WA, USA), and all graphs were generated with JMP 18 (SAS Institute Inc., Cary, NC, USA).

Results

Overview of the Studied Species and Protein Counts

Table 1 lists the 81 species included in this analysis, along with their IDs, taxonomic domain, cell organization type, and the number of exons/proteins in each reference proteome dataset [2].

Principal Component Analysis Results (Contribution Ratios and Eigenvectors)

Table 2 presents the results of the principal component analysis described in the Methods section. It shows the contribution ratios, cumulative contribution ratios, and eigenvectors for the first through the 19th principal components.

Comparison of Principal Component Score Distributions by Species

Figure 1 overlays the distributions of each species for the first through the 19th principal components, with their respective contribution ratios indicated. While the first principal component exhibits considerable variation among species, and the third principal component shows a smaller degree of variation, some species display a bimodal pattern in the second principal component. Nonetheless, all other principal components exhibit nearly identical distributions across species.

To further investigate the behavior of the first, second, and third principal components, I compared them by arranging each species side by side rather than overlaying them; the results are presented in **Figure 2**. The first principal component shows large interspecies variability primarily among Archaea, Bacteria, and certain eukaryotes. In the second principal component, a bimodal peak was observed mainly among Archaea and Bacteria; however, the overall distribution trend, as indicated by the box-and-whisker plots, was nearly constant across species. As for the third principal component, interspecies variability was evident, but overall it appeared to reflect differences among domains rather than those at the species level.

Correlation Analysis of Back-Calculated GC Content, TA Skew, GC Skew, and Principal Component Scores

Figure 3 provides a comprehensive set of two-variable correlation plots between the back-calculated GC content, TA skew, GC skew, and the 1st through 19th principal component scores for all 1.5 million proteins. Because the polarity of each principal component depends on the analysis protocol, the sign of the correlation coefficient is not meaningful here—only its absolute value is relevant. Among these relationships, focusing on those with a correlation coefficient (|r|) of at least 0.5, the strongest correlation (|r| = 0.98) was observed between back-calculated GC content and the first principal component. Additionally, back-calculated TA skew correlated with the second principal component, and back-calculated GC skew correlated with the third principal component. No other pairs showed a correlation coefficient above 0.5.

Discussion

It has long been known, particularly in Bacteria, that the amino acid composition of exomic proteins varies among species and that this phenomenon is attributable to differences in genomic GC content. However, it has remained unclear whether this principle extends to eukaryotic organisms. In this study, using exome datasets from 81 species across all three domains of life (totaling approximately 1.5 million proteins), I examined whether the "species-specific diversity" in amino acid compositions can be statistically explained by factors such as the gene's base composition (e.g., GC content).

To extract interspecies variability, I performed principal component analysis (PCA) on the 20 amino acid composition values of these 1.5 million proteins, weighting each protein by the reciprocal of the total number of proteins in its exome. I then investigated whether the resulting principal component (PC) scores, from PC1 to PC19, captured differences among species. As shown in **Figure 1**, PC1 exhibited large variation among species, whereas PC3 showed moderate variation. Additionally, as illustrated in **Figure 2**, PC1 reflected species-level differences, while the variation in PC3—though smaller in magnitude than that of PC1—appears to reflect differences among the domains (Archaea, Bacteria, and Eukaryotes). Consequently, I concluded that species-specific disparities in amino acid composition are effectively captured primarily by PC1.

Next, I sought to determine whether these species-level differences in amino acid composition indeed stem from the nucleotide composition of the gene. Although the dataset I used contained only amino acid sequences, I back-calculated the gene's nucleotide composition from these

sequences and compared these “back-calculated” values with the principal component scores. The result was that PC1, which I consider to reflect interspecies variation, showed a correlation coefficient of 0.98 with the back-calculated GC content, indicating the two are almost identical. Hence, the hypothesis that species-level diversity in amino acid composition is dictated by the gene’s GC content appears to be valid. Naturally, this back-calculated GC content was derived under the assumption of uniform usage of all synonymous codons, and examining actual GC content would require an analysis of how synonymous codons are truly utilized. However, based on my previous reports indicating that synonymous codon choice is strongly influenced by the gene’s GC content [4], it seems all the more reasonable to attribute interspecies differences in amino acid composition to differences in gene GC content, given the strong correlation between the back-calculated GC content and PC1.

Furthermore, I demonstrated that back-calculated TA skew correlates with PC2, whereas back-calculated GC skew correlates with PC3. In my earlier work, I showed that a gene’s TA skew aligns with the hydrophobicity of the amino acid sequence it encodes, such that high-TA-skew regions tend to encode transmembrane domains, whereas low-TA-skew regions tend to encode intrinsically disordered regions [3]. Accordingly, it seems plausible that PC2 reflects the distributions of hydrophobic and hydrophilic amino acids in protein domains, corresponding to the back-calculated TA skew.

Likewise, the association between PC3 and back-calculated GC skew was also observed. While protein diversity as a whole is known to be immense, the first three principal components alone, whose cumulative contribution ratio is around 36%, appear to be underpinned by relatively low-dimensional characteristics of gene base composition. Because higher-dimensional sequences carry a higher risk of divergence through random mutations, it seems possible that encoding 20 amino acids with four types of nucleotides acts in part to mitigate this risk.

Conclusion

By performing a statistical analysis of the amino acid compositions of approximately 1.5 million proteins from 81 species spanning the three domains of life, I have demonstrated that the primary driver of interspecies differences is variation in GC content at the gene level.

Reference

1. **Du, M.-Z., Zhang, C., Wang, H., Liu, S., Wei, W., & Guo, F.-B.** (2018). The GC Content as a Main Factor Shaping the Amino Acid Usage During Bacterial Evolution Process. *Frontiers Media SA*. https://doi.org/10.3389/fmicb.2018.02948

2. **EMBL-EBI.** (2024). Reference Proteomes (Release 2024_02) [Database]. Retrieved January 7, 2025, from https://www.ebi.ac.uk/reference_proteomes/

3. **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

4. **Esumi, G.** (2023). The Synonymous Codon Usage of a Protein Gene Is Primarily Determined by the Guanine + Cytosine Content of the Individual Gene Rather Than the Species to Which It Belongs To Synthesize Proteins With a Balanced Amino Acid Composition. *Jxiv*. https://doi.org/10.51094/jxiv.561

No.	Scientific Name	Organism ID	Domain	Cell Organization	Protein Count
1	*Halobacterium salinarum* (strain ATCC 700922 / JCM 11081 / NRC-1) (Halobacterium halobium)	64091	archaea	unicellular	2427
2	*Thermococcus kodakarensis* (strain ATCC BAA-918 / JCM 12380 / KOD1) (Pyrococcus kodakaraensis (strain KOD1))	69014	archaea	unicellular	2301
3	*Methanosarcina acetivorans* (strain ATCC 35395 / DSM 2834 / JCM 12185 / C2A)	188937	archaea	unicellular	4468
4	*Methanocaldococcus jannaschii* (strain ATCC 43067 / DSM 2661 / JAL-1 / JCM 10045 / NBRC 100440) (Methanococcus jannaschii)	243232	archaea	unicellular	1787
5	*Saccharolobus solfataricus* (strain ATCC 35092 / DSM 1617 / JCM 11322 / P2) (Sulfolobus solfataricus)	273057	archaea	unicellular	2937
6	*Korarchaeum cryptofilum* (strain OPF8)	374847	archaea	unicellular	1602
7	*Nitrosopumilus maritimus* (strain SCM1)	436308	archaea	unicellular	1795
8	*Mycobacterium tuberculosis* (strain ATCC 25618 / H37Rv)	83332	bacteria	unicellular	3999
9	*Escherichia coli* (strain K12)	83333	bacteria	unicellular	4416
10	*Helicobacter pylori* (strain ATCC 700392 / 26695) (Campylobacter pylori)	85962	bacteria	unicellular	1554
11	*Streptomyces coelicolor* (strain ATCC BAA-471 / A3(2) / M145)	100226	bacteria	unicellular	8039
12	*Neisseria meningitidis* serogroup B (strain MC58)	122586	bacteria	unicellular	2001
13	*Leptospira interrogans* serogroup Icterohaemorrhagiae serovar Lai (strain 56601)	189518	bacteria	unicellular	3676
14	*Fusobacterium nucleatum* subsp. nucleatum (strain ATCC 25586 / DSM 15643 / BCRC 10681 / CIP 101130 / JCM 8532 / KCTC 2640 / LMG 13131 / VPI 4355)	190304	bacteria	unicellular	2046
15	*Pseudomonas aeruginosa* (strain ATCC 15692 / DSM 22644 / CIP 104116 / JCM 14847 / LMG 12228 / 1C / PRS 101 / PAO1)	208964	bacteria	unicellular	5564
16	*Bacillus subtilis* (strain 168)	224308	bacteria	unicellular	4267
17	*Aquifex aeolicus* (strain VF5)	224324	bacteria	unicellular	1553
18	*Bradyrhizobium diazoefficiens* (strain JCM 10833 / BCRC 13528 / IAM 13628 / NBRC 14792 / USDA 110)	224911	bacteria	unicellular	8253
19	*Bacteroides thetaiotaomicron* (strain ATCC 29148 / DSM 2079 / JCM 5827 / CCUG 10774 / NCTC 10582 / VPI-5482 / E50)	226186	bacteria	unicellular	4782
20	*Rhodopirellula baltica* (strain DSM 10527 / NCIMB 13988 / SH1)	243090	bacteria	unicellular	7271
21	*Deinococcus radiodurans* (strain ATCC 13939 / DSM 20539 / JCM 16871 / CCUG 27074 / LMG 4051 / NBRC 15346 / NCIMB 9279 / VKM B-1422 / R1)	243230	bacteria	unicellular	3084
22	*Geobacter sulfurreducens* (strain ATCC 51573 / DSM 12127 / PCA)	243231	bacteria	unicellular	3402
23	*Mycoplasma genitalium* (strain ATCC 33530 / DSM 19775 / NCTC 10195 / G37) (Mycoplasmoides genitalium)	243273	bacteria	unicellular	483
24	*Thermotoga maritima* (strain ATCC 43589 / DSM 3109 / JCM 10099 / NBRC 100826 / MSB8)	243274	bacteria	unicellular	1852
25	*Gloeobacter violaceus* (strain ATCC 29082 / PCC 7421)	251221	bacteria	unicellular	4406
26	*Chlamydia trachomatis* (strain D/UW-3/Cx)	272561	bacteria	unicellular	895
27	*Thermodesulfovibrio yellowstonii* (strain ATCC 51303 / DSM 11347 / YP87)	289376	bacteria	unicellular	1982
28	*Chloroflexus aurantiacus* (strain ATCC 29366 / DSM 635 / J-10-fl)	324602	bacteria	unicellular	3850
29	*Dictyoglomus turgidum* (strain DSM 6724 / Z-1310)	515635	bacteria	unicellular	1743
30	*Synechocystis sp.* (strain PCC 6803 / Kazusa)	1111708	bacteria	unicellular	3508
31	*Chlamydomonas reinhardtii* (Chlamydomonas smithii)	3055	eukaryota	unicellular	18832
32	*Physcomitrium patens* (Spreading-leaved earth moss) (Physcomitrella patens)	3218	eukaryota	multicellular	47782
33	*Arabidopsis thaliana* (Mouse-ear cress)	3702	eukaryota	multicellular	41596
34	*Zea mays* (Maize)	4577	eukaryota	multicellular	63281
35	*Leishmania major*	5664	eukaryota	unicellular	8038
36	*Paramecium tetraurelia*	5888	eukaryota	unicellular	39461
37	*Caenorhabditis elegans*	6239	eukaryota	multicellular	28553
38	*Helobdella robusta* (Californian leech)	6412	eukaryota	multicellular	23328
39	*Ixodes scapularis* (Black-legged tick) (Deer tick)	6945	eukaryota	multicellular	20496
40	*Tribolium castaneum* (Red flour beetle)	7070	eukaryota	multicellular	18505
41	*Anopheles gambiae* (African malaria mosquito)	7165	eukaryota	multicellular	14411
42	*Drosophila melanogaster* (Fruit fly)	7227	eukaryota	multicellular	23539
43	*Ciona intestinalis* (Transparent sea squirt) (Ascidia intestinalis)	7719	eukaryota	multicellular	17311
44	*Branchiostoma floridae* (Florida lancelet) (Amphioxus)	7739	eukaryota	multicellular	38648
45	*Lepisosteus oculatus* (Spotted gar)	7918	eukaryota	multicellular	22463
46	*Danio rerio* (Zebrafish) (Brachydanio rerio)	7955	eukaryota	multicellular	46840
47	*Oryzias latipes* (Japanese rice fish) (Japanese killifish)	8090	eukaryota	multicellular	36138
48	*Xenopus laevis* (African clawed frog)	8355	eukaryota	multicellular	61769
49	*Xenopus tropicalis* (Western clawed frog) (Silurana tropicalis)	8364	eukaryota	multicellular	37693
50	*Gallus gallus* (Chicken)	9031	eukaryota	multicellular	43968
51	*Macaca mulatta* (Rhesus macaque)	9544	eukaryota	multicellular	44416
52	*Gorilla gorilla gorilla* (Western lowland gorilla)	9595	eukaryota	multicellular	44726
53	*Pan troglodytes* (Chimpanzee)	9598	eukaryota	multicellular	48794
54	*Homo sapiens* (Human)	9606	eukaryota	multicellular	104573
55	*Canis lupus familiaris* (Dog) (Canis familiaris)	9615	eukaryota	multicellular	43672
56	*Bos taurus* (Bovine)	9913	eukaryota	multicellular	37871
57	*Mus musculus* (Mouse)	10090	eukaryota	multicellular	63289
58	*Rattus norvegicus* (Rat)	10116	eukaryota	multicellular	49582
59	*Monodelphis domestica* (Gray short-tailed opossum)	13616	eukaryota	multicellular	36221
60	*Thalassiosira pseudonana* (Marine diatom) (Cyclotella nana)	35128	eukaryota	unicellular	11612
61	*Daphnia magna*	35525	eukaryota	multicellular	26600
62	*Plasmodium falciparum* (isolate 3D7)	36329	eukaryota	unicellular	5369
63	*Oryza sativa* subsp. japonica (Rice)	39947	eukaryota	multicellular	49224
64	*Dictyostelium discoideum* (Social amoeba)	44689	eukaryota	unicellular	12746
65	*Nematostella vectensis* (Starlet sea anemone)	45351	eukaryota	multicellular	24445
66	*Monosiga brevicollis* (Choanoflagellate)	81824	eukaryota	unicellular	9156
67	*Phytophthora ramorum* (Sudden oak death agent)	164328	eukaryota	unicellular	15349
68	*Giardia intestinalis* (strain ATCC 50803 / WB clone C6) (Giardia lamblia)	184922	eukaryota	unicellular	4900
69	*Cryptococcus neoformans* var. neoformans serotype D (strain JEC21 / ATCC MYA-565) (Filobasidiella neoformans)	214684	eukaryota	unicellular	6746
70	*Candida albicans* (strain SC5314 / ATCC MYA-2876) (Yeast)	237561	eukaryota	unicellular	6037
71	*Ustilago maydis* (strain 521 / FGSC 9021) (Corn smut fungus)	237631	eukaryota	unicellular	6805
72	*Yarrowia lipolytica* (strain CLIB 122 / E 150) (Yeast) (Candida lipolytica)	284591	eukaryota	unicellular	6454
73	*Schizosaccharomyces pombe* (strain 972 / ATCC 24843) (Fission yeast)	284812	eukaryota	unicellular	5132
74	*Phaeosphaeria nodorum* (strain SN15 / ATCC MYA-4574 / FGSC 10173) (Glume blotch fungus) (Parastagonospora nodorum)	321614	eukaryota	unicellular	15998
75	*Aspergillus fumigatus* (strain ATCC MYA-4609 / CBS 101355 / FGSC A1100 / Af293) (Neosartorya fumigata)	330879	eukaryota	unicellular	9648
76	*Neurospora crassa* (strain ATCC 24698 / 74-OR23-1A / CBS 708.71 / DSM 1257 / FGSC 987)	367110	eukaryota	unicellular	10266
77	*Trichomonas vaginalis* (strain ATCC PRA-98 / G3)	412133	eukaryota	unicellular	50190
78	*Puccinia graminis* f. sp. *tritici* (strain CRL 75-36-700-3 / race SCCL) (Black stem rust fungus)	418459	eukaryota	unicellular	15808
79	*Saccharomyces cerevisiae* (strain ATCC 204508 / S288c) (Baker's yeast)	559292	eukaryota	unicellular	6091
80	*Sclerotinia sclerotiorum* (strain ATCC 18683 / 1980 / Ss-1) (White mold) (Whetzelinia sclerotiorum)	665079	eukaryota	unicellular	14445
81	*Batrachochytrium dendrobatidis* (strain JAM81 / FGSC 10211) (Frog chytrid fungus)	684364	eukaryota	unicellular	8610
				Total	1547370

Table 1. Overview of the Studied Species and Protein Counts

Table 1 lists the 81 species included in this analysis, along with their organism IDs, taxonomic domain, cell organization type, and the number of exons/proteins in each reference proteome dataset.

Componnent	Contribution Ratio	Cumulative Contribution Ratio	Ala	Cys	Asp	Glu	Phe	Gly	His	Ile	Lys	Leu	Met	Asn	Pro	Gln	Arg	Ser	Thr	Val	Trp	Tyr
PC 1	16.1	16.1	-0.383	-0.034	0.038	0.446	0.288	-0.270	-0.106	0.283	0.372	0.025	0.054	0.244	-0.302	-0.046	-0.276	-0.054	-0.095	-0.128	-0.105	0.270
PC 2	10.4	26.5	0.308	-0.031	-0.258	-0.316	0.368	0.214	-0.145	0.229	-0.216	0.326	0.127	-0.149	-0.120	-0.331	-0.130	-0.185	-0.050	0.329	0.248	0.167
PC 3	9.5	36.0	-0.187	0.292	-0.270	-0.442	0.177	-0.107	0.216	-0.042	-0.176	-0.025	-0.041	0.202	0.211	0.154	-0.172	0.328	0.275	-0.243	0.122	0.124
PC 4	6.6	42.6	-0.130	0.150	-0.140	0.133	0.125	-0.256	0.331	-0.102	-0.022	0.382	0.136	-0.177	-0.024	0.389	0.349	-0.134	-0.469	-0.249	0.258	0.036
PC 5	6.1	48.7	-0.159	0.278	0.303	-0.001	0.009	0.238	0.283	-0.099	0.035	-0.474	-0.008	0.024	0.038	-0.322	0.074	-0.219	-0.096	0.055	0.276	0.246
PC 6	5.2	53.9	-0.007	0.288	-0.270	-0.001	-0.091	0.052	-0.008	0.006	0.132	-0.225	0.514	-0.075	-0.062	0.081	-0.058	-0.024	0.012	0.103	-0.494	-0.584
PC 7	4.7	58.6	0.085	-0.330	-0.139	-0.058	0.014	0.215	-0.274	-0.024	0.027	-0.240	0.130	0.142	0.277	0.318	-0.036	-0.121	-0.230	-0.834	0.255	0.275
PC 8	4.7	63.3	0.072	-0.268	0.335	-0.085	-0.018	-0.277	0.201	-0.090	-0.232	0.055	0.132	0.069	-0.237	0.252	-0.114	-0.206	0.396	0.064	0.300	0.064
PC 9	4.3	67.6	-0.154	-0.055	0.112	0.161	0.067	-0.205	-0.454	-0.074	0.045	-0.053	0.199	-0.137	0.040	-0.381	0.256	0.112	0.141	-0.137	0.425	-0.113
PC 10	4.1	71.7	0.053	0.320	-0.043	0.081	-0.100	0.037	-0.463	-0.142	0.052	0.060	-0.222	0.080	-0.224	0.286	-0.178	-0.288	0.032	0.076	0.355	-0.106
PC 11	4.0	75.7	-0.066	-0.207	-0.494	0.196	-0.086	-0.250	0.105	0.068	0.266	-0.122	-0.149	-0.188	0.239	-0.010	0.209	-0.044	0.468	0.137	0.165	0.244
PC 12	3.6	79.3	-0.154	0.013	0.216	0.147	0.198	-0.246	-0.072	-0.110	-0.031	-0.011	0.037	-0.275	0.541	0.153	-0.443	0.070	-0.232	0.374	-0.024	0.044
PC 13	3.5	82.8	-0.112	-0.256	-0.155	0.039	-0.042	0.155	0.357	0.067	0.200	-0.154	-0.092	0.134	-0.153	0.023	-0.212	0.216	-0.194	0.154	0.143	-0.484
PC 14	3.4	86.2	-0.162	-0.099	-0.095	-0.198	-0.020	-0.204	-0.149	-0.046	-0.124	-0.228	-0.046	0.208	-0.142	0.245	0.486	0.160	-0.175	0.357	-0.079	0.217
PC 15	3.1	89.3	-0.091	0.069	0.162	-0.210	0.160	-0.085	-0.094	0.235	-0.041	-0.124	-0.024	0.260	0.320	0.011	0.260	-0.299	0.034	-0.062	-0.026	-0.543
PC 16	3.1	92.4	-0.037	-0.069	-0.011	0.143	0.665	0.200	0.034	-0.187	0.045	-0.264	-0.107	-0.378	-0.318	0.250	0.071	0.077	0.168	-0.077	-0.241	-0.036
PC 17	2.8	95.2	-0.673	-0.023	0.089	0.113	-0.228	0.120	-0.030	0.236	-0.246	0.190	0.026	-0.191	0.045	0.296	0.092	-0.005	0.179	0.023	-0.019	-0.025
PC 18	2.5	97.7	-0.259	-0.099	-0.117	0.174	0.143	0.290	-0.010	-0.671	0.027	0.281	0.088	0.139	0.072	-0.479	-0.017	-0.131	0.114	0.112	-0.206	-0.028
PC 19	2.3	100.0	0.142	0.063	-0.213	0.387	0.097	-0.071	0.045	0.388	-0.648	-0.195	0.004	0.258	-0.010	-0.260	-0.070	0.031	-0.026	-0.020	0.011	-0.004

Table 2. Principal Component Analysis Results: Contribution Ratios, Cumulative Contribution Ratios, and Eigenvectors for the First Through 19th Components

This table shows the principal component analysis (PCA) results derived from the amino acid compositions of approximately 1.5 million proteins, spanning 81 species across all three domains of life. The first column lists the principal components (PC1 to PC19). The second and third columns indicate each component's contribution ratio and its cumulative contribution ratio. The following columns display the component loadings (eigenvector coefficients) for all 20 amino acids. Positive and negative values represent positive and negative contributions, respectively, of each amino acid to the given principal component.

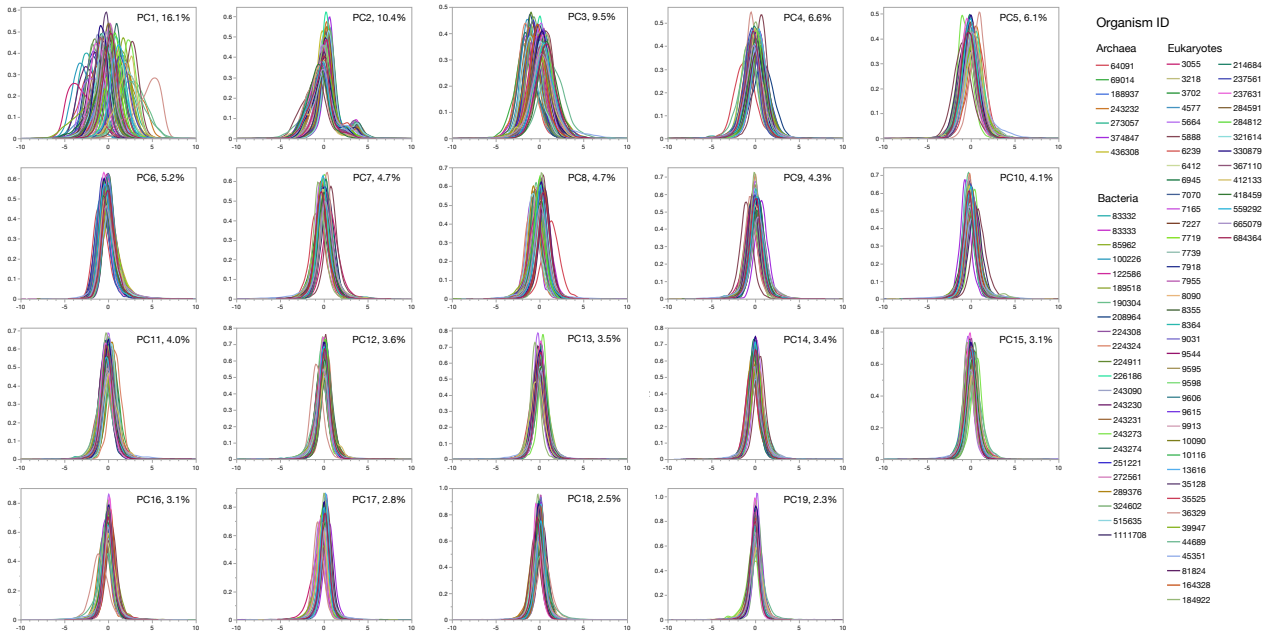

Figure 1. Distributions of the 1st Through 19th Principal Component Scores Across 81 Species

Each panel shows overlaid density plots for a specific principal component, with the component number and the corresponding contribution ratio indicated in the upper-right corner of each panel. The first principal component (PC1) exhibits substantial interspecies variability, while PC3 shows somewhat smaller variability; the remaining principal components display nearly identical distributions across all species. Additionally, some species exhibit a bimodal pattern in the second principal component (PC2).

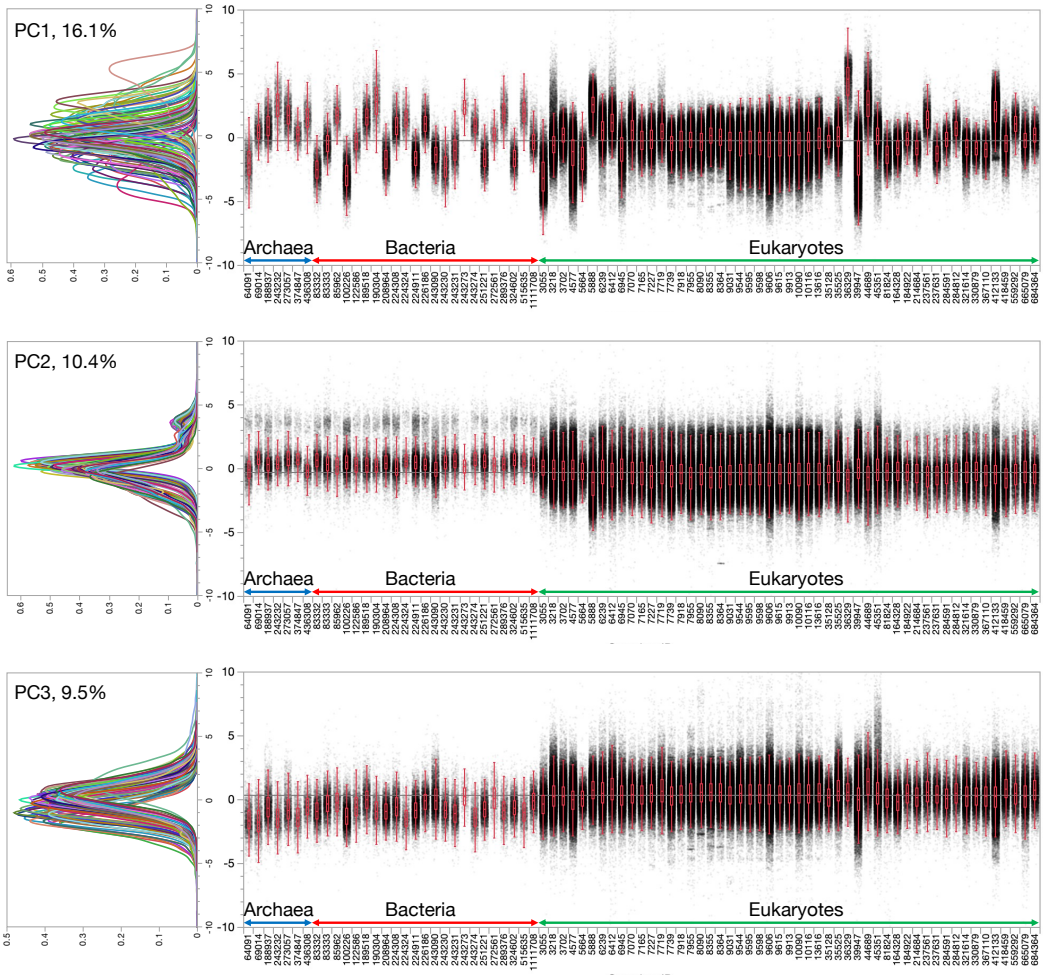

Figure 2. Title: Side-by-Side Comparison of PC1, PC2, and PC3 Distributions by Species

To further investigate the behavior of the first three principal components, each species is plotted individually rather than using overlapping distributions. PC1 shows large variability primarily among Archaea, Bacteria, and some eukaryotic species. In PC2, a bimodal pattern is observed mainly in Archaea and Bacteria, although the overall distribution trend remains nearly constant across species. PC3 reflects distinct domain-level differences, despite noticeable interspecies variation.

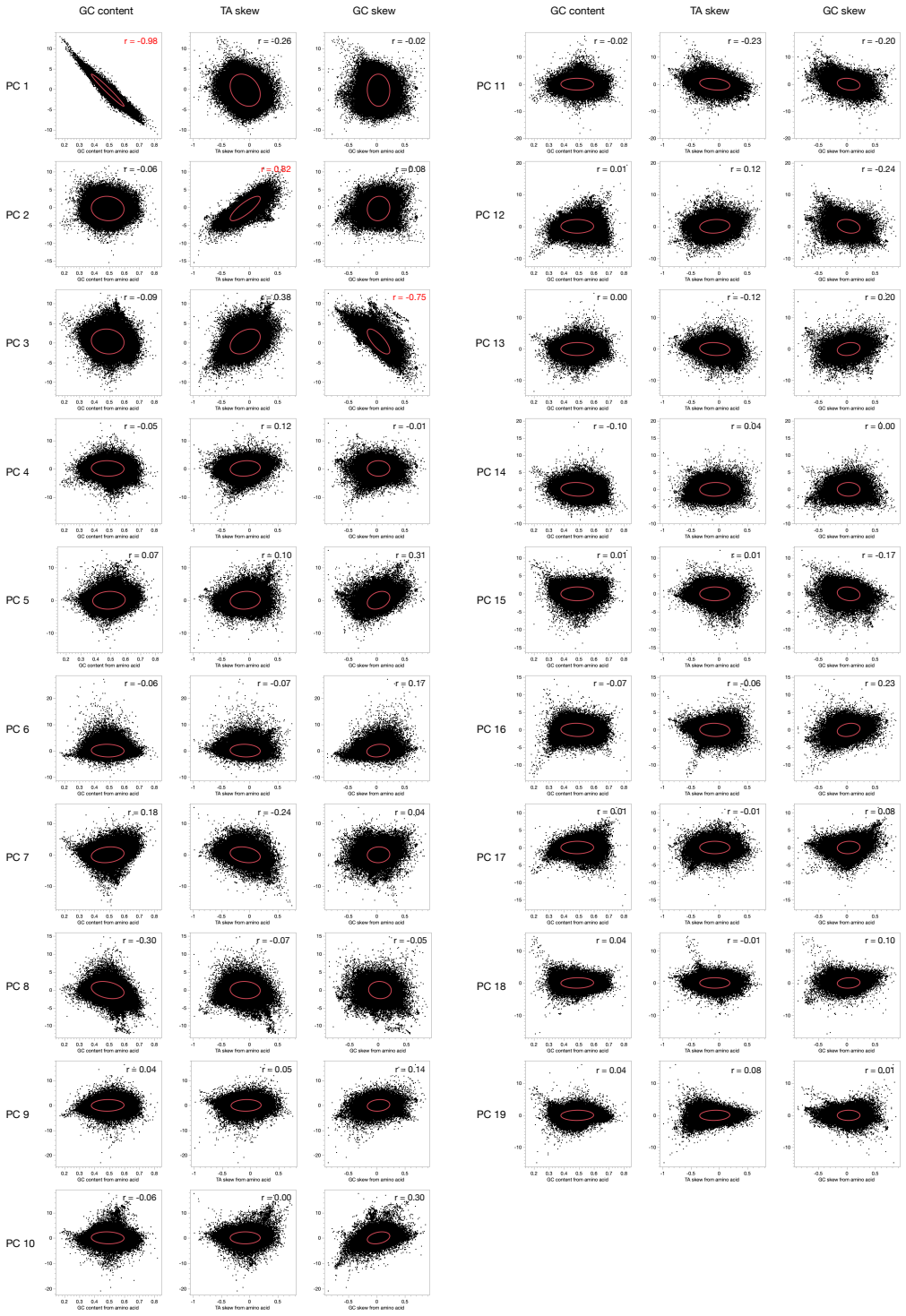

Figure 3. Correlation Analysis Between Back-Calculated GC Content, TA Skew, GC Skew, and Principal Components

These scatter plots show the two-variable correlations for all 1.5 million proteins, comparing each of the 1st through 19th principal component (PC) scores with back-calculated GC content, TA skew, and GC skew. The strongest correlation ($|r| = 0.98$) is observed between back-calculated GC content and PC1, followed by a correlation between TA skew and PC2, and between GC skew and PC3. No other correlations exceed $|r| = 0.5$.