Matters Arising: Re-examining the correlations between codon usage and dihedral bond angles using a population genetics model

Opetunde J. Akeju[1] and Alexander L. Cope[2,3,4,*]

[1]New Jersey City University, Jersey City, New Jersey, United States

[2]Department of Genetics, Rutgers University, Piscataway, New Jersey, United states

[3]Human Genetics Institute of New Jersey, Rutgers University, Piscataway, New Jersey, United States

[4]Robert Wood Johnson Medical School, Rutgers University, Piscataway, New Jersey, United States

[*]Corresponding Author: alexander.cope@rutgers.edu

December 3, 2022

Codon usage bias (CUB), or the non-uniform usage of synonymous codons, has been observed across all domains of life (1). CUB is driven by a combination of both non-adaptive (e.g. mutation biases) and adaptive (e.g. natural selection for translation efficiency/accuracy) evolutionary forces (2; 3; 4; 5; 6). Empirical work has shown that changes to synonymous codon usage can affect co-translational protein folding and various computational studies have sought to determine if there is a general connection between codon usage and protein structure (see (7) for a review). In a recent manuscript, (8) explored the relationship between synonymous codon usage and the dihedral bond angles that form a protein's backbone. Using a method they developed to compare dihedral bond angle distributions across synonymous codons, they detected statistically significant differences between the dihedral bond angle distributions of synonymous codons within the *E. coli* proteome. Although they rightly note that correlation does not imply causation, they hypothesize that differences in dihedral bond angle distributions between synonymous codons could be due to differences in elongation speed (see Figure 5 in (8)). Here, we present results using simulated data suggesting the findings of (8) may be a statistical artifact due to failure to control for a significant factor shaping a gene's synonymous CUB: gene expression.

It is well known that gene expression is the strongest correlate with CUB, with highly expressed genes biased towards codons that are translated more efficiently or accurately and lowly expressed genes biased towards codons favored by mutational biases (5; 6). Analyses attempting to correlate codon usage with other genomic, transcriptomic, or proteomic factors must account for the variation in codon usage due to variation in gene expression. Previous work showed that failure to control for gene expression and other biases known to impact codon usage (e.g. amino acid biases) led to spurious differences in the codon usage biases of signal peptides and protein secondary structure (9; 10). (8) did not control for gene expression in their permutation test for comparing dihedral bond angle distributions (i.e. codons were allowed to be permuted across genes), potentially leading to spurious significant differences between the dihedral bond angles of synonymous codons. We examined if the correlations observed by (8) could be a statistical artifact.

To answer this question, we used a population genetics model of coding sequence evolution, called the Ribosome Overhead Costs version of the Stochastic Evolutionary Model of Protein Production Rate (ROC-SEMPPR) (11). ROC-SEMPPR assumes codon usage is at selection-mutation-drift equilibrium and is able to estimate parameters reflecting natural selection on synonymous codon usage and mutation biases by accounting for variation in gene expression across protein-coding sequences. Importantly, ROC-SEMPPR only considers how codon counts vary across protein-coding sequences. ROC-SEMPPR does not take into account the location of a codon in a protein-coding sequence, making it blind to information relevant to position-specific effects, such as protein bond angles.

The *E. coli* K12 MG1655 protein-coding sequences were downloaded from NCBI-RefSeq (GCF_000005845.2). (8) did not restrict their analysis to a single strain of *E. coli.* Non-K12 MG1655 protein-coding sequences used by (8) were downloaded from the European Nucleotide Archive and appended to the *E. coli* K12 MG1655 protein-coding sequence FASTA file. We note that one protein-coding sequence (ENA AAL21040.1) used by (8) was annotated in the ENA as a *S. enterica* gene, but this was included for completeness. Excluding positions with missing codons in the real data, the simulated data contained 99% of the amino acid sites included in the real data. ROC-SEMPPR was fit to these protein-coding sequences using the AnaCoDa R package (12) to estimate parameters reflecting natural selection on codon usage $\Delta\eta$, mutation bias ΔM, and an evolutionary average estimate gene expression ϕ. We note that ROC-SEMPPR estimates of gene expression were well-correlated with estimates taken from Ribo-seq data (Figure 1A, Spearman $R = 0.48$), suggesting an overall good model fit. Parameters estimated by ROC-SEMPPR were then used to simulate the codon usage of *E. coli* protein-coding sequences. Briefly, for each occurrence of an amino acid in a protein-coding sequence g, a synonymous codon was randomly sampled according to a multinomial distribution with the probability $p_{i,g}$ of observing synonymous codon i (of N synonymous codons) determined by the expression level of the gene ϕ_g and the natural selection $\Delta\eta$ and mutation bias ΔM parameters estimated

by ROC-SEMPPR for the relevant set of synonymous codons (see Equation 1 and see (11) for details).

$$p_{i,g} = \frac{e^{-\Delta M_i - \Delta\eta_i \phi_g}}{\sum_{j=1}^{N} e^{-\Delta M_j - \Delta\eta_j \phi_g}} \tag{1}$$

Note that the amino acid sequences remained the same between the real and simulated data using this procedure to simulate codon usage. As expected, the real and simulated protein-coding sequences had similar codon counts per gene (Figure 1B, Spearman $R = 0.83$).

The computational pipeline and the publicly-available data used to recreate their analysis were downloaded from the sources specified in (8). The pipeline was applied using the default settings to the real data reflecting the true codon usage patterns and the simulated data. To analyze the simulated data with the (8) pipeline, the protein-coding sequences represented in the real data were replaced with their simulated counterparts, retaining all other information as is from the real data, including the bond angles. Although codon usage within the simulated protein-coding sequences is similar to the real sequences, the positions of these synonymous codons within the sequences are effectively randomized compared to the real sequences (Figure 1C). For example, for the 4-codon amino acid valine (V), the percentage of times the same codon occurred at the same position within a protein-coding sequence between the real and simulated data is approximately 25%. For many amino acids, there tends to be a small upward bias relative to the uniform expectation (Figure 1C, dashed lines). This is likely due to the fact that some protein-coding sequences will be strongly biased towards certain synonymous codons and that some codons are rarely-used, such as the isoleucine codon ATA which accounts for only 6% of all isoleucines in the real data. Given the overall lack of agreement between the real and simulated data on a position-specific basis, the simulated data should contain little, if any, correlation between synonymous codon usage and protein bond angles, provided there is a general relationship between the two. Thus, analysis of our simulated data using the (8) pipeline should show little agreement with the analysis of the real data.

Overall, we found striking agreement between the analysis of the real and simulated data. 90.5% of the cases determined to be statistically significant in the real data were also found to be significant in the simulated data, despite the overall randomization of codon usage by position. In agreement with this, the p-values between the two analyses are highly correlated (Figure 1D, Spearman $R = 0.83$). Additionally, the distribution distance (DDist, to use the column name from pipeline output) statistics between the two analyses are also highly correlated (Figure 1E, Spearman $R = 0.98$). In cases where a synonymous codon pair was only significant in one of the datasets, the same synonymous pair tended to have a lower, but not significant, p-value in the other dataset (Figure 1F). Given that the simulated data accurately reflects

gene-specific codon usage patterns, but not position-specific codon usage patterns, it is likely the significant results obtained by (8) reflect a statistical artifact due to gene-specific biases. In the context of our results, the correlation between their distance metric and differences in codon-specific elongation speed (Figure 5 in (8)) suggests they are detecting signals related to selection for translation efficiency, as their analysis did not account for gene expression.

Much like (8), we cannot definitively say if synonymous codon usage determined dihedral bond angles or vice versa. However, our results strongly suggests that the correlations between codon usage and dihedral bond angles are a statistical artifact caused by the correlation between codon usage and gene expression. Importantly, highly expressed genes are also known to be subject to various other biases that might impact dihedral bond angles, including amino acid biases (13). Our findings further emphasize the importance of controlling for gene expression when testing the relationship of biological factors with codon usage.

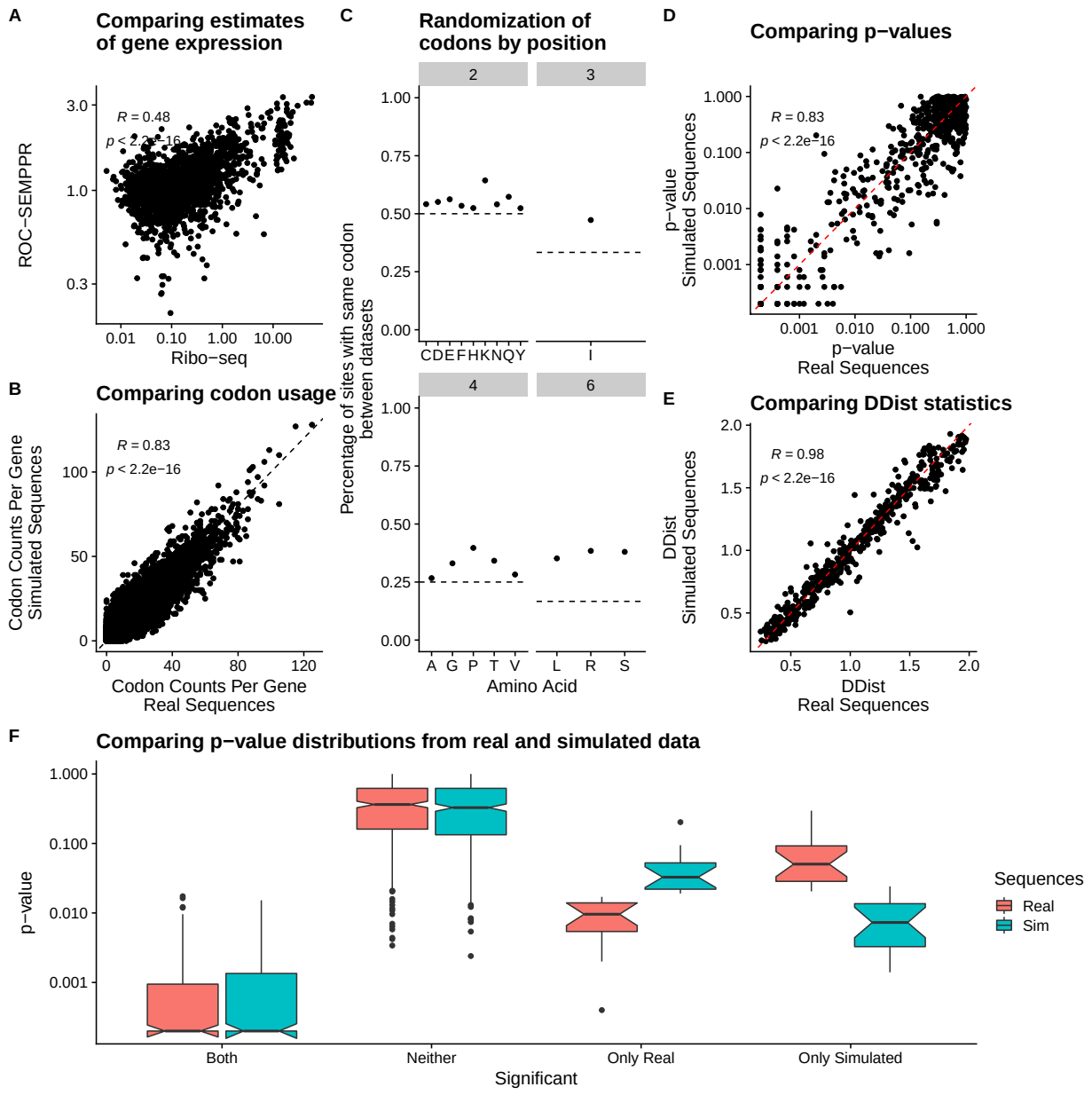

Figure 1: (A) Comparing estimates of ROC-SEMPPR predicted gene expression to Ribo-seq estimates taken from (14). (B) Comparing codon counts across all protein-coding sequences in the real and simulated data. Each dot represents the number of occurrences of codon within a given gene (C) Percentage of amino acid sites that have the same codon at the same position in the same protein-coding sequence between the real and simulated data. The dashed line indicates the expectation under a completely random distribution. (D) Comparison of p-values estimated from the Rosenberg *et al.* pipeline applied to the real and simulated data. (E) Comparison of distribution distance statistics estimated from the Rosenberg *et al.* pipeline applied to the real and simulated data. (F) Comparing p-value distributions for synonymous codon pairs that were significant in both, neither, or only one of the real and simulated data analyses.

Data availability

Relevant data and scripts can be found at https://github.com/acope3/Codon_usage_prot_structure_angles.

Competing interests

The author's declare no competing interest.

Contributions

O.J.A. contributed to the analysis, and the writing and editing of the manuscript. A.L.C. conceptualized the project, contributed to the analysis, and helped write and edit the manuscript.

References

[1] Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nature Reviews Genetics. 2011;12:32-42.

[2] 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. Journal of Molecular Biology. 1981 9;151:389-409.

[3] Gouy M, Gautier C. Codon usage in bacteria: Correlation with gene expressivity. Nucleic Acids Research. 1982 11;10:7055-74. Available from: https://academic.oup.com/nar/article/10/22/7055/1027194.

[4] Bulmer M. The selection-mutation-drift theory of synonymous codon usage. Genetics. 1991;129.

[5] Drummond DA, Wilke CO. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. Cell. 2008;134:341-52.

[6] Shah P, Gilchrist M. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. PNAS. 2011;108:10231-6.

[7] Chaney JL, Clark PL. Roles for Synonymous Codon Usage in Protein Biogenesis. Annu Rev Biophysics. 2015;44:143-66.

[8] Rosenberg AA, Marx A, Bronstein AM. Codon-specific Ramachandran plots show amino acid backbone conformation depends on identity of the translated codon. Nature Communications 2022 13:1. 2022 5;13:1-11. Available from: https://www.nature.com/articles/s41467-022-30390-9.

[9] Cope AL, Hettich RL, Gilchrist MA. Quantifying codon usage in signal peptides: Gene expression and amino acid usage explain apparent selection for inefficient codons. Biochimica et Biophysica Acta - Biomembranes. 2018;1860.

[10] Cope AL, Gilchrist MA. Quantifying shifts in natural selection on codon usage between protein regions: a population genetics approach. BMC Genomics 2022 23:1. 2022 5;23:1-19. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-022-08635-0 http://creativecommons.org/publicdomain/zero/1.0/.

[11] Gilchrist MA, Chen WC, Shah P, Landerer CL, Zaretzki R. Estimating Gene Expression and Codon-Specific Translational Efficiencies, Mutation Biases, and Selection Coefficients from Genomic Data Alone. Genome Biology and Evolution. 2015;7:1559-79.

[12] Landerer C, Cope A, Zaretzki R, Gilchrist MA. AnaCoDa: analyzing codon data with Bayesian mixture models. Bioinformatics. 2018:bty138. Available from: http://dx.doi.org/10.1093/bioinformatics/bty138.

[13] Akashi H, Gojobori T. Metabolic efficiency and amino acid composition in the proteomes of Escherichia coli and Bacillus subtilis. Proceedings of the National Academy of Sciences of the United States of America. 2002 3;99:3695-700. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11904428 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC122586.

[14] Mohammad F, Green R, Buskirk AR. A systematically-revised ribosome profiling method for bacteria reveals pauses at single-codon resolution. eLife. 2019 2;8.