Codon-specific Ramachandran plots show amino acid backbone conformation depends on identity of the translated codon.

Aviv Rosenberg  
Technion - Israel Institute of Technology  
https://orcid.org/0000-0002-3755-6534

Ailie Marx  
Technion - Israel Institute of Technology

Alex Bronstein (✉ bron@cs.technion.ac.il)  
Technion - Israel Institute of Technology

Article

Keywords:

Posted Date: December 2nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1089201/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Codon-specific Ramachandran plots show amino acid backbone conformation depends on identity of the translated codon.

Aviv A. Rosenberg¹#, Ailie Marx¹# and Alex M. Bronstein¹*

¹Computer Science, Technion – Israel Institute of Technology, Haifa, 3200003, Israel

# Equal contribution

*Corresponding author
Abstract

Synonymous codons translate into chemically identical amino acids. Once considered inconsequential to the formation of the protein product, there is now significant evidence to suggest that codon usage affects co-translational protein folding and the final structure of the expressed protein. Here we develop a method for computing and comparing codon-specific Ramachandran plots and demonstrate that the backbone dihedral angle distributions of some synonymous codons are distinguishable with statistical significance for some secondary structures. This shows that there exists a dependence between codon identity and backbone torsion of the translated amino acid. Although these findings cannot pinpoint the causal direction of this dependence, we discuss the vast biological implications should coding be shown to directly shape protein conformation and demonstrate the usefulness of this method as a tool for probing associations between codon usage and protein structure. Finally, we urge for the inclusion of exact genetic information into structural databases.

Introduction

One of the most critical cellular processes is the decoding of genetic information into functional proteins. Transfer RNA (tRNA) molecules recognize codons of the messenger RNA (mRNA) sequence as it passes through the ribosome and deliver specific amino acids sequentially for addition to the growing peptide chain. 61 codons map to 20 amino acids, meaning that most amino acids are encoded by more than one, synonymous, codon. Once considered a silent redundancy of the genetic code, synonymous coding is now known to be functionally important, subject to evolutionary selective pressure and clearly associated with disease. Changes in synonymous coding can alter mRNA splicing, mRNA folding and stability, and can affect translational speed and accuracy and the conformation of the translated protein.

Numerous studies have shown that changes in the rhythm of translation can alter the kinetics of co-translational folding and so, the global conformation of the final protein product. Translation rate is affected by synonymous codon usage which alters mRNA structure and tRNA abundance, the latter
coevolving with codon bias\textsuperscript{15,16,17,18,20}. This mechanism provides an indirect association between codon usage and global protein structure. Nevertheless, whether and how synonymous variants of a gene will alter the conformation of the final folded protein is still poorly predictable and additionally the literature is riddled with reports of single synonymous mutations causing measurable functional effects that are not well-explained by current mechanisms\textsuperscript{21,22,23,24}. Together this suggests that we are far from fully understanding the role of codon usage in orchestrating protein folding.

To the best of our knowledge, no studies have investigated whether the specific backbone torsion of an amino acid is associated with the synonymous codon from which it was translated. To probe for such a direct and local association, we developed a method for estimating and comparing codon-specific backbone dihedral angle distributions, which we term codon-specific Ramachandran plots. Comparing these distributions for pairs of synonymous codons, statistically significant differences are observed, indicating a dependence between the codon identity and the backbone dihedral angle of the amino acid into which that codon is translated.

\textbf{Results}

\textit{Data collection and codon assignment}

Investigating dependence between codon identity and the protein backbone structure is not readily achievable since, regrettably, there is no annotation within the Protein Data Bank (PDB) for the actual genetic template used in producing the protein for crystallization. Automatic assignment of codon identity to each position in a protein structure is a prerequisite to calculate codon-specific Ramachandran plots. It is imperative to stress that any method used for large scale codon reassignment will carry an inherent limitation of being contaminated with uncertainty and error. The main reason is that codon optimization is very commonly used to improve heterologous gene
expression, especially in structure determination which necessitates the production of large amounts of soluble protein. There is not one common approach to codon optimization, and the choice of method often depends on trial and error.

The procedure for computing and comparing codon-specific comparing backbone dihedral angle distributions is displayed in Fig.1 and detailed in the Methods. Briefly, high-resolution PDB structures are retrieved, structures are filtered to remove homology bias and the resulting proteins are grouped according to their unique Uniprot identifier. Genetic sequences were retrieved from ENA records cross-referenced by Uniprot, adopting a conservative approach: locations having more than one genetic variant for a specific residue are excluded from further analysis. For each group, a single protein record was generated with each point in the amino acid sequence annotated with the \( \psi, \phi \) backbone dihedral angles averaged over all the structures in the record, the codon, and DSSP secondary structure assignment. The final data set included 1343 protein chains. We estimated the codon distributions from their samples using kernel density estimation (KDE) on a torus with a Gaussian kernel width of 2°. We used a bootstrapped resampling scheme to estimate multiple realizations of these codon specific distributions. P-values were calculated via permutation test on the \( L_1 \) distance between the estimated densities (steps H-J); the rejection threshold \( p=0.019 \) was established by Benjamini-Hochberg multiple hypothesis correction with the false discovery rate set to \( q=0.05 \).

| Method | X-ray Expr. sys. | Source | Resolution |
|--------|-----------------|--------|------------|
| PDB query | E.coli | E.coli | <1.8Å |

Remove chimeras & redundant chains

Pairwise BLAST scores

Redundant proteins filter

Uniprot and ENA alignment

Figure 1 – Data Collection and Analysis. Querying the PDB for high resolution (<1.8Å), high quality (R\(_{\text{free}}\) < 24%) X-ray crystal structures of *E. coli* proteins expressed in *E.coli* (A), out of which unique chains were extracted (B). To ensure the protein set was non-redundant, pairwise sequence alignment scores were calculated between every pair of unique sequences (C). A farthest point sampling procedure was then employed to produce a sub-set of structures with normalized pairwise similarity not exceeding 0.7 (D). Structures were then grouped according to their unique Uniprot identifier. Genetic sequences were retrieved from ENA records cross-referenced by Uniprot (E), adopting a conservative approach: locations having more than one genetic variant for a specific residue are excluded from further analysis (F). For each group, a single protein record was generated with each point in the amino acid sequence annotated with the \( \phi, \psi \) backbone dihedral angles averaged over all the structures in the record, the codon, and DSSP secondary structure assignment (G). The final data set included 1343 protein chains. We estimated the codon distributions from their samples using kernel density estimation (KDE) on a torus with a Gaussian kernel width of 2°. We used a bootstrapped resampling scheme to estimate multiple realizations of these codon specific distributions. P-values were calculated via permutation test on the \( L_1 \) distance between the estimated densities (steps H-J); the rejection threshold \( p=0.019 \) was established by Benjamini-Hochberg multiple hypothesis correction with the false discovery rate set to \( q=0.05 \).
according to their unique Uniprot entry. For each position in a protein chain, the backbone dihedral angles, $\varphi$ and $\psi$, are calculated; if multiple PDB structures are available, the angles are averaged. Alongside this precise structural information, DSSP secondary structure is designated, and codons are assigned according to the genetic sequences obtained from ENA records cross-referenced in the Uniprot entry. Only locations with unambiguously assigned secondary structures and codons are retained.

We used only well-fitted X-ray crystal structures having a resolution higher than 1.8Å, as a recent study considering alternate backbone conformations found resolutions better than 2.0Å useful for such purposes\(^\text{28}\). To limit codon assignment errors from including codon optimized genes, we selected only structures of *E. coli* proteins expressed in *E. coli*, the most common expression system in the PDB. We purposely did not include all natively expressed proteins from other species, since codon biases differ
between organisms and such generalization could obfuscate the sought for associations between coding and structure.

Codon-specific backbone angle distributions are significantly distinct within the β-sheet mode.

Synonymous codons are known to have distinct propensities to different secondary structures, which is manifested as different probabilities of the corresponding modes in the full codon-specific Ramachandran plots (Fig. 2). The difference in propensity for the main two, α and β, secondary structure modes might therefore dominate the difference between the codon-specific Ramachandran plots of synonymous codons. To factor out this effect, we conditioned the dihedral angle distribution on the secondary structure, effectively restricting it either to the distinct β-sheet or α-helix modes. Select examples of the resulting codon-specific Ramachandran plots, conditioned by these modes, are shown in Fig. 3, while the full set is provided in Supplementary Figs. 1—2. Visually, it is evident that synonymous codons of some amino acids have clearly distinguishable distribution shapes especially in the β-mode.
To quantify those differences and their significance, we used a distribution-free two-sample permutation test with the $L_1$ distance between KDEs serving as the test statistic, and assigned p-values to each synonymous codon pair with respect to the null hypothesis that the two codons have the same underlying distribution of backbone angles. To determine the p-value threshold for statistical significance in a setting where multiple hypotheses are considered together, we employed the Benjamini-Hochberg correction with false discovery rate set to 0.05. This process is shown schematically in Fig. 1 and detailed in the Methods. Matrices visualizing the distances between select pairs of synonymous codon distributions are shown alongside the contour plots in Fig. 3 and for all synonymous codon groups in Supplementary Figs. 3—4.

Note that alongside the 87 synonymous pairs, we also included the 61 comparisons of each codon to itself. The latter served as a control, and indeed, the null hypotheses were not rejected for any of the same codon pairs in either of the secondary structures. No synonymous pairs were rejected in comparisons of the distributions of the $\alpha$-mode, however when comparing distributions for the $\beta$-mode, 57 of the 87 synonymous pairs were rejected.

It is not surprising that $\alpha$-helices, being less flexible than $\beta$-sheets\textsuperscript{34,35}, display less variability in codon-specific Ramachandran plots. The Ramachandran plot defines a richer range of structural contexts than the discrete categories available in DSSP annotation\textsuperscript{36}, especially in the $\beta$-mode. It is therefore possible that some of the differences we observe between codon-specific dihedral angle distributions in the $\beta$-mode are attributable to codon preferences for finer secondary structure categories such as parallel and antiparallel $\beta$-sheets. However, it should be noted that we used a strict conditioning by the secondary structure, taking only the DSSP annotation\textsuperscript{37} E (extended strand – $\beta$-sheet in parallel and/or anti-parallel sheet conformation with minimum length of 2 residues) for the $\beta$-mode, and H ($\alpha$-helix – a 4 turn helix with minimum length of 4 residues), for the $\alpha$-mode.
Distances between dihedral angle distributions of synonymous codons hint at a correlation to features of the translation process. Our findings remain silent regarding the origin of the observed differences in synonymous codon backbone dihedral angle distributions; in particular, the causation direction cannot be established unambiguously. It is tempting to speculate, however, that the translation process plays an active role in the observed effect. To illustrate this speculation, we considered how two features of the translation machinery correlate to the calculated distances between backbone dihedral angle distributions of synonymous codon pairs. Firstly, we demonstrate that the difference in the codon-specific translation speed between a pair of synonymous codons appears to positively correlate to the distance between their dihedral angle distributions (Fig. 4, left). Although ribosome profiling has facilitated measurement of translation speed to exquisite single-codon resolution in human and yeast cells, the application to bacteria has been more problematic. We used the data from Chevance et al.
who developed an in vivo bacterial genetic assay for measuring ribosomal speed independent of the stability of the mRNA transcript or the translated protein product. Note that in order to limit confounding factors, we considered only pairs of codons being translated by the same tRNA.

In a second illustration, we identified codons translated unambiguously by a single tRNA, following Bjork et al., and grouped codon pairs as being translated by either the same or different tRNA molecules. Fig. 4 (right) shows that synonymous codon pairs translated by different tRNAs tend to have a larger distance between their backbone dihedral angle distributions.

While these two trends can by no means be conclusive, they suggest the potential value of the proposed methods in analyzing relations between synonymous coding and the features of the translation process.

Discussion
Codon-specific Ramachandran plots and their comparative analysis could serve as a useful, quantitative tool in future studies looking at the association between coding and local protein structure. It is likely that codon-specific backbone dihedral angle distributions will show even more significant variations when extended to pairs or triplets. Codon pair usage bias has been observed in E. coli and in human disease. It has been suggested that codon translation efficiency is modulated by adjacent single nucleotides, that codon pair order significantly affects translation speed, and, more recently, the case for a genetic code formed by codons triplets has been argued.

The main challenge in extending this method to codon pairs or longer tuples is the relative scarcity of data and the need to compare multi-dimensional density functions characterizing the backbone structure of a tuple of amino acids. Extending the analysis to other expression systems faces a similar data scarcity challenge and considering genes from various source organisms expressed in E. coli, either to probe for evolutionary distinctions between species or to overcome data scarcity when probing codon pairs in a hypothesized translation dependent mechanism, is burdened by the
uncertainty associated with codon (re) assignment. The latter will be overcome when structural
databases start annotating the exact genetic source used for producing protein which is crucial, given
the ever-amounting evidence for the critical functional importance of codon usage

We hope that the associations between synonymous coding and local backbone conformation
revealed through codon-specific Ramachandran plots will spark new investigations which should
directly probe the possible causal relationships that might underpin these observations. The
implications of an active, coding-dependent process would be tremendous, necessitating an
immediate rethink as to how we manipulate the genetic code through codon optimization. This
question could not be timelier, as mRNA vaccines are taking centre stage in global medicine. It would
also affect how we define the role of synonymous variants in health and disease and could open a new
window into understanding protein folding in general. Regardless of the causal direction of the
observed dependence between coding and local structure can potentially improve protein prediction
algorithms since in such a task the causal relationship between the two is superfluous.

Methods

1.1 Data collection

Protein structure data is collected from the Protein Data Bank (PDB) through a structured query
against the search API. We queried for structures meeting the following criteria: (i) Method: X-Ray
Diffraction; (ii) X-Ray Resolution: Less than or equal to 1.8 Å; (iii) \( R_{\text{free}} \): Less than or equal to 0.24, (iv)
Expression system contains the phrase “Escherichia Coli” and (v) Source organism taxonomy ID equal
to 562 (Escherichia Coli).

Each query result contains a list of PDB IDs with an entity number, e.g., 1ABC:1 matching the query
criteria. Each entity within a PDB structure corresponds to one or more identical polypeptide chains
which exist in the structure. Note that a structure may have more than one unique entity (e.g., 1ABC:1
and 1ABC:2) in which case we would obtain both. For each unique entity ID, we select the first of its
matching chains using lexicographic order.
Next, we query the PDB’s entry data API to obtain a mapping from the specific chain to a list of Uniprot IDs. Whilst most chains map to a single Uniprot ID, there are cases where a chain is chimeric i.e. contains sections from multiple different proteins. We discard such chimeric chains and keep only chains which map to a unique Uniprot ID. We align the protein sequence of each chain to the Uniprot record sequence to provide a Uniprot index for each residue in the PDB chain. We used the same pairwise alignment algorithm as described in 1.3.

We then remove homology bias using the procedure described in 1.2. For all the remaining PDB chains, we calculate the backbone angles $(\phi, \psi)$ and use DSSP to assign a secondary structure per residue. Finally, we assign each residue with a codon using the method described in 1.3. The result of this process is what we call a Protein Record for each PDB chain. The Protein Record contains, per residue: corresponding Uniprot ID and residue index, torsion angles $(\phi, \psi)$, secondary structure and codon.

1.2 Redundancy filtering

To remove homology bias from our data, we performed a filtering step. First, each pair of Uniprot sequences are aligned using the BioPython software package, with a match score of 1 and all penalty scores set to zero. Thus, we obtain an alignment score $s_{ij} \geq 0$ between every pair of Uniprot sequences $i, j$. We then calculate normalized scores,

$$\bar{s}_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}s_{jj}}}$$

Note that by definition $0 \leq \bar{s}_{ij} \leq 1$ and $\bar{s}_{ii} = 1$ for every $i, j$. In other words, this normalization ensures that the self-alignment score is 1 and all other scores are normalized the be in $[0,1]$, regardless of the sequence lengths or the alignment penalty values. This normalization also makes it simple to choose a similarity cutoff threshold, since the threshold is chosen in the fixed range $[0,1]$ where 1 equates to an exact match and 0 to a complete mismatch. We chose a normalized similarity threshold of $\tau = 0.7$. 
Using the normalized alignment scores we then employ a farthest-first traversal procedure to sort the Uniprot sequences: the first sequence is selected arbitrarily, and each successively selected sequence is such that it has the lowest maximum normalized alignment score between itself and all previously-selected sequences. Formally, denote by $\mathcal{S}$ and $\mathcal{U}$ the sets of selected and un-selected sequences, respectively. We initialize to $\mathcal{S} = \{0\}$ and $\mathcal{U} = \{1, 2, \ldots, N - 1\}$ where $N$ is the number of Uniprot sequences. At each step $k$ of this traversal, for each unselected sequence $j \in \mathcal{U}$, we calculate its greatest similarity to any of the so-far selected sequences,

$$S_k(j) = \max_{i \in \mathcal{S}} s_i^j.$$

We then choose the sequence which has the lowest maximal similarity to the selected sequences, i.e., we add

$$j = \arg\min_{j' \in \mathcal{U}} S_k(j').$$

We stop the procedure once $S_k(j) > \tau$ for the sequence $j$ that was selected at step $k$, and retain $\mathcal{S}$ as the output filtered set of Uniprot sequences. This ensures that no two sequences in the selected set have a normalized similarity score greater than $\tau$. After performing this procedure, we keep in our dataset only PDB chains that were mapped to one of the Uniprot sequences in $\mathcal{S}$. Any PDB chain mapped to a Uniprot sequence from $\mathcal{U}$ is discarded from analysis. Note that we keep all chains from different PDB structures that correspond to the same selected Uniprot sequence in order to aggregate their backbone angles as explained in 1.4.

1.3 Codon assignment

Since the exact genetic sequence of the protein is not annotated in the PDB we assigned codons from the native sequence. Given a PDB chain, we obtain its unique Uniprot ID from the previous step. We query Uniprot to obtain all cross-referenced IDs to the European Nucleotide Archive (ENA). From the ENA database, we obtain all available genetic sequences for the specific protein and translate each genetic sequence to an amino-acid sequence using the standard genetic code table, and perform
pairwise sequence alignment between the PDB chain’s amino-acid sequence and the translated genetic sequences. The alignment is performed using the BioPython \textsuperscript{47} implementation of the Gotoh global alignment algorithm\textsuperscript{49}. We used BLOSUM80 as the substitution matrix for the alignment, a gap-opening penalty of $-10$ and a gap-extension penalty of $-0.5$.

Following the pairwise alignment of the amino acid sequence to all translated genetic sequences, we obtain the aligned codons from each sequence and assign them to corresponding residues from the PDB chain. This process yields zero or more assigned codons per residue in the PDB chain. In cases where there is more than one codon (i.e., different genetic sequences contributed different codons), we consider the assignment ambiguous and exclude that residue from further analysis.

1.4 Angle aggregation

Since some proteins have been characterized by multiple crystal structures, there residues from different PDB chains which to the same Uniprot ID and location in the Uniprot sequence. For example, in our dataset, the residues 1SEH:A:42, 1RNJ:A:42 and 2HRM:A:42 were all aligned to the Uniprot ID and index P06968:41. We consider such cases as different realizations of the same protein residue and aggregate the backbone angles from such residues, to obtain an “average” measurement of their backbone angles.

When aggregating the angles, we must account for the fact that a torsion angle pair $\varphi = (\varphi, \psi)$ is defined on a torus (i.e. the domain $S^1 \times S^1$ where $S^1$ is a circle). Intuitively, each angle naturally “wraps-around” at $\pm 180^\circ$, and the space spanned by two such angles is a torus. Thus, taking a simple average of each angle separately would not be correct. Instead we use the torus-mean function defined in \textbf{1.7.1}.

1.5 Backbone angle distribution distance

We seek to measure the distance between distributions of backbone torsion angles of synonymous codons in $\alpha$-helix and $\beta$-sheet secondary structure modes. Denote $f(\varphi|c, \mathcal{X})$ the distribution of backbone angles $\varphi$ of codon $c$ in secondary structure $\mathcal{X}$. We denote the distance between the
backbone angle distributions of two synonymous codons $c$ and $c'$ in secondary structure $X$ as $d(c, c')|X$ and estimate them between all pairs of synonymous codons. We include all cases of $c = c'$ as controls. The distance metric we used was the $L_1$ distance,

$$d_1(c, c')|X \triangleq \| f(\cdot|c, X) - f(\cdot|c', X) \|_1 = \int_{[-\pi,\pi]^2} |f(\Phi|c, X) - f(\Phi|c', X)| \, d\Phi.$$  

We also experimented with the $L_2$ and smoothed Wasserstein distances, however empirical tests showed that the $L_1$ metric provided the highest statistical power of all three at reasonable computational costs.

Although the underlying backbone angle distributions $f(\Phi|c, X)$ are unknown, we observe torsion angles sampled from them, since we can assign a codon and secondary structure per residue. This provides us a finite sample $\{\Phi_i \sim f(\cdot|c, X)\}_i$ for each codon $c$ and secondary structure $X$, from which we can estimate the unknown distribution. We use these samples to fit a kernel-density estimate (KDE), $\hat{f}(\Phi|c, X)$, of each distribution, as explained in 1.7.3. The distance metric $d_1(c, c')|X$ is then calculated on the KDEs of each synonymous codon pair. Since the KDEs are discrete, the integration above becomes a sum,

$$\hat{d}_1(c, c')|X = \sum_{k_1,k_2=1}^{K} |f(\Phi_{k_1,k_2}|c, X) - f(\Phi_{k_1,k_2}|c', X)|,$$

where $K$ is the number of KDE bins in each direction and $\Phi_{k_1,k_2} = (\varphi_{k_1}, \psi_{k_2})$ are discrete evenly-sampled grid points. We then use permutation-based hypothesis testing to determine whether the distance we obtained supports the (alternative) hypothesis that the codons have a significantly different distribution, as explained in 1.6.

1.5.1 Preventing bias due to secondary structure preference
A crucial point is that we calculate the distance between synonymous codon distributions separately for each secondary structure. This is done in order to avoid biasing the distance due to the fact that synonymous codons have different propensities for different secondary structures.

To see why this is a problem, consider that the the distribution of a codon’s backbone angles in both helix and sheet together can be written as a bi-modal mixture of the separate distributions:

\[ f(\varphi|c,\{X_\alpha, X_\beta\}) = \gamma_1 \cdot f(\varphi|c, X_\alpha) + \gamma_2 \cdot f(\varphi|c, X_\beta), \]

where we denote \(X_\alpha\) and \(X_\beta\) as \(\alpha\)-helix and \(\beta\)-sheet, respectively, and \(\gamma_1\) and \(\gamma_2\) are the the mixture coefficients representing the prevalence of the codon in each of the secondary structures.

Now assume there exist synonymous codons \(c, c'\) which have exactly the same backbone angle distributions in both helix and sheet, but each has distinct mixture coefficients. Two such codons will clearly manifest the same angles in helices and sheets, so we would like to measure a distance of zero between their backbone angle distributions (theoretically, in the limit of infinite samples). However, measuring the distance between the codons’ distributions on data from both secondary structures together will in-effect be greater than zero (even in the limit of infinite samples), because it would measure the difference between their propensities for the secondary structures, which will be captured as the height of each mode in their bi-modal distributions. Thus, by comparing codon distribution after conditioning on each secondary structure separately, we avoid biasing our distance measurements by this difference of propensity for secondary structures.

### 1.6 Detecting synonymous codons with different angle distributions

Faced with finite-sample estimations of codon backbone angle distributions, \(f(\varphi|c, X)\), we aim to determine whether there exist pairs of synonymous codons \((c, c')\) for which the underlying distributions, \(f(\varphi|c, X')\), are different. The challenge, of course, is that any difference we see in terms of the measured distance between estimated distributions, \(d_s(c, c')|X\), could be due to chance, arising from the the availability of only finite data. Thus, our approach is to determine whether the
distance we measured is large enough such that the probability of obtaining such a distance by chance from identical underlying distributions is extremely small.

For every pair of synonymous codons and secondary structure \((c, c') \mid X\) (where we allow \(c = c'\)), we define a null hypothesis, which states that they have identical underlying backbone angle distributions:

\[
H_{0(c,c')} \mid X : f(\varphi \mid c, X) = f(\varphi \mid c', X)
\]

We used permutation-based hypothesis testing\(^{50}\) to obtain valid p-values for each of these null hypotheses without the need to make assumptions about the backbone angle distributions \(f(\varphi \mid c, X)\) or the distribution of the distance metric \(d_1(c, c') \mid X\) under the null. The permutation testing procedure is detailed in \(1.7.4\). We thus obtain, per secondary structure, a total of 148 p-values: 61 for identical codons, \(c = c'\), and an additional 87 for non-identical but synonymous codons, \(c \neq c'\).

With an hypothesis test, one usually seeks to seeks to control the probability of rejecting the null hypothesis when it is true (type-I error) by choosing a pre-defined significance level such as \(\alpha = 0.05\) and rejecting when \(p < \alpha\). However, in this case we are in a multiple-hypothesis setting, where we consider multiple null-hypotheses simultaneously and seek to determine which of them can be rejected. It is therefore necessary to control the type-I error for the whole set of simultaneous hypothesis tests, not individually per test. Otherwise, for a fixed \(\alpha\), the chance of any type-I error increases with the number of hypotheses, so it is no longer controlled by \(\alpha\). To address the multiple-hypothesis setting, we used the Benjamini-Hochberg method\(^{51}\). Using this approach, a significance threshold is calculated dynamically from the set of all obtained p-values, in a way which controls the False-Discovery Rate (FDR) for the entire set of tests (instead of the type-I error of each individual test). The method allows us to specify the FDR-control parameter, \(q\), and ensures that that over repeated trials the expected value of the proportion between false discoveries (i.e. false rejections of the null hypotheses) and total discoveries (all rejections of null hypotheses) will be \(q\).
1.6.1 Preventing bias due to sample size differences

Synonymous codons are not equally prevalent; some are more abundantly found than others, and in some cases there is a substantial difference in their abundance. This translates into vastly different sample sizes we collect for the backbone angles of each codon. Some rare codons in our dataset had almost two orders of magnitude less data compared to their more abundant synonymous counterparts.

This creates a challenge for comparing distributions estimated from finite samples. Estimated distributions can seem very distinct when using vastly different sample sizes to estimate them, even when the samples come from the same underlying distribution. One way to account for this would be to cross-validate the KDE kernel bandwidth and choose an appropriate value for each sample size. This is challenging however, since we would need to separately cross-validate for all codons, some with very limited data.

Instead, we opted to use a single kernel bandwidth, but fix the sample size for each set of synonymous comparisons. For each amino acid $\mathcal{A}$, and per secondary structure $\mathcal{X}$, we choose a single sample size $N_{\mathcal{A},\mathcal{X}}$ that will be used to estimate the distributions of all its synonymous codons. This single sample size is chosen to be the minimum of the sample sizes from all of its codons. Due to computational constraints, we also set an upper limit $N_{\text{max}}$ to the sample size for each estimated distribution. Thus, the sample size for all codons of amino acid $\mathcal{A}$ was calculated as

$$N_{\mathcal{A},\mathcal{X}} = \min\{N_{\text{max}}, \min_{c \in \mathcal{A}} \{N_{c,\mathcal{X}}\}\},$$

where $N_{c,\mathcal{X}}$ is the sample size for codon $c$ in secondary structure $\mathcal{X}$.

Choosing the smallest sample size mitigates sample-size bias in the estimation of the codon backbone angle distributions. However, this approach causes potential loss of data: in a group of synonymous codons, having one very rare codon would mean that we also use a limited number of samples from the more abundant codons. In order to address the sample size bias on one hand, while exploiting all
available data on the other hand, we employed bootstrapped-sampling scheme on top of the distribution estimation and comparison. For each codon $c \in \mathcal{A}$, we estimate its distribution $B$ times from $N_{\mathcal{A},X}$ samples drawn with replacement from its collected data. This gives us access to at most $B \cdot N_{\mathcal{A},X}$ samples from each codon $c \in \mathcal{A}$, instead of only $N_{\mathcal{A},X}$. We then compare $B$ pairs of distributions for each synonymous codon pair $c, c' \in \mathcal{A}$ using the permutation test, and use the results of all permutations in all bootstrap iterations to calculate the p-value of $(c, c')$.

Statistical tests were performed with $B = 25$ bootstrap iterations with $K = 200$ permutations each for a total of 5000 permutations used for p-value calculation. We used $N_{\text{max}} = 200$ for all comparisons and set an FDR threshold of $q = 0.05$.

Full procedure

The procedure for comparing synonymous codon backbone angle distributions and then selecting which codon pairs have significantly different distributions is described below.

For each synonymous codon pair, $(c, c')$ and secondary structure $X$, we calculate a p-value with respect to the null hypothesis $H_{0,(c,c')|X}$, i.e. that they come from the same underlying distribution:

1. For $b \in \{1, \ldots, B\}$:
   a. Sample $N_{\mathcal{A},X}$ observations randomly from $c$ and from $c'$ (each with replacement).
   b. Denote the sampled observations from $c$ and $c'$ as $C$ and $C'$ respectively.
   c. Apply permutation test procedure (1.7.4) on $C$ and $C'$ for $K$ permutations. The test-statistic $T(X, Y)$ first computes the KDEs of $X$ and $Y$, then calculates the L1 distance between them.
   d. Denote by $\eta_b$ the number of times the base metric no greater than the permuted metric in the current permutation test.

2. Calculate the p-value with respect to $H_{0,(c,c')|X}$:
For each secondary structure $X$, we calculate the significance threshold based on the Benjamini-Hochberg method as follows:

1. Denote $\{p_{i,X}\}_{i=1}^{M}$ the set of $M = 148$ p-values obtained from all pairwise comparisons of synonymous codons in secondary structure $X$.

2. Sort the p-values and denote $p_{(i),X}$ the $i$-th sorted p-value.

3. Calculate the threshold p-value index for an FDR of $q$, which is the largest p-value smaller than the adaptive threshold of $q \cdot i/M$:

$$i_0 = \max \left\{ i : p_{(i),X} \leq q \cdot \frac{i}{M} \right\}.$$

4. Set the adaptive significance threshold: $\alpha_M = p_{(i_0),X}$.

5. Reject the $i$-th null-hypotheses if $p_{(i),X} < \alpha_M$.

Finally, the set of synonymous codon pairs corresponding to the rejected null hypotheses are deemed to have significantly different backbone angle distributions.

1.7 Mathematical Tools

1.7.1 Torus mean. Given a set of $N$ points on a torus $\{\varphi_i\}_{i=1}^{N}$ where $\varphi_i = (\varphi_i, \psi_i) \in S^1 \times S^1$, we would like to calculate the mean of these points, $\overline{\varphi}$ in a way which accounts for the wrap-around of each angle at $\pm 180^\circ$. We define function which approximates a centroid on a torus, by calculating the average angle with circular wrapping in each direction separately. We denote this function as $\overline{\varphi} = \text{torm}(\{\varphi_i\})$. For example, if $\varphi_1 = (170, 170)$ and $\varphi_2 = (-170, -130)$ then we expect $\text{torm}(\{\varphi_1, \varphi_2\}) = (180, -160)$. We define the function as follows
\[ \mathbf{\hat{p}} = \text{torm}(\{\mathbf{\varphi}_i\}_{i=1}^N) = (\hat{\varphi}, \hat{\psi}) \]

where \( \text{atan2}(y,x) \) is a signed version of \( \arctan(y/x) \) which uses the sign of both arguments to unambiguously recover the sign of the original angle \( \theta \) such that \( y = \sin \theta \) and \( x = \cos \theta \).

1.7.2 Torus distance. Given two points on the torus, \( \mathbf{\varphi}_1 = (\varphi_1, \psi_1) \) and \( \mathbf{\varphi}_2 = (\varphi_2, \psi_2) \), we would like to measure the distance, in angles between these points. Calculating a simple Euclidean distance doesn’t account for the fact that each angle wraps around at \( \pm 180^\circ \). For example, the Euclidean distance between \( \mathbf{\varphi}_1 = (170,0) \) and \( \mathbf{\varphi}_2 = (-160,0) \) would produce \( 330^\circ \), while it’s simple to see that when considering the wrap-around at \( 180^\circ \), the distance should be \( 30^\circ \). We therefore define the torus distance function as follows:

\[ tord(\mathbf{\varphi}_1, \mathbf{\varphi}_2) = \sqrt{\arccos^2 \cos(\varphi_1 - \varphi_2) + \arccos^2 \cos(\psi_1 - \psi_2)}. \]

1.7.3 Kernel density estimation. We used two-dimensional kernel density estimation (KDE) in order to estimate backbone angle distributions from finite samples. Given samples \( \{\mathbf{\varphi}_i\}_{i=1}^N \) from torsion angles of a codon \( c \) in secondary structure \( \mathcal{X} \), we calculate

\[ f(\mathbf{\varphi}|c, \mathcal{X}) = \frac{\gamma}{N} \sum_{i=1}^N K(tord(\mathbf{\varphi}, \mathbf{\varphi}_i)), \]

where \( \mathbf{\varphi} \) represents points on a discrete grid, \( K \) is a scalar kernel function, \( tord(\cdot,\cdot) \) is the torus wrap-around distance defined in 1.7.2, and \( \gamma \) is a constant factor which normalizes the KDE so that it sums to one. The KDE was evaluated on a discrete grid of size \( 128 \times 128 \), which corresponds to a bin width of \( 360/128 \approx 2.8^\circ \). By applying the kernel to the wrap-around distance, we correctly account for the distance on the torus between each sample and each grid point. We used a simple univariate Gaussian kernel, \( K(x) = \exp(-x^2/2\sigma^2) \), with a variance of \( \sigma = 2 \) (equivalent to the kernel bandwidth). We did not adjust the kernel bandwidth for each estimated distribution according to the number of data.
points. Instead, we used a fixed bandwidth for all KDEs, and made sure to always compare KDEs calculated from the same number of samples.

1.7.4 Permutation-based two-sample hypothesis test. Given two statistical samples, \( X = \{x_i\}_{i=1}^{N_X} \) and \( Y = \{y_i\}_{i=1}^{N_Y} \) containing \( N_X \) and \( N_Y \) observations respectively, we wish to test whether the observations in both samples were obtained from the same underlying data distribution. A powerful and well-known approach to do this, is by conducting a two-sample statistical hypothesis test, with the null hypothesis that \( X \) and \( Y \) are sampled from the same distribution, i.e., \( H_0: P_X(x) = P_Y(y) \).

Such a test allows one to determine whether there is sufficient evidence to reject the null hypothesis, while limiting the chance of a type-I error (false positive, or rejecting \( H_0 \) when it is true) to be at most \( 0 < \alpha \ll 1 \). Denote by \( T(X, Y) \in \mathbb{R} \) a test statistic of our choosing, which numerically summarizes the differences between \( X \) and \( Y \), such that the smaller the value of \( T(X, Y) \), the more \( X \) and \( Y \) are deemed similar. Further denote by \( \hat{t} = T(X, Y) \) the value of this test statistic when evaluated on the samples at hand. To perform the hypothesis test, a p-value is calculated, which is the probability of obtaining a result at least as large as \( \hat{t} \) under the assumption that \( H_0 \) is true: \( p = \Pr[T \geq \hat{t} | H_0] \). The null hypothesis \( H_0 \) is then rejected if \( p < \alpha \), thereby limiting the probability of type-I error to be \( \alpha \).

A point of difficulty with this approach is that calculating the p-value requires access to the distribution of the test statistic under the null. Specifically, one needs to know the CDF of the random variable \( T|H_0 \), which can typically only be known by making some assumptions about the data (e.g., that it follows a known distribution) and choosing a test-statistic for which the distribution of \( T|H_0 \) can be computed analytically under the assumptions about the data (such as the t-statistic).

For the application of comparing codon backbone angle distributions, it is crucial to avoid any assumptions about the data distributions from which the samples are obtained. Any such assumption would be unfounded, and may bias our test to the point of making it useless. Moreover, we require the ability to choose any test-statistic \( T \), because this allows us to compare multiple options and choose a highly-discriminative test-statistic for the type of data at hand.
In order to avoid making assumptions about the data or compromising on the choice of test-statistic, we employed a permutation-based two-sample hypothesis test, where the distribution of $T|H_0$ can be estimated for any choice of $T$ by randomly permuting the observations’ labels. The procedure can be described as follows:

1. Inputs: samples $X = \{x_i\}_{i=1}^{N_X}$, $Y = \{y_i\}_{i=1}^{N_Y}$, test-statistic $T(X, Y) \in \mathbb{R}$, number of permutations $K$.

2. Compute the base statistic value: $\hat{T} = T(X, Y)$.

3. Pool the observations: $Z = \{z_1, ..., z_{N_X+N_Y}\} = \{x_1, ..., x_{N_X}, y_1, ..., y_{N_Y}\}$.

4. Compute a random permutation $\pi$ of $\{1, ..., N_X+N_Y\}$, such that $\pi(i)$ is the $i$-th element of this permutation.

5. For $k \in \{1, ..., K\}$:
   a. Permute the pooled observations: $Z^\pi = \{z_{\pi(1)}, ..., z_{\pi(N_X+N_Y)}\}$.
   b. Split the permuted observations:
      \[
      X^\pi = \{z_{\pi(1)}, ..., z_{\pi(N_X)}\}, \\
      Y^\pi = \{z_{\pi(N_X+1)}, ..., z_{\pi(N_X+N_Y)}\}
      \]
   c. Compute the permuted statistic value: $\tilde{T}_k = T(X^\pi, Y^\pi)$.

6. Calculate $\eta = \sum_{k=1}^{K} I[\hat{T} \leq \tilde{T}_k]$, the number of times that the base statistical was no greater than the permuted statistic.

7. Calculate the p-value $p = \frac{1+\eta}{1+K}$.

8. Output: $p$ and $\eta$. 
The key observation behind this approach is that under the null, we can treat $X$ and $Y$ as labels which are randomly assigned to observations from the same data distribution. Therefore, by permuting the labels and calculating the permuted test-statistic, we are obtaining samples of $T|H_0$. If $H_0$ is indeed true, we expect that $\hat{\tau} \approx \tilde{\tau}_k$, thereby yielding $p \approx 0.5$ as $K \to \infty$. Conversely, if $H_0$ is false, we would expect that $\hat{\tau} > \tilde{\tau}_k$, and then $p \to 0$ as $K \to \infty$. In practice, the number of permutations $K$ is limited by computational constraints. Nevertheless, since the smallest p-value which can be obtained is $p_{\text{min}} = 1/(1 + K)$, we know an upper limit for the number of necessary permutations for a given significance level (in case of a single test).

**Figure 5 – Example of test statistic distribution in the permutation test.** Codon pairs are compared using the $L_1$ distance statistic between their dihedral angle KDEs in the β-sheet secondary structure mode. For each pair, depicted is the 1-cumulative distribution function (CDF) of the difference between the $L_1$ distance between the pair of KDEs and one between the pair of KDEs constructed with permuted labels. The intersection of 1-CDF with the vertical axis yields the p-value of the test. When comparing a codon to itself (I-ATT, I-ATT), the null hypothesis holds, and the difference is expected to be positive half of the times (p-value≈0.5). The indistinguishable pair I-ATT, I-ATC produces a high p-value, while the more clearly distinguishable pair I-ATT, I-ATA yield a very low p-value. Two non-synonymous codons (I-ATT, A-GCG) appear perfectly distinguishable. Distributions were calculated using 100 bootstrap samples with 200 permutations in each.
Figure 6 – p-values obtained comparing pairs of synonymous codons in the β-sheet and α-helix modes. The total set of hypothesis tests included the 87 synonymous codon pairs with the addition of 61 comparisons of the codon with itself for control. The rejection threshold corresponding to false discovery rate q=0.05 was established using the Benjamini-Hochberg procedure (red curve). The set of tests on which the null was rejected is marked in green. For the full list of rejected pairs, refer to Supplementary Figure 1.

Acknowledgments

The authors would like to thank Yaniv Romano for his helpful discussions on statistical methods.

Author Contribution

AM posed the original hypothesis; AR, AM and AB designed the studies, interpreted the results and wrote the manuscript; AR and AB developed all the computational methods and performed the analyses.

AR and AM contributed equally to this work.

Competing Interests Statement

All authors declare having no competing interests.
References

1. Chen R, Davydov E. V., Sirota M., Butte A. J. Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLoS One*; 5:e13574 (2010)

2. Sharma, Y., Miladi, M., Dukare, S., Boulay, K., Caudron-Herger, M. et al. A pancancer analysis of synonymous mutations. *Nature Communications*, 10:2569 (2019)

3. Walsh, I., Bowman, M., Soto Santarriaga, I., Rodriguez, A. and Clark, P. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proceedings of the National Academy of Sciences*. 117 (7) 3528-3534 (2020)

4. Komar, A. The Ying and Yang of Codon Usage. *Hum Mol Genet* ;25(R2):R77-R85 (2016)

5. Kimchi-Sarfaty, C., Mi Oh, J., Kim, I-W., Sauna, Z. E., Calcagno, A. M., et al. A “silent” polymorphism in the MDR1 gene changes substrate specificity. *Science* 315, 525–528 (2007)

6. Mueller, W. F., Larsen, L. S., Garibaldi, A., Hatfield, G. W. and Hertel, K. J. The Silent Sway of Splicing by Synonymous Substitutions. *The Journal of biological chemistry*, 290(46), 27700–27711 (2015)

7. Pagani, F., Raponi, M. and Baralle, F. E. Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc. Natl. Acad. Sci*. 102, 6368–6372 (2005)

8. Zhou, X., Zhou, W., Wang, C., Wang, L., Jin, Y., et al. A Comprehensive Analysis and Splicing Characterization of Naturally Occurring Synonymous Variants in the ATP7B Gene. *Frontiers in genetics* 11, 592611 (2021)

9. Purvis I.J., Bettany A.J., Santiago T.C., Coggins J.R., Duncan K., et al. The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. A hypothesis. *J. Mol. Biol.* 193:413–417 (1987)
10. Zhao, F., Yu, C. H., & Liu, Y. Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells. *Nucleic acids research*, 45(14), 8484–8492. (2017)

11. Akashi H. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 136(3):927–935 (1994)

12. Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134(2):341–352 (2008)

13. Liu, Y. A code within the genetic code: codon usage regulates co-translational protein folding. *Cell Commun Signal* 18:145 (2020)

14. Buhr F., Jha S., Thommen M., Mittelstaet J., Kutz F., et al. Synonymous codons direct cotranslational folding toward different protein conformations. *Mol. Cell*. 61:341–351 (2016)

15. Riba, A., Di Nanni, N., Mittal, N., Arhne, E., Schmidt, A., Zavolan, M. Protein synthesis rates and ribosome occupancies reveal determinants of translation elongation rates, *Proc. Natl. Acad. Sci*. 116 (30) 15023-15032 (2019)

16. Nackley, A. G., Shabalina, S.A., Tchivileva, I. E., Satterfield, K., Korchynskyi, O., et al. Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science* 314, 1930–1933 (2006)

17. Bartoszewski, R. A., Jablonsky, M., Bartoszewska, S., Stevenson, L., Dai, Q., Kappes, J., Collawn, J. F., and Bebok, Z. A synonymous single nucleotide polymorphism in ΔF508 CFTR alters the secondary structure of the mRNA and the expression of the mutant protein. *J. Biol. Chem*. 285, 28741–28748 (2010)

18. Bulmer, M. Coevolution of codon usage and transfer RNA abundance. *Nature* 325, 728–730 (1987)
19. Ikemura, T. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. *J Mol Biol* 151, 389–409 (1981)

20. Yulong, W., Silke, J. and Xia, X. An improved estimation of tRNA expression to better elucidate the coevolution between tRNA abundance and codon usage in bacteria. *Scientific Reports* 9, 3184 (2019)

21. Karakostis, K., Gnanasundram, S.V., López, I., Thermou, A., Wang, L., *et al.* A single synonymous mutation determines the phosphorylation and stability of the nascent protein. *Journal of Molecular Cell Biology*, 11(3), 187–199 (2019)

22. Rajeshbhai Patel, U., Sudhanshu, G. and Chatterji, D. Unraveling the Role of Silent Mutation in the ω-Subunit of Escherichia coli RNA Polymerase: Structure Transition Inhibits Transcription. *ACS Omega.*, 4(18): 17714–17725 (2019)

23. Simhadri, V.L., Hamasaki-Katagiri, N., Lin, B.C., Hunt, R., Jha, S., *et al.* Single synonymous mutation in factor IX alters protein properties and underlies haemophilia B. *J Med Genet.* 54(5):338-345 (2017)

24. Chevance F. and Hughes K. Case for the genetic code as a triplet of triplets. *Proc Natl Acad Sci USA* 114(18):4745-4750 (2017)

25. Angov, E., Hillier, C. J., Kincaid, R. L. and Lyon, J. A. Heterologous Protein Expression Is Enhanced by Harmonizing the Codon Usage Frequencies of the Target Gene with those of the Expression Host. *PLoS ONE* 3(5): e2189 (2008)

26. Fu, H., Liang, Y., Zhong, X., Pan, Z., Huang, L., *et al.* Codon optimization with deep learning to enhance protein expression. *Sci Rep* 10, 17617 (2020)

27. Ranaghan, M.J., Li, J.J., Laprise, D.M. and Garvie, C.W. Assessing optimal: inequalities in codon optimization algorithms. *BMC Biol.*;19(1):36. (2021)
28. Keedy D.A., Fraser J.S. and van den Bedem H. Exposing Hidden Alternative Backbone Conformations in X-ray Crystallography Using qFit. *PLoS Comput Biol* 11(10): e1004507 (2015)

29. Plotkin, J. B. and Kudla, G. Synonymous but not the same: the causes and consequences of codon bias. *Nature Rev. Genet.* 12, 32–42 (2011)

30. Adzhubei, A. A., Adzhubei, I. A., Krasheninnikov, I. A. and Neidle, S. Non-random usage of 'degenerate' codons is related to protein three-dimensional structure. *FEBS Lett.* 399 (1-2):78-82 (1996)

31. Gu, W., Zhou, T., Ma, J., Sun, X. and Lu, Z. The relationship between synonymous codon usage and protein structure in Escherichia coli and Homo sapiens. *Bio Systems.* 73(2):89-97 (2004)

32. Gupta, S. K., Majumdar, S., Bhattacharya, T.K. and Ghosh, T.C. Studies on the Relationships between the Synonymous Codon Usage and Protein Secondary Structural Units, *Biochemical and Biophysical Research Communications:* 269 (3): 692-696 (2000)

33. Saunders, R. and Deane, C. M. Synonymous codon usage influences the local protein structure observed, *Nucleic Acids Res.* 38(19): 6719–6728 (2010)

34. Emberly, E. G., Mukhopadhyay, R., Tang, C., Wingreen, N. S. Flexibility of β-sheets: Principal component analysis of database protein structures. *Proteins: Struct., Funct., Bioinf.* 55, 91–98 (2004)

35. Emberly, E. G., Mukhopadhyay, R., Wingreen, N. S., Tang, C. Flexibility of α-helices: Results of a statistical analysis of database protein structures. *J. Mol. Biol.* 327, 229–237 (2003)

36. Hollingsworth, S. A., and Karplus, P. A. A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins. *Biomolecular concepts,* 1(3-4), 271–283 (2010)

37. Kabsch W., Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* 22 (12): 2577–2637 (1983)

38. Mohammad F, Green R, Buskirk AR. A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. *Elife.* 8:e42591 (2019)
39. Chevance, F.F., Le Guyon, S., Hughes, K.T. The effects of codon context on in vivo translation speed. *PLoS Genet.* 10(6):e1004392 (2014)

40. Björk GR, Hagervall TG. Transfer RNA Modification: Presence, Synthesis, and Function. *EcoSal Plus*; 6(1) (2014)

41. Yarus, M. and Folley, L.S., Sense codons are found in specific contexts. *J. Mol. Biol.* 182; 529-540 (1985)

42. Alexaki, A., Kames, J., Holcomb, D.D., Athey, J., Santana-Quintero, L.V., Lam, P.V.N., Hamasaki-Katagiri, N., Osipova, E., Simonyan, V., Bar, H., Komar, A.A. and Kimchi-Sarfaty, C. Codon and Codon-Pair Usage Tables (CoCoPUTs): Facilitating Genetic Variation Analyses and Recombinant Gene Design. *J Mol Biol.* 431(13):2434-2441 (2019)

43. Diambra, A. Differential bicodon usage in lowly and highly abundant proteins. *PeerJ.*, 5 (2017)

44. Cutler RW, Chantawannakul P. Synonymous codon usage bias dependent on local nucleotide context in the class Deinococci. *J Mol Evol.* 67(3):301-14 (2008)

45. Sussman, J.L., Lin, D., Jiang, J., Manning, N.O., Prilusky, J., *et al.* Protein Data Bank (PDB): Database of Three-Dimensional Structural Information of Biological Macromolecules *Acta Crystallographica Section D: Biological Crystallography* 54 (6): 1078–84 (1998)

46. Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., *et al.* UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Research* 32 (suppl_1): D115–19 (2004)

47. Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., *et al.* Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics *Bioinformatics* 25 (11): 1422–23 (2009)

48. Leinonen, R., Akhtar, R., Birney, E., Bower, L., Cerdeno-Tárraga, A. *et al.* The European Nucleotide Archive *Nucleic Acids Research* 39 (suppl_1): D28–31 (2010)

49. Gotoh, O. 1990. Optimal Sequence Alignment Allowing for Long Gaps *Bulletin of Mathematical Biology* 52 (3): 359–73 (1990)
50. Chung, E. Y., and Romano, J. P. Exact and Asymptotically Robust Permutation Tests. *Annals of Statistics* 41 (2): 484–507 (2013)

51. Benjamini, Y. and Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1): 289–300 (1995)

52. Efron, B., and Tibshirani, R. J. *An Introduction to the Bootstrap*. CRC press (1994)

53. Simonoff, J. S. *Smoothing Methods in Statistics*. Springer Science & Business Media (2012)
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryFiguresreduced.pdf