Statistical tests to detect differences between codon-specific Ramachandran plots

Javier González-Delgado[1,2], Pablo Mier[3], Pau Bernadó[4], Pierre Neuvial[2] and Juan Cortés[1]

[1] *LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France.*

[2] *Institut de Mathématiques de Toulouse, Université de Toulouse, CNRS, Toulouse, France.*

[3] *Institute of Organismic and Molecular Evolution, Faculty of Biology, Johannes Gutenberg University Mainz, Germany.*

[4] *Centre de Biologie Structurale, Université de Montpellier, INSERM, CNRS, France.*

In their 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. 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, 5], 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. However, their statistical methodology is formally incorrect, casting doubt on the obtained results. The origin of the incorrectness is described in the following section. Then, using an appropriate methodology, we reanalyzed the data presented in [1]. Our results confirm the influence of the codon on the distribution of the dihedral angles, but differ from those of Rosenberg *et al.* in the strength of significance of the differences depending on the secondary structure type. Finally, we assessed whether these findings may be affected by the structural classification or the local sequence context. These additional analyses show that codon-specific effects have similar significance in different areas of Ramachandran space, although the effect may be stronger for a particular type of secondary structure, such as β-strands compared to α-helices. They also indicate that synonymous codon effects are stronger when considered in the context of the local sequence.

Incorrectness of the original methodology

The goal of Rosenberg *et al.* was to assess the effect of synonymous codons on the distribution of (ϕ, ψ) dihedral angles by comparing codon-specific Ramachandran plots. Keeping the notation of [1], if (c, c') denotes a pair of synonymous codons and $\mathcal{X}$ a type of secondary structure, they 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, Rosenberg *et al.* 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

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

as a p-value for $H_{0,(c,c')|\mathcal{X}}$. We can reformulate (1) in order to gain insight into its statistical behavior. First, let us define

$$p_b = \frac{1 + n_b}{1 + K}\,, \tag{2}$$

which is a well-defined p-value for the b-th permutation test. Letting

$$\overline{p}_B = \frac{1}{B}\sum_{b=1}^{B} p_b\,, \tag{3}$$

it can be shown (see SI) that

$$|p_{(c,c'),\mathcal{X}} - \overline{p}_B| \leq \frac{1}{K}\,. \tag{4}$$

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.

However, $\overline{p}_B$ is not a valid p-value (see SI for a formal proof). Let us recall that a p-value p is statistically valid if and only its distribution under the null hypothesis is Super-Uniform. A random variable is said to be Super-Uniform if its cumulative distribution function (CDF) F is upper bounded by that of the Uniform distribution (denoted by U$[0,1]$ below), that is:

$$F(x) \leq x \text{ for all } x \text{ in } [0,1] \tag{5}$$

(see e.g. [6, Section 3.3]). Moreover, the closer the p-value distribution under the null hypothesis is to U$[0,1]$, the more powerful the corresponding test is. Condition (5) is satisfied for classical permutation p-values such as p_b (with the CDF getting closer to the U$[0,1]$ distribution 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 (e.g. [7, Theorem 27.1]), the distribution of $\overline{p}_B$ would be asymptotically Gaussian $\mathcal{N}(1/2, 1/\sqrt{12B})$ as B tends to infinity. This distribution does not verify (5), and therefore tests based on such a distribution are mathematically invalid. 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 $B = 25$ in [1]), the null distribution of (1) deviates only slightly from the asymptotic independence setting. This is illustrated in Figure 1, where the null distribution of (1) is simulated using the parameters chosen in [1]. Details on the simulation and further analyses of the effect of K and B are included in the SI.

The empirical distribution of $p_{(c,c'),\mathcal{X}}$ presented in Figure 1 does not satisfy Condition (5). 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 [8]. In particular, the Benjamini-Hochberg procedure [9] used in [1] needs the p-values to be Super-Uniform under the null hypothesis to control the False Discovery Rate (FDR). Consequently, using and adjusting (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*.

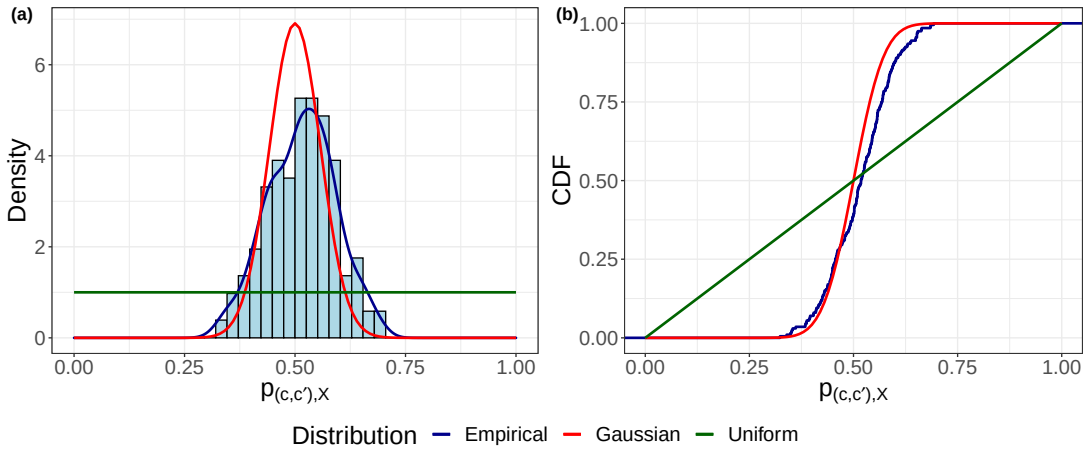

Figure 1: Simulation of the null distribution of $p_{(c,c'),\mathcal{X}}$ for $K = 200$ and $B = 25$, chosen in [1]. Left panel (a): histogram and kernel density estimate. Right panel (b): empirical Cumulative Distribution Function (CDF). Red lines: asymptotic Gaussian distribution $\mathcal{N}(1/2,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

Goodness-of-fit between codon-specific (ϕ, ψ) distributions

Beyond the above-mentioned methodological issues, the approach proposed in [1] presents several practical limitations. It needs, on the one hand, a prior parametric estimation of the underlying densities, whose parameters would need to be optimized. On the other hand, it requires 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] is set to $N_{\max} = 200$, whereas, for instance, the median sample size for α-helical conformations is 1414 and only 1.16% of the samples have sizes below $N_{\max}$.

In our recent work [10], we defined two two-sample goodness-of-fit tests for probability distributions supported on the two-dimensional flat torus, in order to study local changes on polypeptide backbone conformations. Both approaches are non-parametric and they use the information provided by entire datasets. The test statistic is based on the 2-Wasserstein distance, which integrates the geometry of the underlying space and provides strong theoretical guarantees and attractive empirical performance [11]. Here, we implemented the first of the testing procedures defined in [10], called N_g-geod, to detect differences between the codon-specific Ramachandran plots provided in [1]. For each amino acid, we tested all the pairwise differences between all the (ϕ, ψ) distributions of synonymous codons. To facilitate the comparison with the results in [1], we kept only pairs of samples with sizes $n, m \geq 30$ and we divided all conformations according to their secondary structure according to DSSP [12]: Extended strand (E) and α-helix (H). We also performed the analysis for all the conformations not belonging to any of these classes, which we named Others. The same multiplicity correction as in [9] was performed to the computed p-values. The results are presented in Figure 2.

The p-value distributions presented in Figure 2 indicate that significant differences between codon-specific Ramachandran plots are found for a substantial number of tested hypotheses: 78% for H, 87% for E and and 92% for Others. The results for α-helical structures strongly differ from those presented in [1], where no significant difference was retrieved (see Figure 4 in the original study). In addition, the proportion of significant differences for E is also considerably higher than in [1], where only 39% of the synonymous pairs were identified as structurally distinct. We believe that the observed discrepancies

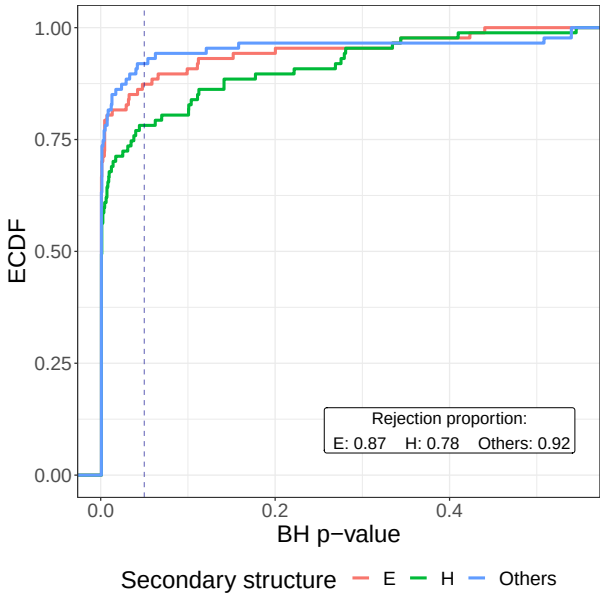

Figure 2: Empirical cumulative distribution function (ECDF) of corrected p-values corresponding to testing the equality of (ϕ, ψ) distribution pairs corresponding to different synonymous codons, for conformations in extended strand (E, red), α-helix (H, green) and other (Others, blue) secondary structures. The dashed blue line corresponds to a target FDR set to 0.05, determining the proportion of rejections among each set of tested hypotheses.

originate from the above-discussed methodological incorrectness of the methods applied in the original study, and in particular the substantial lack of power of the chosen statistic.

Results presented in Figure 2 clearly show how the effect of codon on the (ϕ, ψ) distribution is stronger for less rigid structural elements, as suggested in [1]. Indeed, we observe that the null hypothesis is more strongly rejected for extended strand structures than for α-helical ones, but even more strongly in regions that do not belong to any of these categories (blue curve in Figure 2). Outside H and E structures, (ϕ, ψ) angles are less constrained, making them potentially more sensitive to the translated codon. These differences in dihedral angle restrictions can be illustrated by measuring the dispersion of (ϕ, ψ) samples belonging to each secondary structure. We defined an estimator D measuring the concentration of one sample around its torus barycenter, which confirmed the previous statements. See, for instance, the average values for the three secondary structures: $\bar{D}_{\text{Others}} = 0.06 > \bar{D}_{\text{E}} = 0.01 > \bar{D}_{\text{H}} = 0.002$. Details and further analyses are provided in SI.

The results presented above, as those of Rosenberg *et al.*, were based on the structural classification provided by DSSP. We performed the same analyses using a less restrictive classification, only considering conformational regions on the Ramachandran space based on non-overlapping angular intervals and disregarding the formation of hydrogen bonds. The corresponding results (see SI) show that the differences on the rejection power between extended and helical conformations disappear in this case. This indicates that codon-specific effects have similar significance in different areas of Ramachandran space, although the effect may be stronger for a particular type of secondary structure, such as β-strands compared to α-helices.

Rosenberg *et al.* considered codon-specific Ramachandran plots corresponding to amino acids with arbitrary neighbors. However, the invalidity of Flory's Isolated Pair Hypothesis [13] and the interdependence of neighbor effects have been demonstrated in several experimental [14,15] and theoretical/computational

studies [16,17]. The consideration of neighboring residues is particularly relevant here, because the dataset in [1] exhibits important discrepancies in the proportion of left and right neighboring amino acid types among synonymous codons. After repeating the same analysis but fixing the identities of left and right neighbors, the overall conclusions were not substantially different. However, further statistical analyses showed that the codon effect on (ϕ, ψ) distribution is stronger when amino acid triplets are considered, as hypothesized in [1]. This implies that by ignoring the identities of the neighboring amino acids, the effect of codon on (ϕ, ψ) distributions is underestimated, and thus distorts subsequent analyses to understand the effect of codon changes for a particular amino acid. This is discussed more in detail in SI, where the corresponding analyses are presented.

Discussion and conclusions

Although quantitatively different, our results, based on an appropriate statistical methodology, confirm those presented in the original study by Rosenberg *et al.* [1], indicating that the nature of the codon has some influence on the fine details of the local conformation in proteins. While the correlation between synonymous codons and secondary structure in the translated proteins is a well known phenomenon, differences at the (ϕ, ψ) level for the most populated conformational states remain an intriguing and somehow counterintuitive observation. In fact, we cannot exclude artifacts related to the procedures. In particular, the nature of the dataset used could explain some of the subtle differences that we have observed. This dataset was derived from a limited set of *Escherichia coli* proteins for which the structure has been experimentally determined, and it was assumed that the gene used for the production of the protein was the same as in the original organism, which is a reasonable assumption in this case, but probably not in general. Moreover, high-resolution crystallographic structures are elucidated in a highly-packed context and at low temperature, severely reducing their inherent conformational fluctuations.

We believe that the detailed understanding of the codon effect on the fine structural features of proteins will only be achieved when extensive structural databases including the corresponding gene sequence are available. With the availability of extensive and accurate datasets, the comparative analysis of codon-specific Ramachandran plots at the amino acid and/or triplet level will be possible using the statistical methods presented here, thus enabling an unambiguous assessment on the influence of the gene sequence on polypeptide structure.

Software availability

The code to reproduce the analyses presented here is available at https://github.com/gonzalez-delgado/synco. The two testing procedures defined in [10] for assessing differences between (ϕ, ψ) distributions are implemented at https://github.com/gonzalez-delgado/wgof_torus.

References

[1] A. A. Rosenberg, A. Marx, and A. M. Bronstein, "Codon-specific ramachandran plots show amino acid backbone conformation depends on identity of the translated codon," *Nature Communications*, vol. 13, may 2022.

[2] F. Pagani, M. Raponi, and F. E. Baralle, "Synonymous mutations in cftr exon 12 affect splicing and are not neutral in evolution," *Proceedings of the National Academy of Sciences*, vol. 102, no. 18, pp. 6368–6372, 2005.

[3] F. Buhr, S. Jha, M. Thommen, J. Mittelstaet, F. Kutz, H. Schwalbe, M. V. Rodnina, and A. A. Komar, "Synonymous codons direct cotranslational folding toward different protein conformations," *Molecular Cell*, vol. 61, no. 3, pp. 341–351, 2016.

[4] M. Orešič and D. Shalloway, "Specific correlations between relative synonymous codon usage and protein secondary structure11edited by g. von heijne," *Journal of Molecular Biology*, vol. 281, no. 1, pp. 31–48, 1998.

[5] R. Saunders and C. M. Deane, "Synonymous codon usage influences the local protein structure observed," *Nucleic Acids Research*, vol. 38, pp. 6719–6728, 06 2010.

[6] E. L. Lehmann, J. P. Romano, and G. Casella, *Testing statistical hypotheses*, vol. 3. Springer, 2005.

[7] P. Billingsley, *Probability and measure*. John Wiley & Sons, 1995.

[8] E. Roquain, "Type i error rate control for testing many hypotheses: a survey with proofs," *Journal de la Société Française de Statistique*, vol. 152, no. 2, pp. 3–38, 2011.

[9] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[10] J. González-Delgado, A. González-Sanz, J. Cortés, and P. Neuvial, "Two-sample goodness-of-fit tests on the flat torus based on wasserstein distance and their relevance to structural biology," 2021. arXiv:2108.00165.

[11] G. Peyré and M. Cuturi, "Computational optimal transport: With applications to data science," *Foundations and Trends® in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[12] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[13] P. J. Flory and M. Volkenstein, "Statistical mechanics of chain molecules," *Biopolymers*, vol. 8, no. 5, pp. 699–700, 1969.

[14] K.-I. Oh, Y.-S. Jung, G.-S. Hwang, and M. Cho, "Conformational distributions of denatured and unstructured proteins are similar to those of 20 x 20 blocked dipeptides," *J. Biomol. NMR*, vol. 53, pp. 25–41, 2012.

[15] R. Schweitzer-Stenner and S. E. Toal, "Anticooperative nearest-neighbor interactions between residues in unfolded peptides and proteins," *Biophys. J.*, vol. 114, no. 5, pp. 1046–1057, 2018.

[16] D. Ting, G. Wang, M. Shapovalov, R. Mitra, M. I. Jordan, and R. Dunbrack, "Neighbor-dependent ramachandran probability distributions of amino acids developed from a hierarchical dirichlet process model," *PLoS Comput. Biol.*, vol. 6, no. 4, p. e1000763, 2010.

[17] J. González-Delgado, P. Bernadó, P. Neuvial, and J. Cortés, "Statistical proofs of the interdependence between nearest neighbor effects on polypeptide backbone conformations," *Journal of Structural Biology*, vol. 214, no. 4, p. 107907, 2022.

Author contributions

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.

Competing interests

The authors declare no competing interests.