The dependence of the amino acid backbone conformation on the translated synonymous codon is not statistically significant

Javier González-Delgado, Pablo Mier, Pau Bernadó, Pierre Neuvial, Juan Cortés

The dependence of the amino acid backbone conformation on the translated synonymous codon is not statistically significant

Javier González-Delgado[a], **Pablo Mier**[b], **Pau Bernadó**[c], **Pierre Neuvial**[d], **and Juan Cortés**[e,*]

This manuscript was compiled on June 15, 2025

The correlation between synonymous codon usage and secondary structure in translated proteins has been widely demonstrated. This usage plays a capital role in tuning translational rates and protein folding kinetics, indirectly influencing multiple biological processes. A recent report (1) suggests that the translated synonymous codon influences the (ϕ, ψ) dihedral angles within secondary structure elements. If true, this conclusion would have strong consequences in several scientific fields, including structural biology and protein design, where results would depend on DNA sequence rather than protein sequence. Here, we show that the original statistical methodology used in the referred study was formally incorrect. Furthermore, when using a correct approach, we demonstrate that the influence of the codon on the distribution of the dihedral angles is not statistically significant for any type of secondary structure.

Synonymous codons | Protein structure | Angle distributions | Statistical tests

Significance Statement

Understanding how the genetic code influences protein structure is essential across biology, from gene expression to protein engineering. A recent study claimed that synonymous codons —different nucleotide triplets encoding the same amino acid— can affect the distribution of backbone dihedral angles (ϕ, ψ) within secondary structure elements, challenging the assumption that structure depends solely on amino acid sequence. We reanalyzed this claim using a statistically rigorous approach and found no significant codon-dependent effects on dihedral angle distributions across secondary structure types. Our results indicate that, based on available data, synonymous codon usage does not alter backbone geometry in folded proteins, reinforcing key assumptions in structural biology, protein design, and molecular evolution.

In a recent work, Rosenberg *et al.* (1) studied the dependence between the identity of synonymous codons and the distribution of the backbone dihedral angles of the translated amino acids. In the past, it has been shown that the use of synonymous codons is highly relevant in multiple biological processes including, among others, mRNA splicing, translational rates and protein folding (2, 3). While the correlation between synonymous codons and secondary structure in translated proteins has been widely studied (4–6), Rosenberg *et al.* evaluated the effect of codon identity on a finer scale, analyzing whether the distribution of (ϕ, ψ) dihedral angles within secondary structure elements is significantly altered when synonymous codons are used. Their conclusion, showing significant differences, particularly for amino acid residues involved in β-strands, would represent a new paradigm for the role played by synonymous codons in defining protein structure. In this work, we show that the statistical methodology used in (1) is formally incorrect. Besides, it is based on density estimates that might be imprecise for small sample sizes, yielding misleading comparisons. Using an appropriate methodology, we reanalyze the data presented in (1) and show that the influence of the codon on the distribution of the dihedral angles is not statistically significant for any of the secondary structures, contradicting the conclusion of (1). These results are corroborated by repeating the analysis on structures for the same set of proteins extracted from the AlphaFold Database (7), and shown to be robust with respect to the definition of secondary structural classes and also when considering the nature of the neighbor residues.

Results

Limitations of the original methodology. Keeping the notation of (1), if (c, c') denotes a pair of synonymous codons and $\mathcal{X}$ a type of secondary structure, Rosenberg *et al.* aimed at testing the null hypothesis $H_{0,(c,c')|\mathcal{X}}$ that both codon-specific distributions are the same. To do so, the authors introduced a metric to quantify differences between the distributions corresponding to different codons. Then, to assess the significance of such differences, they proposed to draw $B = 25$ pairs of bootstrapped samples, and to compare them with their synonymous codon counterparts using a permutation test procedure, with $K = 200$ permutations. For each bootstrap sample $b \in \{1, \dots, B\}$, if n_b denotes the number of permutations where the permuted metric is larger than the base metric (obtained from non-permuted data), they proposed the quantity

$$p_{(c,c'),\mathcal{X}} = \frac{1 + \sum_{b=1}^{B} n_b}{1 + BK} \qquad [1]$$

Author affiliations: [a]Université de Rennes, ENSAI, CNRS, CREST-UMR 9194, F-35000 Rennes, France; [b]Andalusian Centre for Developmental Biology (CABD, UPO-CSIC-JA), Faculty of Experimental Sciences (Genetics Area), University Pablo de Olavide, 41013 Seville, Spain; [c]Centre de Biologie Structurale, Université de Montpellier, INSERM and CNRS, F-34090 Montpellier, France; [d]Institut de Mathématiques de Toulouse, UMR5219, Université de Toulouse, CNRS, UPS, F-31062 Toulouse Cedex 9, France; [e]LAAS-CNRS, Université de Toulouse, CNRS, F-31400 Toulouse, France.

All the authors designed the studies, interpreted the results and wrote the manuscript; J.G. developed all the computational methods and performed the analyses; J.G. and P.N. carried out the theoretical analyses.

The authors declare no competing interests.

[*]To whom correspondence should be addressed. E-mail: juan.corteslaas.fr

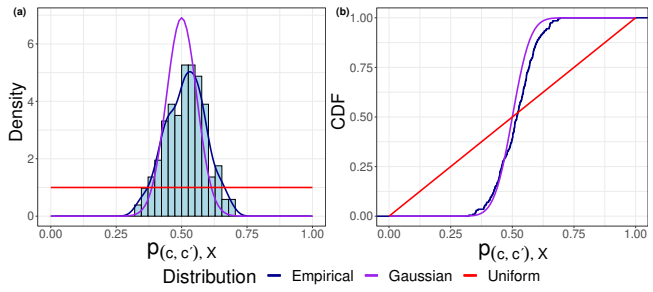

Fig. 1. Simulation of the null distribution of $p_{(c,c'),\mathcal{X}}$ for $K = 200$ and $B = 25$, chosen in (1). Left panel (a): histogram and kernel density estimate. Right panel (b): empirical Cumulative Distribution Function (CDF). Purple lines: asymptotic Gaussian distribution $\mathcal{N}(1/2, 1/\sqrt{12B})$; red lines: uniform distribution on $[0, 1]$.

as a p-value for $H_{0,(c,c')|\mathcal{X}}$. We can reformulate Eq. (1) in order to gain insight into its statistical behavior. First, let us define $p_b = (1 + n_b)/(1 + K)$, which is a well-defined p-value for the b-th permutation test. Letting $\overline{p}_B = B^{-1} \sum_{b=1}^{B} p_b$, it can be shown that

$$|p_{(c,c'),\mathcal{X}} - \overline{p}_B| \leq \frac{1}{K}\,. \qquad [2]$$

That is, for sufficiently large K, $p_{(c,c'),\mathcal{X}}$ is approximately the empirical mean of the B p-values associated to individual permutation tests. Details can be found in Section A of the Supplementary Information (SI).

However, $\overline{p}_B$ is not a valid p-value (see Section A in SI for a formal proof). Let us recall that a p-value is statistically valid if and only if its distribution under the null hypothesis is Super-Uniform, that is, if its cumulative distribution function (CDF) F is upper bounded by that of the Uniform distribution (denoted by U[0, 1] below): $F(x) \leq x$ for all x in $[0, 1]$. Moreover, the closer the p-value distribution under the null hypothesis is to U[0, 1], the more powerful the corresponding test is. Super-Uniformity is satisfied for classical permutation p-values such as p_b (with the CDF getting closer to U[0, 1] as K increases), but not for averages of p-values like $\overline{p}_B$. Instead, all the p_b could be correctly aggregated by taking their minimum and correcting the result for multiple testing (Bonferroni aggregation).

If the p_b were independent, then, by the Central Limit Theorem, the distribution of $\overline{p}_B$ would be asymptotically Gaussian $\mathcal{N}(1/2, 1/\sqrt{12B})$ as B tends to infinity. This distribution is not Super-Uniform, and therefore tests based on such a distribution are mathematically incorrect. In the setting of (1), the p_b are not independent since they have been computed by bootstrapping from one initial sample. However, for small values of B (including the choice of $B = 25$ in (1)), the null distribution of Eq. (1) deviates only slightly from the asymptotic independence setting. This is illustrated in Fig. 1, where the null distribution of Eq. (1) is simulated using the parameters chosen in (1). Details on the simulation and on the effect of K and B are included in Section B of the SI.

The empirical distribution of $p_{(c,c'),\mathcal{X}}$ presented in Fig. 1 is not Super-Uniform. Moreover, it is extremely conservative for large values of the statistic realization that is, low p-values, yielding an important number of false negatives and thus ignoring substantial differences appearing between the compared samples.

Finally, since the scores $p_{(c,c'),\mathcal{X}}$ are not valid p-values, they cannot be incorporated in a multiple testing procedure such as the Benjamini-Hochberg (BH) procedure (8) used in (1) for False Discovery Rate (FDR). Consequently, using and adjusting Eq. (1) for multiplicity will yield misleading analyses of the overall behaviour of all the null hypotheses and therefore, inaccurate results when the specificities of individual amino acids are studied *a posteriori*.

Beyond the above-mentioned methodological issues, the approach proposed in (1) presents several practical limitations. It needs, on the one hand, a substantial reduction of sample sizes, which may imply an important loss of information in some cases and thus a substantial power reduction. Indeed, the maximum sample size in (1) was set to $N_{\max} = 200$, whereas, for instance, the mean sample size for α-helical conformations was 724 and only 18% of the samples had sizes below $N_{\max}$. On the other hand, it requires a prior parametric estimation of the underlying densities, whose parameters would need to be optimized. In (1), the authors opted to fix the same bandwidth for all comparisons. However, too small bandwidths can lead to density undersmoothing, especially for small sample sizes. This would yield biased kernel estimates whose comparison might lead to false positives.

Correct goodness-of-fit tests show no significant differences between (ϕ, ψ) distributions. Using a well-suited statistical test (9), we tested all the pairwise differences between the (ϕ, ψ) distributions for all pairs of synonymous codons, for each amino acid. As in (1), we kept data points where codons were unambiguously assigned. Similarly, we repeated the same redundancy filtering based on averaging (ϕ, ψ) points with identical UniProt ID and sequence position. Note that an alternative filtering approach, which kept every first redundant point in the dataset, produced similar results. To facilitate the comparison with the results in (1), we kept only pairs of samples with sizes $n, m \geq 30$ and we classified all conformations according to their secondary structure according to DSSP (10): extended strand (E) and α-helix (H). We also performed the analysis for all the conformations not belonging to any of these two classes, which we named *Others*. As in (1), the BH multiplicity correction (8) was performed. When representing the Empirical Cumulative Distribution Function (ECDF) of the p-values, points laying above the line of slope $1/\alpha$ are considered rejections for a target FDR of α according to the BH procedure. The results are presented in Fig. 2(a).

The p-value distributions presented in Fig. 2(a) indicate that, for all the tested hypotheses, differences between codon-specific Ramachandran plots are not significant at level $\alpha = 0.05$. Indeed, the three depicted ECDF lay below the line of slope $1/\alpha$ and therefore no rejections are produced for a target FDR of α. Note that our results for H structures agree with those presented in (1), for which no significant discrepancies were retrieved. However, results for E structures strongly contradict those in (1), for which significant conformational differences were found for 66% of the synonymous codon pairs tested. Therefore, the main conclusion in (1) is firmly refuted when an appropriate statistical approach is used and, as a consequence, no significant effect of the translated codon identity on the amino-acid backbone conformation can be inferred from the current data.

The discrepancies between the two analyses are most probably due to the above-discussed incorrectness of the methods applied in the original study and, especially, from

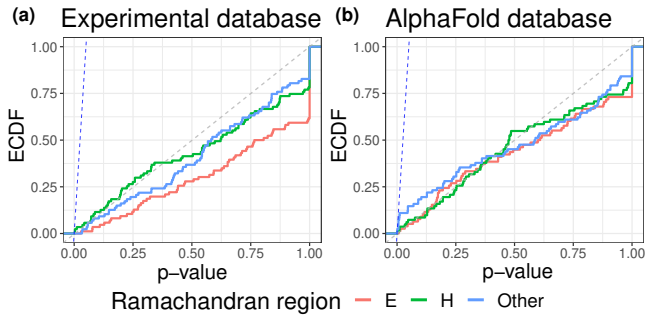

Fig. 2. Empirical cumulative distribution function (ECDF) of corrected p-values corresponding to testing the equality of (ϕ, ψ) distribution pairs of synonymous codons in the (a) experimental and (b) AlphaFold database, for conformations in extended strand (E, red), α-helix (H, green) and other (*Others*, blue) secondary structures. The dashed blue line of slope $1/\alpha$ corresponds to a target FDR set to $\alpha = 0.05$ for the BH correction. The dashed gray line represents the CDF of a Uniform distribution.

the potential use of biased density estimates to describe small (ϕ, ψ) samples. Indeed, when we looked at the codon pairs whose (ϕ, ψ) distributions were found significantly different in (1), we observed that their sample sizes concentrated around the smallest values in the dataset (see Section C in SI). This correspondence between significant differences and small sample sizes is counterintuitive for a well-defined statistical test. When the sample size is small and there is limited information about the underlying distribution, the null hypothesis is not rejected unless the evidence against it is very strong. Similarly, p-values closer to zero are often found for larger sample sizes, where the statistical power (the test's ability to detect differences) is higher. The opposite phenomenon found here suggests that false positives might be appearing due to misleading comparisons of small (ϕ, ψ) samples. As we discussed in the previous section, this may be caused by undersmoothed kernel density estimates computed with too small bandwidths for that setting.

To substantiate our conclusions, we performed additional analyses in alternative settings. First, we used structures extracted from the AlphaFold Database (7) for the same set of sequences, keeping residues with pLDDT values larger than 90. Results, shown in Fig. 2(b), qualitatively match those in Fig. 2(a) and therefore support the aforementioned conclusions. Our findings are equally consistent when using a less restrictive structural classification, based on non-overlapping regions on the Ramachandran space (see Section D in SI) (11, 12). Similarly, fixing the identities of neighboring amino acids to account for neighbor effects yielded the same conclusions (see Section E in SI).

Concluding remarks

The work of Rosenberg *et al.* introduced a new paradigm in biology: the nature of the codon influences the (ϕ, ψ) angles of protein secondary structures. While the correlation between synonymous codons and secondary structure in the translated proteins is a well known phenomenon (4–6), differences at the (ϕ, ψ) level for the most populated conformational states emerged as an intriguing and controversial observation (13). The conclusions reached from their work could have major impact on one of the paradigms of structural biology, which should shift from protein-sequence to DNA-sequence structure encoding, an information that is not currently stored in structural databases.

With the present study, we have demonstrated the incorrectness of the statistical methodology proposed in (1) to compare probability distributions. This, together with the use of density estimates that are not appropriately tuned for small sample sizes, makes the approach in (1) unsuitable to correctly compare codon-specific Ramachandran plots. When using our previously developed statistical tests (9) on the same database, no significant differences between the structures encoded by synonymous codons could be detected. Importantly, we demonstrated that this observation is robust with respect to the origin of the 3D structure, the definition of the structural classes and the type of the flanking residues. Therefore, the ensemble of our results unambiguously show that, based on available data, a significant influence of the codon usage in the distribution of backbone dihedral angles in proteins cannot be inferred.

It is worth mentioning, however, that our results have been derived from a limited set of *Escherichia coli* proteins for which the structure had been experimentally determined, and assuming that the gene used for the production of the protein was the same as in the original organism. We believe that a general understanding of the of codon-specific (ϕ, ψ) plots can only be achieved by using extensive structural databases, including the corresponding gene sequence, and applying robust statistical methods, such as the one presented here.

Materials and Methods

We implemented the first of the two-sample goodness-of-fit tests for probability distributions supported on the two-dimensional flat torus presented in (9). The approach is non-parametric and based on the 2-Wasserstein distance.

Software availability

The code to reproduce the analyses presented in Fig. 1 and 2, as well as the additional analyses described throughout the text, is available at https://github.com/gonzalez-delgado/synco.

1. AA Rosenberg, A Marx, AM Bronstein, Codon-specific ramachandran plots show amino acid backbone conformation depends on identity of the translated codon. *Nat. Commun.* **13** (2022).
2. F Pagani, M Raponi, FE Baralle, Synonymous mutations in cftr exon 12 affect splicing and are not neutral in evolution. *Proc. Natl. Acad. Sci.* **102**, 6368–6372 (2005).
3. F Buhr, et al., Synonymous codons direct cotranslational folding toward different protein conformations. *Mol. Cell* **61**, 341–351 (2016).
4. M Orešič, D Shalloway, Specific correlations between relative synonymous codon usage and protein secondary structure. *J. Mol. Biol.* **281**, 31–48 (1998).
5. R Saunders, CM Deane, Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Res.* **38**, 6719–6728 (2010).
6. S Pechmann, J Frydman, Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. & Mol. Biol.* **20**, 237–243 (2013).
7. M Varadi, et al., AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2021).
8. Y Benjamini, Y Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Royal Stat. Soc. Ser. B* **57**, 289–300 (1995).
9. J González-Delgado, A González-Sanz, J Cortés, P Neuvial, Two-sample goodness-of-fit tests on the flat torus based on Wasserstein distance and their relevance to structural biology. *Electron. J. Stat.* **17**, 1547 – 1586 (2023).
10. W Kabsch, C Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
11. V Ozenne, et al., Mapping the potential energy landscape of intrinsically disordered proteins at amino acid resolution. *J. Am. Chem. Soc.* **134**, 15138–15148 (2012) PMID: 22901047.
12. A Estaña, et al., Predicting secondary structure propensities in idps using simple statistics from three-residue fragments. *J. Mol. Biol.* **432**, 5447–5459 (2020).
13. OJ Akeju, AL Cope, Re-examining correlations between synonymous codon usage and protein bond angles in E. coli. *Genome Biol. Evol.* p. evae080 (2024).