Amino-acid composition is a selection target for coding-sequence retention: evidence from out-of-frame translation comparisons in *Escherichia coli*

Esumi, Genshiro

Department of Pediatric Surgery,

Hospital of the University of Occupational and Environmental Health, Kitakyushu, Japan

Abstract

Protein evolution is a major engine of biological diversification, but the selective pressures that determine which sequences are retained as genes remain debated. In previous work, we showed that, across the three domains of life, **per-protein amino-acid residue fractions** in reference proteomes —defined here as the set of proteins encoded by the annotated coding sequences (CDS) of a reference genome—consistently form species-specific, bell-shaped distributions well approximated by binomial expectations for each of the 20 amino acids. If a genome can, in principle, encode a wide range of amino-acid compositions yet the realized set of coding sequences occupies only a narrow region of that space, this would imply that amino-acid composition itself is a target of selection during gene retention. Here we test this idea in *Escherichia coli*. Using this definition, we compared the **distributions of per-CDS residue fractions** for the reference proteome's native (+1) translations with those for **genome-encoded out-of-frame translations** obtained by re-parsing the **unaltered** CDS in non-native frames (+2, +3 on the plus strand; −1, −2, −3 on the reverse complement). We then located the reference proteome within the composition space spanned by these alternatives. The reference proteome was concentrated within a markedly **narrower, shifted region** of composition space than that spanned by the out-of-frame translations, with especially strong separations for **cysteine, aspartate, glutamate, and arginine**. Thus, despite the genome's capacity to encode diverse compositions, the reference proteome lies within a restricted subset—**consistent with amino-acid residue composition being an important target of selection** in coding-sequence retention.

Keywords: Amino-acid composition, Amino-acid composition space, De novo gene, Reference proteome, Mutual-constraint hypothesis

E-mail: esumi@clnc.uoeh-u.ac.jp

Introduction

Protein evolution, driven by genetic mutation, is a major engine of biological diversification, yet the selective pressures that determine which of the resulting sequences are retained as genes remain debated [1,2,3,4,5]. In previous work, we showed that, across the three domains of life, **per-protein amino-acid residue fractions** in reference proteomes—here defined as the set of proteins encoded by the annotated coding sequences (CDS) of a reference genome—**consistently** form species-specific, bell-shaped distributions well approximated by binomial expectations for each of the 20 amino acids [6,7]. Notably, the modal (peak) values of these distributions differed across species; although not definitive, this pattern suggests species-specific constraints on amino-acid composition. If it can be established that a genome can, in principle, encode a **wide range of amino-acid compositions** yet a species' reference proteome occupies only a **narrow region of that space**, this would necessarily imply that amino-acid composition is itself a **target of selection during gene retention**.

Protein-coding sequences are publicly available as CDS entries in genomic databases [8]. Because DNA is translated to amino acids via the genetic code, re-parsing an **unaltered** CDS in alternative reading frames—including on the reverse-complement strand—yields **genome-encoded out-of-frame translations**. Although these predicted polypeptides are ordinarily not expressed, they provide a tractable, internally controlled proxy for the spectrum of amino-acid compositions that the same genome could, in principle, produce. Comparing the composition profile realized by the reference proteome with that of these out-of-frame alternatives therefore offers a direct way to ask whether proteome-level composition is constrained relative to genome-encoded possibilities.

Here, using *Escherichia coli* K-12 as a case study, we implement this comparison [8]. We treat the set of proteins encoded by the annotated CDS as the **reference (native +1) proteome**, translate each CDS in its native frame, and—**without altering the nucleotide sequence**—generate five out-of-frame translations by re-parsing the same sequences in the +2 and +3 frames on the plus strand and the −1, −2, and −3 frames on the reverse-complement strand [9]. We then compared the distributions of **CDS-level amino-acid fractions** between the **native translations** and the **out-of-frame sets**, for each of the 20 amino acids. This design preserves nucleotide-level composition and local sequence context while altering only the reading frame, allowing us to test whether the *E. coli* reference proteome occupies a **restricted subset** of the amino-acid composition space accessible from its genome.

Materials and Methods

Organism and CDS dataset

As a representative species we analyzed *Escherichia coli* K-12. Coding sequences (CDS) were obtained from NCBI RefSeq for the K-12 reference genome assembly **ASM584v2** (RefSeq assembly accession **GCF_000005845.2**). We downloaded the RefSeq "**CDS from genomic**" FASTA file and used this annotated CDS set for all analyses [8].

Construction of reading-frame alternatives

We adopted the conventional six-frame nomenclature: **+1, +2, +3** denote the three frames on the plus strand (starting at nucleotide positions 1, 2, and 3 of the CDS), and **−1, −2, −3** denote the three frames on the reverse-complement strand (starting at positions 1, 2, and 3 of the reverse-complement sequence) [9].

For each annotated CDS, the **native frame** was treated as **+1**. The **five non-native frames** (**+2, +3, −1, −2, −3**) were generated by re-parsing the unaltered nucleotide sequence in the specified strand/ frame. No insertions or deletions were introduced; only strand and frame were changed. Terminal overhangs shorter than a full codon were discarded so that only complete, non-overlapping triplets were counted. Throughout, "**reference proteome**" refers to the +1 translations, and "**out-of-frame sets**" to the +2, +3, −1, −2, and −3 translations.

Translation mapping and counting

Triplets were mapped to amino acids using **NCBI translation table 1** (the Standard Code) [10]. For *E. coli* K-12, the customary choice is **NCBI translation table 11** (Bacterial, Archaeal, and Plant Plastid Code), whose only relevant difference here is the **optional start-codon recoding**: at the initiation position certain alternative starts are translated as Met (e.g., **GTG/GUG**, **TTG/UUG**, and occasionally **CTG/CUG**). In our dataset, the vast majority of CDS began with **AUG/ATG** (**n = 3,893** of **4,318**), with smaller numbers beginning with **GUG/GTG** (**n = 338**) or **UUG/TTG** (**n = 80**). Because start-codon recoding affects **at most one residue per CDS**, and to apply a uniform rule across native and out-of-frame readings (which lack an initiation context), we **did not apply start-specific recoding**: initial **GUG/GTG** and **UUG/TTG** were mapped to **Val** and **Leu**, respectively, just as at internal positions. Functionally, this choice is equivalent to using table 11 with start-codon recoding disabled.

The three termination codons (**TAA, TAG, TGA**) were treated as stops and **excluded from counts**. Recoding events such as selenocysteine (**UGA**) and programmed frameshifting were **not modeled**. For each CDS in each frame set, we counted occurrences of the **20 canonical amino acids** across all complete, non-overlapping codons.

Amino-acid composition per CDS

CDS-level amino-acid composition was computed as fractions. For amino acid a in **CDS** i,

$$f_{i,a} = \frac{\text{count}_i(a)}{\sum_{b \in \{20 \text{ amino acids}\}} \text{count}_i(b)},$$

yielding a **20-component composition vector** that **sums to 1** for every CDS under the native frame and under each non-native frame. (Stops were excluded from both numerator and denominator.)

Software and comparisons

Initial parsing of the FASTA file and generation of frame-specific codon counts were performed in **Microsoft Excel® for Mac (v16.100.4)**. Composition vectors were imported into **JMP® Pro 18.2.0** (SAS Institute) for tabulation and visualization. For each residue, we compared the **distributions of per-CDS residue fractions** for the native reference proteome (+1 frame) with those for each out-of-frame set (+2, +3, −1, −2, −3). Within each panel, density curves were **normalized to unit area** to enable cross-frame comparison of distributional shape and location. Apart from the exclusion of stops and the omission of CDS–frame combinations with no sense codons (see above), no additional filtering was applied; software defaults were used unless otherwise noted.

Results

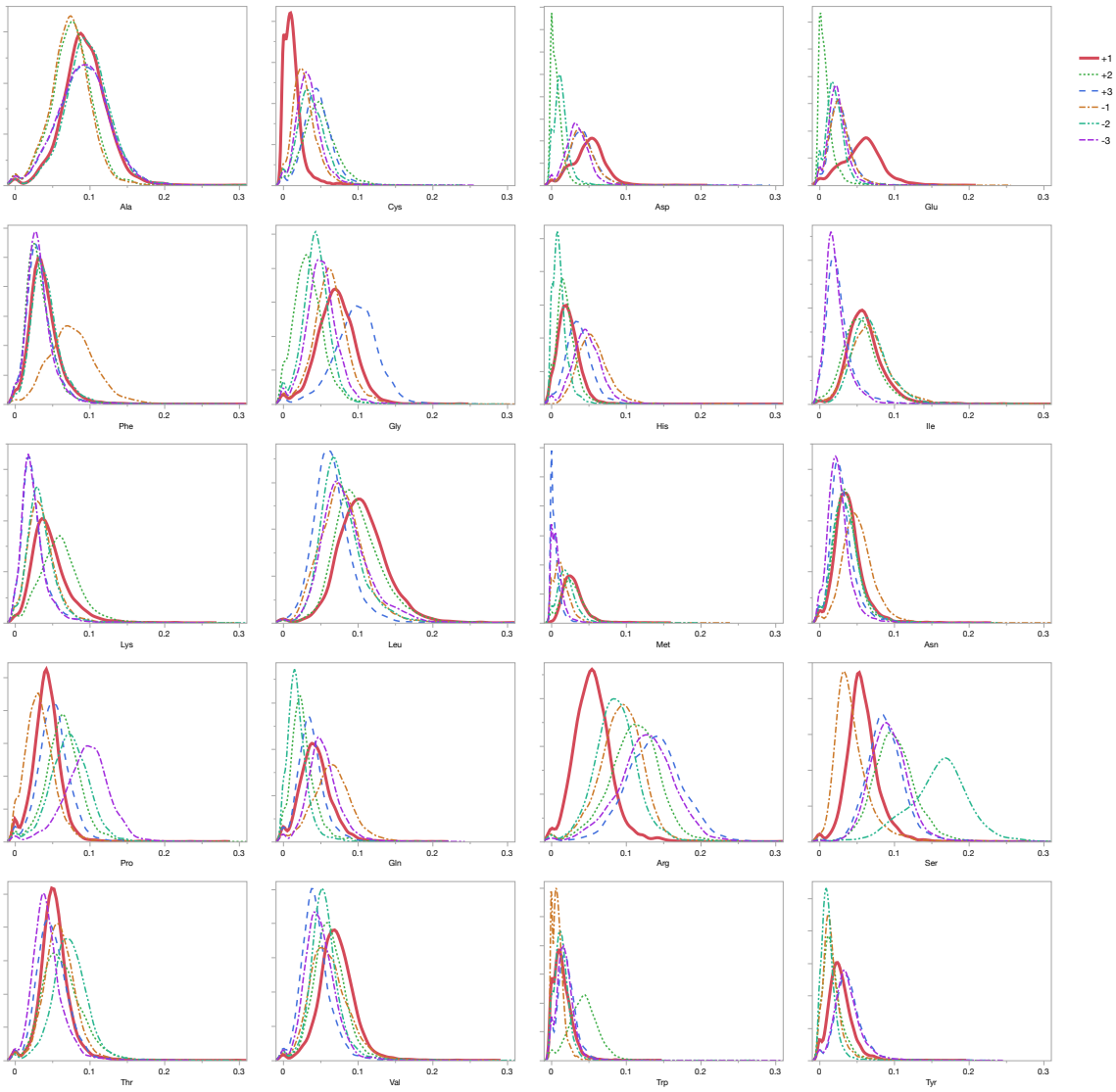
Reference-proteome–wide per-CDS residue-fraction distributions under native and out-of-frame readings

The *Escherichia coli* K-12 dataset comprised **4,318** annotated coding sequences (CDS; mean length **932.6 bp**, median **825 bp**). For each CDS, we computed the per-CDS fraction of each of the 20 amino acids in the native frame (+1) and in the five non-native frames (+2, +3, −1, −2, −3) by re-parsing the unaltered nucleotide sequence. For every residue, we visualized the distribution of these per-CDS fractions for the native reference proteome (solid red line; Figure 1) overlaid with those for each out-of-frame set (distinct dashed lines; Figure 1). Within each panel, curves were normalized to unit area to facilitate cross-frame comparison of distributional shape and location.

Across residues, distributions were broadly bell-shaped, yet systematic, frame-dependent shifts were evident. Notably, the native reference proteome exhibited consistently **lower cysteine and arginine** fractions than any of the five out-of-frame sets, whereas **glutamate and aspartate** were consistently **higher** than in all alternatives. Although several residues showed only minor frame dependence, the overall pattern indicates that the native reference proteome occupies a **restricted region of composition space** relative to **that spanned by its genome-encoded reading-frame alternatives**.

Figures

Figure 1. Per-CDS residue-fraction distributions for the native reference proteome versus out-of-frame translations in *Escherichia coli* K-12



Legend. Each panel corresponds to one of the 20 genetically encoded amino acids. Within a panel, curves show the distribution of **per-CDS residue fractions** in the native reference proteome (**solid red**, +1 frame) overlaid with those for the five non-native frames (**distinct dashed lines**; +2, +3, −1, −2, −3). Non-native sets were generated by re-parsing the **unaltered** CDS nucleotide sequences in the specified strand/frame; no insertions or deletions were introduced. Curves are **normalized to unit area** within each panel to facilitate cross-frame comparison of distributional shape and location. **X-axis**: residue fraction per CDS (0–0.3). **Y-axis**: density (unit area). **Dataset**: 4,318 CDS (mean length **932.6 bp**, median **825 bp**). Consistent with prior work, the reference-proteome curves (solid red) are **bell-shaped** for all 20 residues [6,7]. Notable trends include **lower cysteine and arginine** and **higher glutamate and aspartate** in the native reference proteome relative to all out-of-frame sets.

Discussion

Recent work on de novo genes suggests that the emergence of coding sequences from previously noncoding DNA is not uncommon in evolution [1,2,3]. Yet the selective filters that allow such sequences—and variants arising from mutations of existing genes—to persist as genes remain debated [1,4,5]. Proposed criteria range from molecular properties (e.g., biosynthetic costs, toxicity, aggregation propensity) to systems-level requirements (e.g., functional integration and maturation) [4,5,11,12,13]. However, whether amino-acid composition per se constitutes a direct target of selection has not, to our knowledge, been systematically assessed.

In previous work, we analyzed reference proteomes from 81 species spanning the three domains of life and showed that, for each of the 20 genetically encoded amino acids, **per-protein residue fractions** form bell-shaped, species-specific distributions well approximated by binomial expectations [6,7]. To our knowledge, this reference-proteome–level regularity had not been previously documented. One natural implication is that proteomes are subject to **proteome-scale constraints on amino-acid composition**. An alternative, however, is that such patterns simply mirror features of the genetic code (e.g., unequal codon degeneracy) or reflect **non-random properties of genomes** (GC content, codon usage, k-mer structure) [10,14,15], without invoking selection on amino-acid composition per se. Discriminating among these possibilities requires asking whether the **composition space actually realized by a reference proteome** is restricted relative to the **composition space that the genome could in principle encode**.

A brute-force enumeration of all potential de novo open reading frames (ORFs) is computationally onerous and depends on operational choices. As a tractable proxy, we re-parsed each annotated CDS in the five **non-native reading frames** to generate **genome-encoded reading-frame alternatives**, without modifying the underlying nucleotide sequence. This construction preserves base composition and local sequence context while altering only the mapping from triplets to amino acids, thereby providing a conservative comparison set for the reference proteome's native (+1) translations.

In *Escherichia coli* K-12, the native reference proteome occupied a **narrow, displaced subset** of the composition landscape spanned by these frame alternatives. Thus, the native reference proteome is unlikely to be a typical sample from the genome-encoded alternatives. Instead, the results support a model in which coding sequences are preferentially **retained** when their amino-acid compositions **approximate a species-typical profile**—a macro-scale **"reference-proteome–conformance"** constraint that operates alongside, and upstream of, functional selection.

We next ask why protein and reference-proteome amino-acid compositions are constrained—**by what forces, and toward what profile**. If reference-proteome composition is constrained in this way, then the **cellular proteome**—the abundance-weighted set of actually expressed proteins—will be indirectly constrained as well, because it is drawn from, and limited by, the reference set. In previous work (preprint), we proposed a **mutual-constraint hypothesis**: the largest immediate resource for proteome synthesis is the cell's own **proteome-derived amino-acid pool** (recycling of degradation products) [6]. Consequently, (i) the **reference proteome** constrains the feasible

composition of the **cellular proteome**, while (ii) the **cellular proteome**, through turnover and recycling, constrains the effective amino-acid supply that feeds back on which coding sequences are sustainable over time. The **external compositional constraint** documented here—namely, that native translations lie within a restricted subset of genome-encoded composition space—provides an anchoring boundary condition for this mutual-constraint framework. Thus, we suspect that the observed constraint on amino-acid composition reflects a **mutual constraint** between the cellular proteome's **abundance-weighted** amino-acid composition and the reference proteome's **per-CDS residue-composition profile**.

An immediate next question concerns the strength of this constraint. Intuitively, stronger constraint should manifest as **sharper (less dispersed) per-residue distributions** around the species-specific mode. Consistent with our prior observation, the empirical distributions are well approximated by **binomial forms**. Although we do not quantify the strength here, the fact that a binomial distribution arises when sampling n residues at random from a population with fixed residue frequency p suggests a simple interpretation: the reference-proteome distributions may reflect sampling around a **target composition p** set by the **cellular proteome**—that is, by the abundance-weighted amino-acid profile of the recycled resource pool generated by proteome turnover. Under this reading, the reference proteome is constrained by (and in turn constrains) the cellular proteome's composition, consistent with the mutual-constraint hypothesis. This interpretation remains provisional; **future work** should explicitly account for gene-length heterogeneity and test for (over)dispersion relative to binomial expectations (e.g., via beta-binomial fits) across conditions and taxa.

Taken together with our earlier observation that reference proteomes across 81 species exhibit similar bell-shaped per-protein distributions [6,7], the present results suggest that **amino-acid composition is a general target of selection** that constrains gene retention across lineages. While code-level structure and genome-level supply biases undoubtedly contribute, the consistent restriction of the realized composition space relative to genome-encoded alternatives points to a **widespread, upstream compositional filter** shaping which sequences ultimately persist as genes.

Conclusion

By comparing **genome-encoded, unexpressed out-of-frame translations** with the native (+1) translations that constitute the **reference proteome**, we show that in *Escherichia coli* K-12 the reference-proteome composition occupies a **narrow subset** of the amino-acid composition space accessible from the genome. **Amino-acid composition** therefore represents a previously underappreciated **upstream requirement**, constraining the emergence and persistence of protein-coding genes and, by extension, the species' reference proteome.

Reference

1. Van Oss, S. B., & Carvunis, A.-R. (2019). De novo gene birth. PLOS Genetics, 15(5), e1008160. https://doi.org/10.1371/journal.pgen.1008160
2. Carvunis, A.-R., Rolland, T., Wapinski, I., Calderwood, M. A., Yildirim, M. A., Simonis, N., Charloteaux, B., Hidalgo, C. A., Barbette, J., Santhanam, B., Brar, G. A., Weissman, J. S., Regev, A., Thierry-Mieg, N., Cusick, M. E., & Vidal, M. (2012). Proto-genes and de novo gene birth. Nature, 487(7407), 370–374. https://doi.org/10.1038/nature11184
3. Zhao, L., Svetec, N., & Begun, D. J. (2024). De Novo Genes. Annual Review of Genetics, 58(1), 211–232. https://doi.org/10.1146/annurev-genet-111523-102413
4. Schmitz, J. F., & Bornberg-Bauer, E. (2017). Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. F1000Research, 6, 57. https://doi.org/10.12688/f1000research.10079.1
5. Iyengar, B. R., & Bornberg-Bauer, E. (2023). Neutral Models of De Novo Gene Emergence Suggest that Gene Evolution has a Preferred Trajectory. Molecular Biology and Evolution, 40(4). https://doi.org/10.1093/molbev/msad079
6. Esumi, G. (2023). The distributions of amino acid compositions of proteins in an organism's proteome uniformly approximate binomial distributions [Preprint]. Jxiv. https://doi.org/10.51094/jxiv.408
7. Esumi, G. (2025). Chicken Eggs Are a Practical and Common Exome-Matched Diet for Multicellular Eukaryotic Organisms [Preprint]. Jxiv. https://doi.org/10.51094/jxiv.1056
8. National Center for Biotechnology Information (NCBI). (2025). NCBI Datasets Taxonomy: Escherichia coli K-12 (TaxID 83333). National Library of Medicine (US). Retrieved September 29, 2025, from https://www.ncbi.nlm.nih.gov/datasets/taxonomy/83333/
9. Mir, K., Neuhaus, K., Scherer, S., Bossert, M., & Schober, S. (2012). Predicting Statistical Properties of Open Reading Frames in Bacterial Genomes. PLoS ONE, 7(9), e45103. https://doi.org/10.1371/journal.pone.0045103
10. Elzanowski, A., & Ostell, J. (2024, September 23). The Genetic Codes. National Center for Biotechnology Information (NCBI). https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi
11. Akashi, H., & Gojobori, T. (2002). Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proceedings of the National Academy of Sciences, 99(6), 3695–3700. https://doi.org/10.1073/pnas.062526999
12. Wilson, B. A., Foy, S. G., Neme, R., & Masel, J. (2017). Young genes are highly disordered as predicted by the preadaptation hypothesis of de novo gene birth. Nature Ecology & Evolution, 1(6), 0146. https://doi.org/10.1038/s41559-017-0146
13. Heames, B., Buchel, F., Aubel, M., Tretyachenko, V., Loginov, D., Novák, P., Lange, A., Bornberg-Bauer, E., & Hlouchová, K. (2023). Experimental characterization of de novo proteins and their unevolved random-sequence counterparts. Nature Ecology & Evolution, 7(4), 570–580. https://doi.org/10.1038/s41559-023-02010-2
14. Kariin, S., & Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. Trends in Genetics, 11(7), 283–290. https://doi.org/10.1016/S0168-9525(00)89076-9
15. Campbell, A., Mrázek, J., & Karlin, S. (1999). Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proceedings of the National Academy of Sciences, 96(16), 9184–9189. https://doi.org/10.1073/pnas.96.16.9184