Codon Bias and Amino Acid Symmetry in Simulated RNA Sequences

A Statistical Exploration of Translation under Nucleotide Constraints

Kilian BARREIRO

May 2025

Overview

We developed a simulation framework in R to generate synthetic RNA sequences across five nucleotide profiles (balanced, GC-rich, AU-rich; light/high variants). We simulated sequences between 100 and 1000 nt long and translated them using the standard genetic code. The main goal was to explore several interactions between variables such as codon redundancy, nucleotide (nt) biases, melting temperatures and amino acid (AA) distributions.

We found a strong positive correlation between codon redundancy and AA frequency (ρ up to 0.75, $p < 2.2 \times 10^{-16}$), particularly under balanced and GC-rich conditions. Under AU-rich profiles, this relationship weakens. Nucleotide bias also alters AA composition: GC-rich codons favour GC-rich AAs (e.g., Proline), while AU-rich conditions enrich AU-biased AAs (e.g., Isoleucine), with shifts up to $\Delta = +0.109$.

Regarding strand symmetry (i.e. similarity in amino-acid composition between original and complementary strands), although the nucleotide bias significantly affects mean compositional differences between strands ($p = 4.6 \times 10^{-5}$), these differences remain small and statistically non-significant at the amino acid level. Moreover, differences converge toward zero as sequence length increases, indicating that overall strand symmetry is preserved despite compositional biases. AA compositional symmetry was largely independent of the codon reading frame shift.

Finally, a linear model revealed that amino acid proportions are significantly influenced by both profiles and mean melting temperature, with GC-rich scenarios showing higher values at elevated temperatures, indicating a strong interaction between sequence composition and melting temperature. This might be in line with known thermodynamic stability of G/C base pairs.

Methods

RNA sequences were generated as a string of nitrogenous bases (A, U, C, G), which were split into codons and translated into 21 amino acids (using the standard genetic code), using a custom function, *genRnaSeq*, which allows control over sequence length, GC content (and so AU), and complementary sequence creation via several parameters (all functions parameters and details are available at `https://github.com/KilianBARREIRO/RNA-World`). Unless specified, every analysis has been performed with starting at frame 1 (i.e. first nucleotide, shift parameter set on 1)

We chose to consider five scenarios (cited as profiles below) to modify sequences nitrogenous bases composition, including one random with uniform distribution, and four with predefined GC probabilities: 0.8 (highly GC-favored), 0.6 (moderately GC-favored), 0.4 (moderately AU-favored), and 0.2 (highly AU-favored) (nb: custom probability profiles can be integrated to give more freedom to the user, see *custom_profile* parameter).

Codon Redundancy Drives Amino Acid Proportions

We generated RNA sequences with lengths ranging from 100 to 500 nucleotides, in steps of 50, using 100 replicates per length, for each profile. To evaluate the influence of codon redundancy on amino acid composition in simulated RNA sequences, we computed Spearman rank correlation coefficients between the number of synonymous codons per amino acid and their observed proportions for each replicate in the translated sequences.

To statistically assess whether codon redundancy consistently drives amino acid frequency, we applied a one-sided one-sample t-test to the distribution of correlation coefficients, testing whether the mean correlation was significantly greater than zero, indicating a positive association. Normality pre-requisites and details of test results are displayed in the Table **??** (Annexe). Additionally, 95% confidence intervals for the mean correlation were computed using the standard error and the Student's t-distribution.

Amino Acid Symmetry Between Original and Complementary Strands

Effect of Frame Shift and Nucleotide Bias on AA Symmetry (ANOVA)

To assess which way would be worth exploring, we evaluated the effect of both nucleotide bias profile (*profile* parameter) and reading frame shift (setting *shift* parameter to 1, 2 or 3) on the compositional symmetry between original and complementary RNA strands. For each combination, 100 replicates of sequences with length 500 nt were simulated and amino acid proportion differences were computed per replicate. The mean absolute difference per replicate was used as the response variable in a two-way ANOVA testing the main effects of *Profile*

and *Shift* as well as their interaction. Post-hoc tests and visualisation were used to confirm the stability of the symmetry across shifts.

Complementary Strand Symmetry

To check how amino acid distribution varies between the original and complementary strand, we generated RNA sequences ranging from 100 to 1000 nucleotide lengths (step of 50), using 100 replicates per length, under every nucleotide distribution profile. For each sequence, the original and complementary strands were translated into amino acids, and the difference in amino acid proportions was averaged per length and profile.

To assess whether these differences were statistically significant, paired two-sided t-tests were performed across replicates for each amino acid at each sequence length. Results have been displayed for a 500-nucleotide length in table **??**. To account for multiple comparisons, p-values were adjusted using the Benjamini-Hochberg procedure to control the false discovery rate.

Summary statistics were calculated per amino acid and sequence length, including the mean difference, standard deviation, and adjusted p-values. Additionally, 95% confidence intervals were derived using the standard error of the mean and the Student's t-distribution. Figure **??** (Annexe) illustrates the results for length 500 nt; a full summary across all profiles is provided in Table 1.

Intrinsic Codon Composition Explains AA Usage Biases

To investigate whether the observed amino acid proportions are linked to the nucleotide composition of their codons, we used a reference table (see github code) containing the average GC and AU content, calculated as the proportion of G, C, respectively A and U nucleotides throughout all the codons coding for an AA. for each amino acid based on codon usage. For each simulation scenario, we compared the observed amino acid proportions with their theoretical GC or AU content using scatter plots, faceted by bias profile.

To assess how nucleotide composition biases affect amino acid usage, we compared the mean amino acid proportions obtained under four biased nucleotide profiles (GC-favored and AU-favored, in both "light" and "high" variants) against those obtained under a balanced nucleotide profile. For each condition, amino acid proportions were averaged across 100 replicates of 500-nucleotide-long sequences. Using a reference table of codon GC content, we identified amino acids with extreme compositional shifts. The top three most enriched and most depleted amino acids (compared to the balanced profile) were extracted for each biased condition. These were summarised in a table, including their respective GC content, to highlight how codon composition correlates with over- or under-representation in biased contexts. A complementary scatterplot was produced to visualise the global association between amino acid GC content and their observed frequencies across bias profiles.

Melting temperature as a categorical factor acts differently along biased profiles

For each profile, sequences ranging in length from 100 to 500 nucleotides were generated with 100 replicates per condition. The data were merged with melting temperature information per amino acid The melting temperature (Tm) for each amino acid was computed per AA, averaged on the number of codons coding for this AA and based on the simplified formula hereafter :

$$Tm = 2 \times (A + U) + 4 \times (G + C)$$

To investigate the relationship between melting temperature and sequence proportion across different profiles, the mean melting temperature was discretized into categorical bins. This approach was chosen to account for potential nonlinearities in the effect of temperature and to facilitate interpretation of interactions with the categorical variable Profile. A linear model was then fitted with the proportion of the target sequence (Proportion) as the dependent variable, and *MeanTm* (as a factor), Profile, and their interaction as independent variables. The choice of a linear model was motivated by the continuous nature of the response variable (Proportion of AA) and the interest in estimating linear relationships and interactions between categorical factors. Model diagnostics were conducted to verify the assumptions of normality, homoscedasticity, and independence of residuals.

An analysis of variance (ANOVA) was performed to test the significance of main effects and interaction terms.

Results

Codon Redundancy Positively Correlates with Amino Acid Frequencies

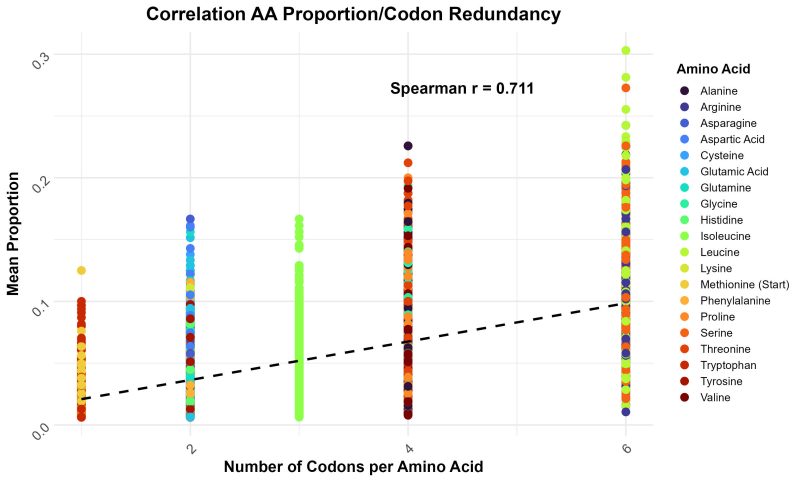

Figure 1: Correlation between amino acid proportion and codon redundancy for balanced profile.

Profile	Mean Spearman	95% CI lower	p-value
Balanced	0.71	0.70	$<$ 2.2e-16
GC-favored (light)	0.75	0.74	$<$ 2.2e-16
GC-favored (high)	0.67	0.66	$<$ 2.2e-16
AU-favored (light)	0.55	0.54	$<$ 2.2e-16
AU-favored (high)	0.16	0.15	$<$ 2.2e-16

Table 1: Statistical summary for every considered profile

The **Balanced** profile yielded a mean Spearman correlation of **0.71** (95% CI: 0.70–0.72), confirming a strong positive association between codon redundancy and amino acid frequency. Interestingly, a **light GC-favored** profile showed an even higher correlation (**0.75**, 95% CI: 0.74–0.76). This may reflect a stronger alignment between codon usage and amino acid frequency under GC-favoring conditions, possibly due to constrained codon sets.. In contrast, a **strong GC bias** (GC-favored high) led to a slightly lower mean correlation (**0.67**), likely due to reduced diversity in the codon pool. The **AU-favored light** profile showed a weaker correlation (**0.55**), and the **AU-favored high** profile further reduced it to **0.16**, reflecting the dominance of a smaller subset of AU-rich codons and the loss of codon-level variability (see Table 1).

Across all tested profiles, one-sided t-tests confirmed that the mean Spearman correlation was significantly greater than zero ($p < 2.2 \times 10^{-16}$ in all cases),

indicating that codon redundancy remains a key driver of amino acid frequency, even under nucleotide biases, though the nature of the bias modulates the magnitude of this effect.

Original VS Complementary strands differences in Amino Acid Proportions

Nucleotide Bias, Not Reading Frame, Drives Strand Asymmetry The two-way ANOVA (Table **??**) revealed a significant effect of nucleotide bias profile on amino acid compositional differences between original and complementary strands ($F(4, 1485) = 6.35$, $p = 4.6 \times 10^{-5}$), indicating that certain profiles increase these differences. Conversely, the shift of the reading frame had no significant impact ($F(2, 1485) = 0.19$, $p = 0.82$), nor was there a significant interaction between shift and profile ($F(8, 1485) = 0.62$, $p = 0.76$). Both Shift and Profile were treated as fixed factors in the model. This strongly suggests that strand compositional symmetry is robust to shifts in codon reading frame (see Figure **??**, Annexe).

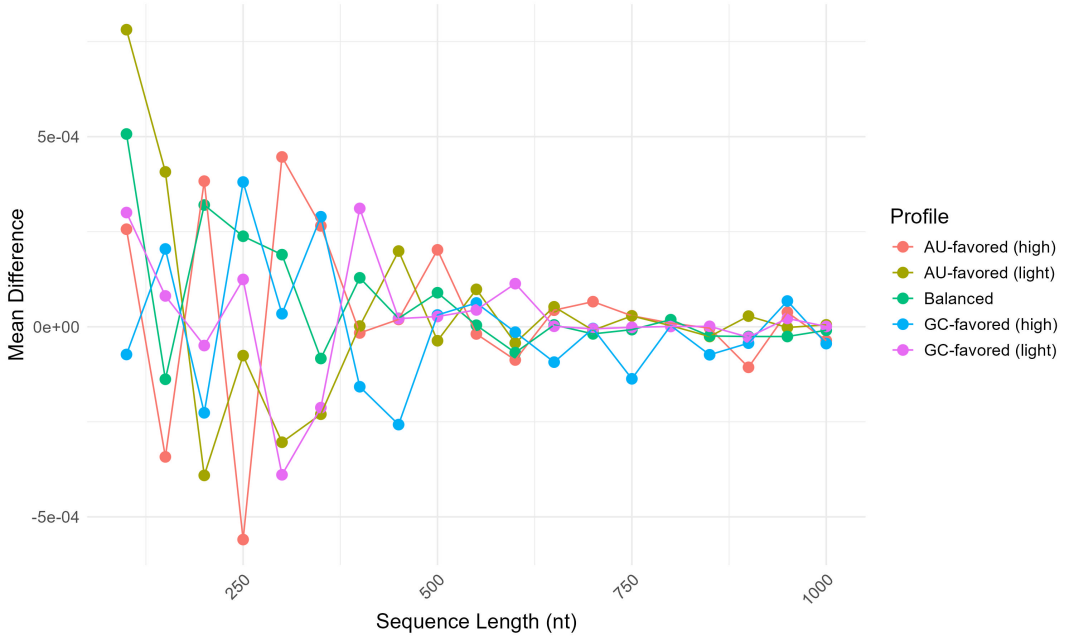

Figure 2: Mean difference between Original and Complementary amino acid proportions

Detailed Amino Acid-Level Analysis of Strand Differences Across all tested sequence lengths (100 to 1000 nt, step 50), the average differences in amino acid proportions between the original and complementary strands tend to get closer to zero (see Figure 2).

Among all profiles and for length 500, as shown in Table **??**, the maximum AA proportion difference was 0.008 (Leucine, AU-favored (High) profile), and no amino acid showed statistically significant proportion differences (all adjusted p-values > 0.05), indicating strong compositional symmetry between strands under uniform and biased nucleotide distribution. At the same length for the Balanced profile, Figure **??** (Annexe) illustrates the mean differences and associated confidence intervals for each amino acid. 85% of intervals included zero, and effect sizes were small, reinforcing the absence of directional bias in the codon-to-amino-acid mapping under these conditions.

These results suggest that these sequences, under our conditions, do not exhibit meaningful compositional divergence between original and complementary strands when no nucleotide bias is introduced. A complete table has been displayed in the Annexe.

Codon Composition Bias Explains Amino Acid Shifts Across Profiles

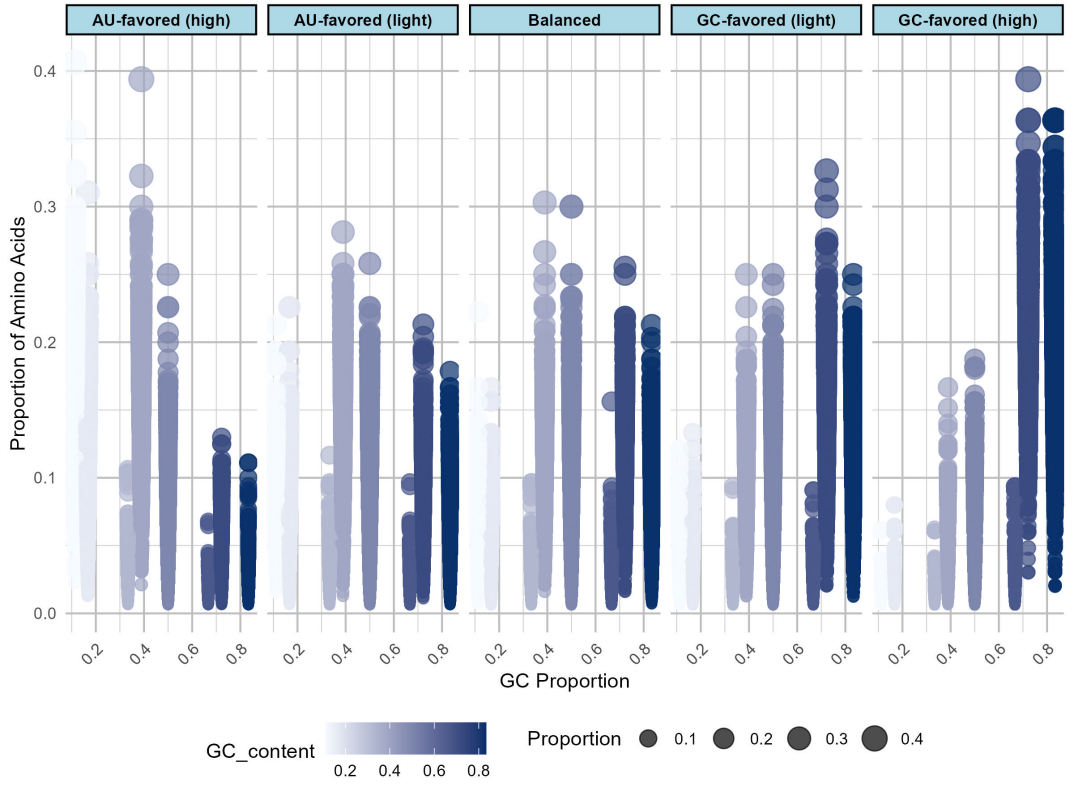

Figure 3: GC proportion vs AA proportion

The accompanying scatterplot (Figure 3) illustrates this global trend, showing that the proportional shift in amino acids aligns with their GC codon con-

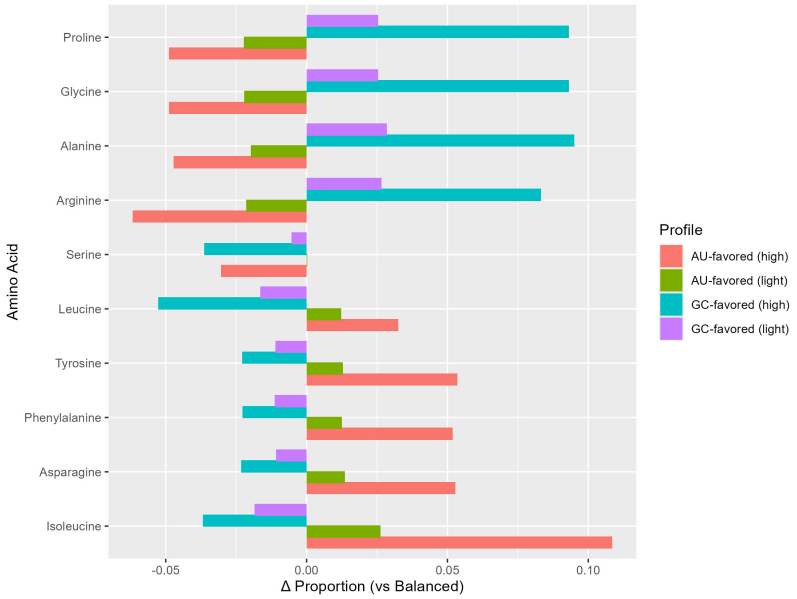

Figure 4: Most highly deviated amino acids from the balanced codon usage profile, illustrating shifts in amino acid frequencies under AT- and GC-biased scenarios.

tent across the tested profiles. This trend supports the hypothesis that nucleotide composition directly influences amino acid usage by favouring codons with matching base content.

As shown in figure 4 and detailed in Table **??**, GC-rich amino acids such as Alanine, Glycine, and Proline became substantially over-represented in GC-biased profiles, particularly under the "high" GC-favored condition (e.g., ΔProportion up to +0.095). Conversely, AU-rich amino acids such as Isoleucine, Asparagine, and Tyrosine dominated under AU-biased simulations, with the most extreme over-representation observed for Isoleucine in the high AU-favored condition (Δ Proportion = +0.109). Amino acids with intermediate GC content showed smaller or inconsistent deviations. These results confirm that codon GC/AU composition is a key determinant of amino acid usage in synthetic RNA sequences when nucleotide biases are introduced (For AU driven AA proportion see figure **??**.

This confirms that the codon composition strongly determines amino acid usage when the sequence composition is biased. supporting the model's accuracy by showing that GC/AU-balanced amino acids remain unaffected.

Melting Temperature

The linear model examining the effects of mean melting temperature factor (MeanTm), profile (Profile), and their interaction on the response variable *Proportion* was highly significant ($F(39, 155698) = 6625$, $p < 0.001$, see Table 2) and explained 62.4% of the variance (adjusted $R^2 = 0.6239$). The ANOVA

Source	Df	Sum Sq	Mean Sq	F value	Pr(>F)
MeanTm	7	76.858	10.980	14058.52	$< 2.2 \times 10^{-16}$
Profile	4	2.730	0.682	873.76	$< 2.2 \times 10^{-16}$
MeanTm:Profile	28	122.204	4.364	5588.29	$< 2.2 \times 10^{-16}$
Residuals	155698	121.600	0.0008		

Table 2: Analysis of variance for the linear model with melting temperature and profile interaction variables

further confirmed significant main effects of melting temperature (MeanTm), Profile, and their interaction (all render $p < 0.001$), indicating that the effects of temperature on the response varied depending on the profile.

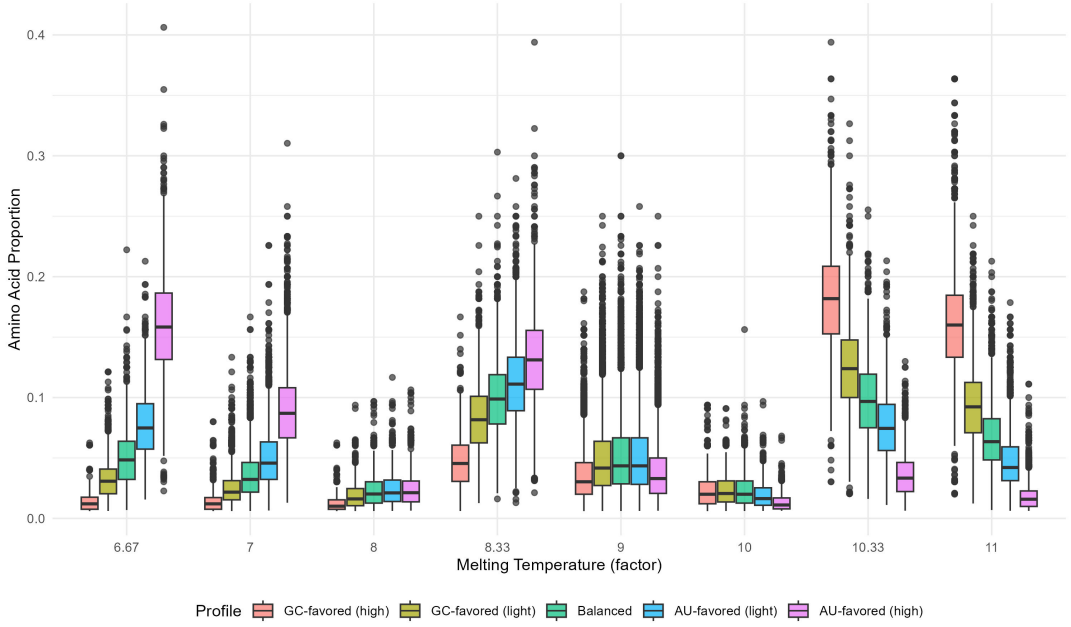

Figure 5: Distribution of AA proportions by Melting temperature and profile

Figure 5 displays boxplots of the response variable across temperature levels and profiles, visually supporting the statistical findings. Scenarios biased in favor of GC-rich codons show higher median amino acid proportions at elevated temperatures, whereas AU-favored or balanced profiles exhibit different or less pronounced patterns. These differences highlight the significant interaction between temperature and profile, suggesting that GC-enriched scenarios may be more responsive or adapted to higher thermal contexts.

It is worth noting that melting temperature is derived from codon usage metrics and thus partially overlaps with nucleotide composition, which may explain the strength of the observed associations. This reinforces the idea that temperature integrates effects already captured by compositional variables.