Quantifying shifts in natural selection on codon usage between protein regions: A population genetics approach

Alexander L. Cope[1,*,+] and Michael A. Gilchrist[1,2,3]

[1]Genome Science and Technology, University of Tennessee, Knoxville

[2]Department of Ecology and Evolutionary Biology, University of Tennessee, Knoxville

[3]National Institute for Mathematical and Biological Synthesis, Knoxville, TN

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

[+]Current: Department of Genetics, Rutgers University

Last compiled on Monday 1st March, 2021 at 15:46.

Keywords: codon usage bias, protein folding, protein secondary structure, population genetics

Abstract

Codon usage bias (CUB), the non-uniform usage of synonymous codons, occurs across all domains of life. Adaptive CUB is hypothesized to result from selection for efficient ribosome movement and/or accurate translation. Empirical and computational evidence indicate changes to missense error rate and translation speed can impact the ability of a protein to reach its native state. Given the critical link between folding and function, numerous studies have analyzed the relationship between codon usage and protein structure. Results have often been contradictory due to the various methods for quantifying differences in codon usage and not appropriately controlling for confounding factors (e.g. amino acid biases). To avoid these shortcomings, we take an explicit population genetics approach to quantify codon-specific shifts in natural selection related to protein structure. Our results reveal a weak relationship between codon usage and protein structure, indicating that differences in selection between structures are very subtle and/or occur intermittently. While the magnitude of differences in selection appear slight, our results indicate the relationship between codon usage and protein structure is more complex than previously believed. Our work demonstrates the statistical power and benefits of studying selective shifts on codon usage or other genomic features from an explicitly evolutionary approach.

Introduction

Patterns of codon usage bias (CUB), or the non-uniform usage of synonymous codons, vary both within and across species (Plotkin and Kudla, 2011). Although non-adaptive evolutionary forces, such as mutation bias, genetic drift, and recombination, are well-known to shape codon usage patterns, natural selection also plays a significant role. The correlation between codon frequency and tRNA abundances (Ikemura, 1981) and stronger codon bias in highly expressed genes (Bulmer, 1991; Gilchrist et al., 2015; Shah and Gilchrist, 2011) suggests selection against translation inefficiency is a major factor shaping genome-wide codon patterns.

Although adaptive CUB is thought to be primarily driven by selection for translation efficiency, some research indicates potential selective advantages of inefficient codons (sometimes referred to as "non-optimal" or "rare" codons) at certain sites within a protein (Chaney and Clark, 2015). Given

that codon usage patterns are strongly shaped by amino acid biases, mutation biases, and gene expression, it is important for researchers investigating possible adaptive codon usage patterns to ensure that these patterns cannot be explained by non-adaptive factors. Gould and Lewontin (1979) highlighted the bias of biologists towards adaptationist storytelling, arguing that non-adaptive evolutionary forces (e.g. genetic drift) and other constraints (e.g. developmental constraints) should be given greater consideration. With the ushering in of the genomic-era, similar arguments have been made about the importance of testing hypotheses related to selection on and adaptation of genomic features in the context of the evolutionary null, i.e. the expectation in the absence of selection (Lynch, 2007; Koonin, 2016).

A significant portion of research on codon usage has focused on the relationship between codon usage and protein folding. It is generally accepted that misfolded proteins have impaired function, can aggregate within the cell, and possibly disrupt key cellular processes (Drummond and Wilke, 2008; Geiler-Samerotte et al., 2011; Gidalevitz et al., 2009). Missense errors are hypothesized to increase the frequency of protein misfolding; thus, regions important for the protein folding are expected to be under stronger selection for translationally-accurate codons (Drummond and Wilke, 2008). Codon usage is also thought to modulate protein folding via changes in the elongation rates at key steps during co-translational folding (O'Brien et al., 2014). Empirical evidence indicate changes to elongation rates via synonymous codon usage can alter co-translational protein folding in organisms ranging from bacteria to multicellular eukaryotes (Buhr et al., 2016; Fu et al., 2016; Holtkamp et al., 2015; Walsh et al., 2020; Yu et al., 2015; Zhao et al., 2017). Importantly, synonymous changes to codon usage are known to impact cellular fitness and have been implicated in human diseases (Kimchi-Sarfaty et al., 2007; Simhadri et al., 2017) through altered protein folding. Understanding the role of codon usage in protein folding also has biotechnological significance, as a recombinant protein is often expressed in an organism with a drastically different CUB, potentially perturbing co-translational folding (Buhr et al., 2016).

Given that protein secondary structures generally differ in physicochemical properties and ability to co-translationally fold, researchers hypothesized that different structures will exhibit different codon usage patterns (Chaney and Clark, 2015; Pechmann and Frydman, 2013; Saunders and Deane, 2010). Numerous studies examining the relationship between codon usage and protein sec-

ondary structure, but a general relationship, if any, remains unclear (Brunak and Engelbrecht, 1996; Gupta et al., 2000; Pechmann and Frydman, 2013; Saunders and Deane, 2010; Tao and Dafu, 1998; Thanaraj and Argos, 1996; Chaney et al., 2017). Analyses of codon usage patterns at higher levels of protein structure have also been inconsistent (Homma et al., 2016; Zhou et al., 2015; Chaney et al., 2017; Jacobs and Shakhnovich, 2017; Saunders and Deane, 2010).

A likely cause for inconsistent results between studies are the differences in metrics for quantifying codon usage. Metrics based on tRNA abundances, such as the tRNA Adaptation Index (tAI) (dos Reis et al., 2004), often do not consider relative differences between synonymous codons, but absolute differences across all codons. This makes many metrics based on tRNA abundances particularly prone to amino acid biases when comparing codon usage patterns (Chaney and Clark, 2015; Cope et al., 2018). Furthermore, while many studies often delineate codons into subsets of "optimal" and "non-optimal" codons, the criterion for classification often varies between studies (Pechmann and Frydman, 2013; Zhou et al., 2009). Previous work concluded identifying the optimal codon for an amino acid can be problematic if the effects of mutation bias are ignored (Shah and Gilchrist, 2011; Gilchrist et al., 2015; Wallace et al., 2013). When determining codon optimality using tRNA-based metrics, this can lead to the odd situation where all synonyms for an amino acid fall into one group (Pechmann and Frydman, 2013). In some cases, studies examining adaptive codon usage related to protein folding have failed to control for common confounding factors, such as amino acid biases and gene expression (Chaney and Clark, 2015). Although other studies have attempted to control for these confounding factors, these approaches were somewhat *ad-hoc*.

Recent work has relied on comparative approaches to examine the functional relationship between codon usage and protein structure, appropriately recognizing that purifying selection would lead to conserved codon usage patterns (Chaney et al., 2017; Pechmann and Frydman, 2013), although these approaches do not explicitly model evolutionary processes (selection, mutation, drift, etc.) Alternative to species-based comparative approaches are population genetics approaches which explicitly attempt to model such evolutionary processes. As noted by Lynch (2007), "nothing in evolution makes sense except in the light of population genetics." Population genetics based approaches have been used in various context to examine selection on codon usage (Bulmer, 1991; Gilchrist, 2007; Shah and Gilchrist, 2011; Gilchrist et al., 2015; Lawrie et al., 2013; Machado et al.,

2020). One particularly powerful population genetics approach is the Ribosomal Overhead Cost Stochastic Evolutionary Model of Protein Production Rates (ROC-SEMPPR), which is able to separate out the effects of mutation and selection on codon usage by accounting for the natural variation in intergenic gene expression (Shah and Gilchrist, 2011; Gilchrist et al., 2015; Wallace et al., 2013). Unlike many approaches which either average codon usage over regions using heuristic metrics (e.g. CAI) or delineate codons as either optimal or non-optimal, ROC-SEMPPR provides estimates of selection and mutation bias at a codon-level resolution.

ROC-SEMPPR was originally developed for estimating selection and mutation biases based on genome-wide codon frequencies, but recent work has used ROC-SEMPPR to investigate both intragenic and intragenomic differences in codon usage patterns (Cope et al., 2018; Landerer et al., 2020). ROC-SEMPPR is implemented in a Bayesian framework (Gilchrist et al., 2015; Landerer et al., 2018), allowing for model comparisons using information criterion. Here, as a proof of principle, we tested for differences in selection on codon usage related to protein secondary structures and IDRs in *S. cerevisiae* and *E. coli*, two common model organisms for studying codon usage bias. Although model comparisons indicate selection on codon usage differs across protein structures in both species, these differences are relatively small, suggesting that differences in selection related to folding of these different structures is likely rare, weak, or both. We find some evidence of position-specific shifts in natural selection within helices of both *S. cerevisiae* and *E. coli*. However, we also find that previous conclusions regarding enrichment of slow translating codons near the start of helices (Pechmann and Frydman, 2013) were the result of biases in amino acid usage and/or failing to control for the effects of gene expression on codon usage in helices. Overall, this work demonstrates the power of population genetics approaches for testing hypotheses related to intragenic differences in selection on codon usage.

Materials and Methods

Protein-coding sequences (CDS) and amino acids sequences for *S. cerevisiae* S288c (GCF_000146045.2) and *E. coli* K12 MG1655 (GCF_000005845.2) were downloaded from NCBI Refseq. Previous analysis of CUB in *E. coli* indicated approximately 750 genes had outlier codon usage patterns, many

of which were hypothesized to be due to horizontal gene transfer (dos Reis et al., 2003). Fitting these outlier genes with ROC-SEMPPR revealed selection on codon usage within these genes was anti-correlated with the remaining genes (Cope et al., 2018). Here, our analysis of *E. coli* excludes all of these outlier genes.

Identifying protein secondary structure

Our analysis makes use of both protein secondary structures determined empirically via methods like X-ray crystallography, and computationally via methods like PsiPred (Jones, 1999). The empirical data is a more conservative dataset, with fewer proteins available but more accurate and reliable designations of protein secondary structures. Predicted secondary structures allow us to analyze a larger set of proteins, increasing our statistical power, but includes the possibility that our analysis will be confounded by biases in the prediction algorithm. The current implementation of PsiPred has an overall accuracy score of 84% (Buchan and Jones, 2019), but secondary structure prediction algorithms generally struggle with accurately identifying the termini of secondary structures (Saunders and Deane, 2010; Pechmann and Frydman, 2013). Therefore, analyses of codon usage at secondary structure termini were based exclusively on empirically-determined secondary structures.

Empirically-determined protein secondary structures and corresponding protein sequences were obtained from the Protein Data Bank (PDB). Residues were grouped into three overarching structural groups based on their DSSP classification: helix (DSSP H, G, and I), sheet (DSSP E and B), and coil (DSSP S, T, and '.'). This classification system is consistent with secondary structure prediction algorithms (Jones, 1999) and other analyses of codon usage patterns based on empirically-determined secondary structures (Adzhubei et al., 1996; Thanaraj and Argos, 1996; Pechmann and Frydman, 2013; Saunders and Deane, 2010; Tao and Dafu, 1998). Note that the classification symbol (.) is a catchall containing any amino acids not matching any other DSSP classifications. Protein sequences from PDB were aligned to the *S. cerevisiae* and *E. coli* proteomes using BLAST. Sequences were considered mapped to the proteomes if the PDB sequence covered 80% of the length of the protein and had a percent identity score of 95% or higher. This provided us with 1,097 and 1,285 protein sequences with empirically-determined secondary structures in

S. cerevisiae and *E. coli*, respectively. This dataset was used for comparing selection on codon usage between and within secondary structures.

Protein secondary structures were predicted for all nuclear protein sequences for *S. cerevisiae* and for 1,742 proteins in *E. coli* using the PsiPred software (Jones, 1999) at default settings. PsiPred combines the secondary structural classifications of DSSP into helices (H), sheets (E), and coils (C).

Identifying structured and intrinsically-disordered regions

Unlike protein secondary structures, empirically-determined intrinsically-disordered regions (IDRs) are rare. The DisProt database includes only 134 proteins with IDRs for *S. cerevisiae*. Thus, our analysis of codon usage patterns in IDRs and structured regions in *S. cerevisiae* and *E. coli* relied on predicted IDRs using IUPRED2 (Mészáros et al., 2018), which provides a quasi-probability of the an amino acid falling into a disordered region, using default settings. An amino acid with a quasi-probability of greater than 0.5 is more likely to be disordered, while a quasi-probability less than 0.5 is more likely to be structured; thus, amino acids with a score less than or equal to 0.5 were classified as structured, while amino acids with a score greater than 0.5 were classified as disordered, consistent with the analysis done by Zhou et al. (2015).

Analysis with ROC-SEMPPR

All analyses of codon usage bias was performed using ROC-SEMPPR with the R package AnaCoDa (Landerer et al., 2018). We note ROC-SEMPPR assumes weak selection. To meet this assumption, serine was split into separate codon groups: the 4 codon group TCN (S_4) and the 2 codon group AGN (S_2).

For any amino acid with n_{aa} synonymous codons, the probability of observing codon i in gene g can be described by the equation

$$p_i = \frac{e^{-\Delta M_i - \Delta\eta_i \phi_g}}{\sum_j^{n_{aa}} e^{-\Delta M_j - \Delta\eta_j \phi_g}}$$

where ΔM represents mutation bias, $\Delta\eta$ represents natural selection, and ϕ represents the evolu-

tionary average protein production rate of gene g. Note that Δ indicates the mutation bias and natural selection parameters are relative to a reference codon. More specifically, ROC-SEMPPR's mutation bias ΔM parameter represents the log of the ratio of mutation rates between two synonymous codons (Gilchrist et al., 2015). Although originally described as being proportional to relative differences in translation efficiency between two synonymous codons (Gilchrist et al., 2015), $\Delta\eta$ more generally reflects the population genetics parameter sN_e, where N_e is the effective population size and s represents the selection coefficient relative to the reference codon. Assuming a gene with an average protein production rate ϕ, $\Delta\eta$ represents the strength and direction of natural selection for a codon, proportional to the strength of genetic drift. A deeper understanding of the model parameters can be obtained by considering the cases of no protein production $\phi = 0$ and average protein production $\phi = 1$. We note that ROC-SEMPPR scales ϕ such that the $E[\phi] = 1$. In the case of no protein production, natural selection on codon usage is completely absent, resulting in mutation biases determining the synonymous codon frequencies. In case of average protein production, the synonymous codon frequencies will reflect the relative strengths and directions of mutation bias and natural selection (proportional to drift). For example, if the mutation bias is stronger and in the opposite direction of natural selection (i.e. mutation and selection favor different codons), then the mutationally-favored codon is expected to be more frequent in an average expression gene. Importantly, ϕ scales the strength of natural selection such that strong mutation biases can be (but not necessarily will be) overwhelmed in highly expressed genes. Previous work indicates ROC-SEMPPR's parameter estimates correlate well with empirical measurements in *S. cerevisiae* and *E. coli* (Gilchrist et al., 2015; Cope et al., 2018).

As a first step, ROC-SEMPPR was fit to all protein-coding sequences in *S. cerevisiae* and *E. coli* to obtain gene-specific estimates of protein production rates ϕ and codon-specific estimates of mutation bias ΔM, as in Cope et al. (2018). Protein-coding sequences were then partitioned based on the codons corresponding secondary structural category (based on empirically-determined or predicted structures). In order to determine if codon usage patters are statistically different between protein secondary structures, structural groupings were combined, e.g. helices and sheets were combined into one group as opposed to treating them as separate group. These structural groupings were then further merged into different models such that each structure category was

represented once, either as a standalone group (e.g. helix) or grouped with another secondary structure (e.g. helix and sheet as one group). To be clear, "different models" simply refers to different ways in which different secondary structures were grouped. This ensured the sequence data (i.e. the number of codons, protein-coding sequences, etc.) is the same across all models, making them directly comparable. We note that the first 35 codons of all genes were excluded to help reduce the impact of a weaker selection on codon usage at the 5'-end of genes (Gilchrist et al., 2009).

A similar analysis to the one outlined above for comparing codon usage between secondary structures was performed based on the predictions using IUPRED2, in which we compared a model which had structured and disordered regions as separate groupings to a model which treated them as one grouping. Finally, an analysis was performed which combined information from PsiPred and IUPRED2 to classify amino acids based on both methods for classifying structural information. This allowed us to distinguish coils which may be found as part of a larger structured domain from coils part of intrinsically-disordered regions.

To examine variation in codon usage within secondary structures, empirically-determined secondary structures were divided into the N-terminus and C-terminus regions, with all codons in between being classified as the core of the secondary structures. To assess robustness of our results, we varied the minimum number of amino acids for a secondary structure between 4 and 10 and varied the termini length to be 2 or 3 amino acids. To test the hypothesis presented by Pechmann and Frydman (2013) in *S. cerevisiae*, helices of minimum length 6 amino acids were split up into the second and third codons (relative to the start of the helix) and the remainder of the helix.

For statistically comparing codon usage patterns, ROC-SEMPPR model fits were compared using the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002). Briefly, DIC is a Bayesian information criterion which tries to balance the overall model fit to the data as determined by the posterior distribution and the number of parameters used to fit the data. If the level or nature of selection on codon usage differs between two structures, then it is expected a model treating these structures as separate groupings will have a better (lower) DIC score than model fits treating the structures as single (or merged) groupings. We follow the general rules of thumb for comparing models using information criterion (Burnham and Anderson, 2004). A model that

differs from the minimum DIC model by fewer than 2 DIC units has substantial statistical support. A difference in the range of the 2-4 DIC units are still considered to have strong support, while a difference of 4-7 DIC units are considered to have less support. However, a model that differs from the minimum DIC model by 10 or more DIC units can generally be disregarded. We note that all ΔDIC values will represent $\text{DIC}_i - \text{DIC}_{\text{Best}}$, where DIC_i is the DIC score of the ith model and DIC_{Best} is the minimum DIC score (i.e. the best model) of the models under consideration. This means that all reported ΔDIC values will fall into the range $[0, \infty)$.

Comparing models via DIC indicates differences in selection on codon usage between structural regions, but does not tell us how they differ. Similar to Cope et al. (2018), we broadly compared codon-specific estimates of selection $\Delta\eta$ between structural groupings using a model-II regression, which accounts for errors in both the independent and dependent variables (Sokal and Rohlf, 1995). In this work, we used the Deming Regression, as implemented in the R package **deming**. A Deming regression on $\Delta\eta$ estimated from different structural grouping with a slope of $\hat{\beta} = 1$ (i.e. $y = x$) would suggest there is not a general shift in natural selection on codon usage between the two groupings that can be described by the functional relationship $\Delta\eta_B = \beta\Delta\eta_A$. On the other hand, a Deming regression slope $\hat{\beta}$ significantly different from 1 is consistent with an overall shift in natural selection on codon usage between the two structural groupings being compared. For each amino acid, its corresponding $\Delta\eta$ values were scaled relative to its optimal synonymous codon, i.e. the one most favored by natural selection based on fitting the null model where selection on codon usage does not vary across structural categories; specifically, models Y_0 and E_0 for *S. cerevisiae* and *E. coli*, respectively. As a result, the null model optimal codon $\Delta\eta$ value is always 0 and its synonyms are always $\Delta\eta > 0$, unless there's a structure-specific shift in the optimal codon. In this case, there can be $\Delta\eta$ values less than 0. Importantly, the Deming regression only summarizes a possible overall shift in the strength or direction of selection on codon usage between two structural groupings, but does not rule out the possibility that selection is different between specific codons. This information can be obtained by comparing the $\Delta\eta$ estimates for a codon across the different protein structures. We focus on the codons for which the $\Delta\eta$ 95% posterior probability intervals do not overlap between structural groupings, as we are most confident in the sign of any shift in selection.

Estimates of elongation and missense error rates

To obtain estimates of elongation rates and missense error rates, we used the simple, mechanistic model presented in Shah and Gilchrist (2010). Briefly, this model estimates codon-specific average elongation rates and missense error rates given a set of tRNA abundances and a list of cognate and near-cognate tRNAs for each codon (Fluitt et al., 2007). For *E. coli*, the default values described in Shah and Gilchrist (2010) were used. For *S. cerevisiae*, the average elongation rate was set to 9 amino acids per second and the background nonsense error rate was set to 10^{-3} per second, consistent with previous work (Gilchrist and Wagner, 2006; Gilchrist et al., 2009). Empirical estimates of codon-specific average waiting times (based on ribosome densities determined via ribosome profiling) for *S. cerevisiae* were taken from Weinberg et al. (2016).

Evaluating effects of confounding factors in analyses of position-specific codon usage

Previous work found that positions 2 and 3 (relative to the start) of helices in *S. cerevisiae* were enriched in non-optimal codons (Pechmann and Frydman, 2013). To determine if this pattern can be generated by amino acid biases or gene expression, the yeast genome was independently simulated 100 times using **AnaCoDa** (Landerer et al., 2018) under the ROC-SEMPPR model with the genome-wide estimates of selection $\Delta\eta$ and mutation bias ΔM, as well as the gene-specific estimates of protein production rates ϕ, such that the nature of selection was the same for every codon across the genome. Proteins used by Pechmann and Frydman (2013) for their analysis of position-specific codon usage were pulled and all helices were aligned by position. For each position, enrichment of non-optimal codons based on the nTE metric (as defined in Pechmann and Frydman (2013)) was tested using a Fisher's exact test. This generated a distribution of 100 odds ratios per position, which were then compared to the reported odds ratios in Pechmann and Frydman (2013).

Results

Based on predicted secondary structures from PsiPred (Jones, 1999), we found that the best supported model in *S. cerevisiae* allowed selection on codon usage to differ between helices, sheets, and coils (Table 1, Model Y_1). This model is 68 DIC units better than the next best model assuming no difference in selection between helices and sheets (Y_2), and 621 DIC units better than the null model assuming no difference between any . Overall, we obtained similar results when using empirically-determined secondary structures, but the best two models were ambiguous as to whether selection differed between helices and sheets (Table S1, Y_1 vs Y_2, ΔDIC = 1.16).

Similar results were obtained for *E. coli*. The best overall model fit allowed for selection to differ between helices, sheets, and coils, with a 61 DIC unit improvement over the model assuming no differences between helices and sheets (E_2) and 471 DIC unit improvement over the null model (E_0). Unlike *S. cerevisiae*, similar model fits in *E. coli* using empirical data clearly differentiated between the best two models (Table S2, E_1 vs E_2, ΔDIC = 19.74).

Table 1: Comparison of model fits examining variation in codon usage between predicted protein secondary structures. The null model (Y_0) assumes no differences in selection between secondary structures). H: helix. E: sheet. C: coil. $\Delta\text{DIC} = \text{DIC}_i - \text{DIC}_{\text{Best}}$.

		Groupings			
Species	Model	I	II	III	ΔDIC
S. cerevisiae	Y_1	H	E	C	0
	Y_2	HE	–	C	68
	Y_3	H	EC	–	285
	Y_4	HC	E	–	425
	Y_0	HEC	–	–	621
E. coli	E_1	H	E	C	0
	E_2	HE	–	C	61
	E_3	H	EC	–	251
	E_4	HC	E	–	358
	E_0	HEC	–	–	471

Comparing selection $\Delta\eta$ on codon usage between secondary structures

Given the model fits indicate differences in selection on codon usage between protein secondary structure, we compared estimates of $\Delta\eta$ to determine the differences in strength and direction of selection. Comparing selection estimates $\Delta\eta$ with a Deming regression revealed no significant differences between any of the three secondary structures (Figure 1) in either species, indicating no <u>overall</u> selective shifts between any of the three secondary structures.

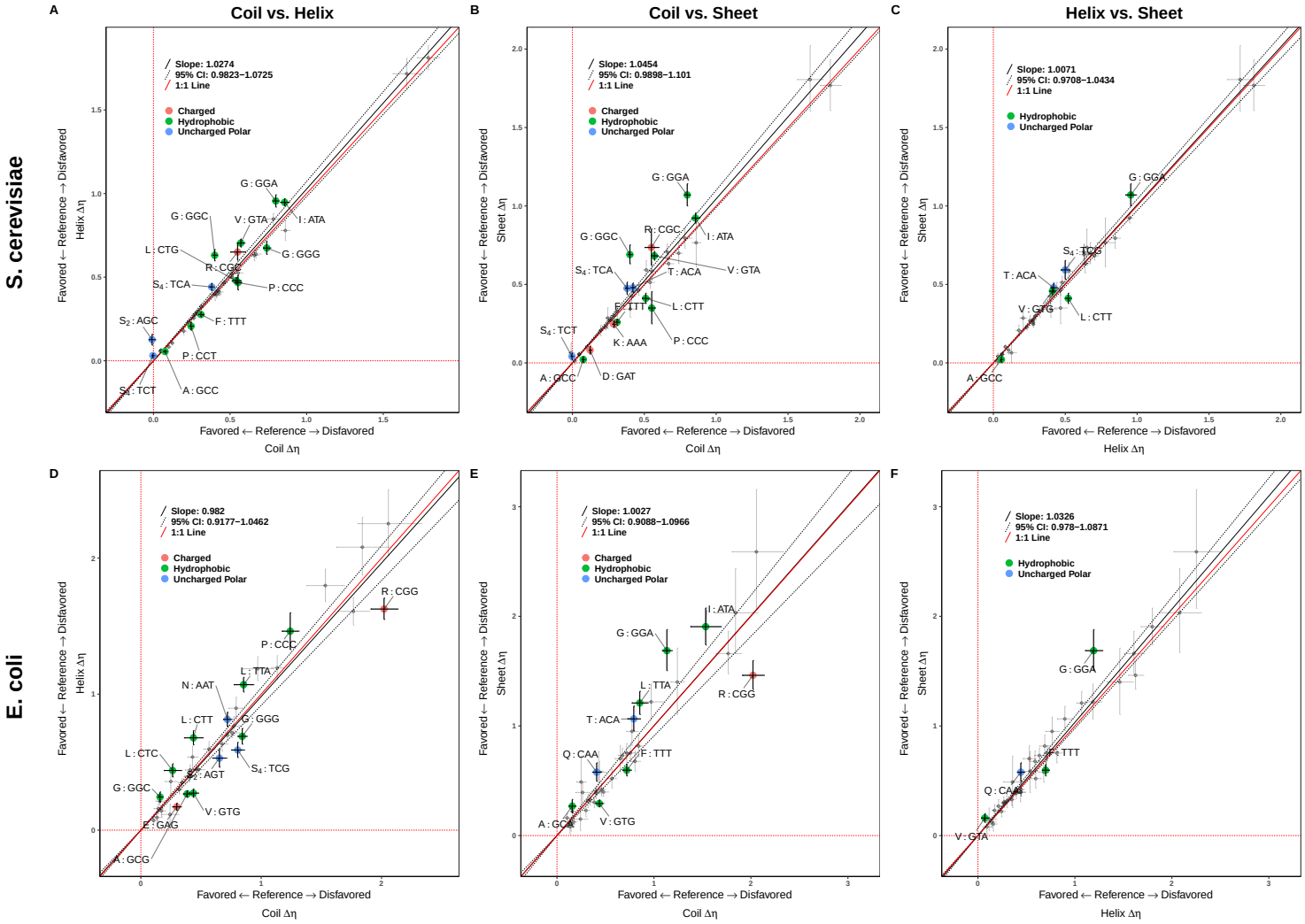

Figure 1: Comparison of selection estimates $\Delta\eta$ between different protein secondary structures. $\Delta\eta$ values are scaled relative to the genome-wide optimal codon. Points represent $\Delta\eta$ values for each codon, while error bars represent the 95% posterior probability intervals. Codons showing significant selective shifts are colored by amino acid property. Negative values indicate a change in the selectively-favored codon relative to the genome-wide optimal. The Deming regression slope and 95% confidence intervals are represented by solid and dashed black lines, respectively. **(A,D)** Coils vs. Helices. **(B,E)** Coils vs. Sheets. **(D,F)** Helices vs. Sheets.

.

Although no overall selective shifts on codon usage were detected, examination of the 95% posterior probability intervals for $\Delta\eta$ reveals clear differences in the strength of selection related to individual codons. We find that for most of the 18 amino acids with multiple synonyms, selection differs between secondary structures for at least one codon (i.e. its $\Delta\eta$ 95% posterior probability

intervals do not overlap). For both *S. cerevisiae* and *E. coli* we observe shifts in the following 10 amino acids: alanine (A), glycine (G), serine (S_4 and S_2), phenylalanine (F), leucine (L), valine (V), arginine (R), proline (P), isoleucine (I), and threonine (T) (Figure 1). In addition, *S. cerevisiae* shows selective shifts in aspartic acid (D) and lysine (K), while *E. coli* shows shifts in glutamic acid (E), glutamine (Q), and asparagine (N). Most of these codons code for non-polar, hydrophobic amino acids and most of these shifts occurred between coils and the other two secondary structures. Importantly, these differences mostly reflect changes in the average strength of selection and not a switch in the optimal codon. An exception to this appears to be serine (S_4 and S_2) in coils of *S. cerevisiae*. While codons TCT and AGC are disfavored in helices and/or sheets, this is not the case for coils in which there appears to be no differences in the preference for these two codons and the genome-wide optimal codons (TCC and AGT, respectively, Figure 1A,B). There appear to be very few differences in selection between helices and sheets in both species. This weak effect likely explains the ambiguous results of the model comparisons when using empirically-determined secondary structures in *S. cerevisiae*, as this dataset represents approximately $1/6^{\text{th}}$ of the *S. cerevisiae* protein-coding sequences.

Intrinsically-disordered Regions Show Distinct Patterns of Selection on Codon Usage

Consistent with Zhou et al. (2015), we found that whether a region was structured or intrinsically-disordered better explained intragenic codon usage patterns than protein secondary structures in *S. cerevisiae* (Table 2, Models Y_5 vs. Y_1). Consistent with this, selection was 9% weaker, on average, in IDRs compared to structured regions (Deming Regression $\hat{\beta} = 0.9053$, 95% CI: 0.8229 – 0.9876, Figure 2A). In contrast, splitting codons into structured regions and IDRs in *E. coli* did a worse job of explaining intragenic codon usage patterns than secondary structures (Table 2, Model E_1 vs. E_5). This was unsurprising given the rarity of IDRs in prokaryotic proteomes (Basile et al., 2019). Despite being a worse fit compared to the secondary structure model in *E. coli*, the structured regions vs. IDR model is still a significant improvement over the null model (Tables 1 and 2, Model E_5 vs. E_0, ΔDIC = 129). Even though the Deming regression slope comparing

structured regions and IDRs in *E. coli* was of similar magnitude to the same slope estimated for *S. cerevisiae*, it was not significant (Figure 2B, Deming Regression $\hat{\beta} = 0.9327$, 95% CI: 0.8491 – 1.0163).

Although selection on codon usage was weaker, on average, in IDRs of *S. cerevisiae*, we note a subset of amino acids demonstrate the opposite pattern in which selection against certain codons appears to be stronger: alanine (A), histidine (H), lysine (K), proline (P), and threonine (T) (Figure 2A). In addition, serine (S_4 and S_2) shows shifts in the selectively-favored codon in IDRs relative to structured regions, with the former showing preference for TCT and AGC. This is similar to the results observed in coils for *S. cerevisiae*, but in this case, the selective shifts clearly indicate one of the codons is preferred over the other (i.e. the 95% posterior probability intervals do not overlap with 0, Figure 2A). Alanine (A), serine (S_4), and threonine (T) showed a similar pattern in *E. coli* (Figure 2B). Interestingly, many of these amino acids have a higher propensity for forming disordered regions and/or serve as sites for phosphorylation (Pal Singh, 2015; Uversky, 2015).

We found that categorizing the predicted structured regions based on the corresponding secondary structure predictions improved the overall model fits in both species (Models Y_6 and E_6), indicating differences in codon usage between secondary structures are not solely due to the presence of IDRs, contrary to the findings of Zhou et al. (2015). Comparing selection estimates $\Delta\eta$ from the secondary structures (with IDRs removed) to $\Delta\eta$ estimated from IDRs suggests selection on codon usage is, on average, stronger in coil and sheet secondary structures compared to IDRs in *S. cerevisiae* (Figure 2A,C). Although a comparison of helices to IDRs has a slope estimate consistent with stronger selection in helices, this slope is not statistically significant (Deming regression $\hat{\beta} = 0.9046$, 95% CI: 0.8046 – 1.0027). In *E. coli*, all three Deming regression slopes are biased downwards, but only the comparison between coils and IDRs is statistically significant (Figure 2F–H). When comparing $\Delta\eta$ between secondary structures after removing IDRs, we found that many codons still exhibited significant selective shifts between secondary structures (Figure S1). Notably, the observed selective shifts in *S. cerevisiae* on codons TCT (S_4) and AGC (S_2) in coils appears weakened or missing when IDRs are removed, suggesting the previously observed results were primarily driven by differences in selection in IDRs.

Table 2: Model comparisons of structure categorizations based on the Deviance Information Criterion (DIC), where the smallest value is considered the best model fit. For simplicity, only models which are an improvement over the model over the best secondary structure model (Table 1) are shown, with the exception of the Structured and IDR model for *E. coli*. Struct H,E,C and IDR H,E,C indicates the secondary structure predicted by PsiPred was further split into regions predicted to be structured and disordered by IUPRED2. Struct = structured region. IDR = intrinsically disordered region. H: helix. E: sheet. C: coil

		Groupings						
Species	Model	I	II	III	IV	V	VI	ΔDIC
S. cerevisiae	Y_{10}	Struct H	IDR H	E	–	Struct C	IDR C	0
	Y_9	Struct H	IDR H	Struct E	IDR E	Struct C	IDR C	0.46
	Y_8	H	–	E	–	Struct C	IDR C	83
	Y_7	H	–	Struct E	IDR E	Struct C	IDR C	84
	Y_6	Struct H	–	Struct E	–	Struct C	IDR	117
	Y_5	Structured	–	–	–	–	IDR	466
	Y_1	H	–	E	–	C	–	686
E. coli	E_6	Struct H	–	Struct E	–	Struct C	IDR	0
	E_8	H	–	E	–	Struct C	IDR C	29
	E_{10}	Struct H	IDR H	E	–	Struct C	IDR C	30
	E_7	H	–	Struct E	IDR E	Struct C	IDR C	45
	E_9	Struct H	IDR H	Struct E	IDR E	Struct C	IDR C	47
	E_1	H	–	E	–	C	–	77
	E_5	Structured	–	–	–	–	IDR	419

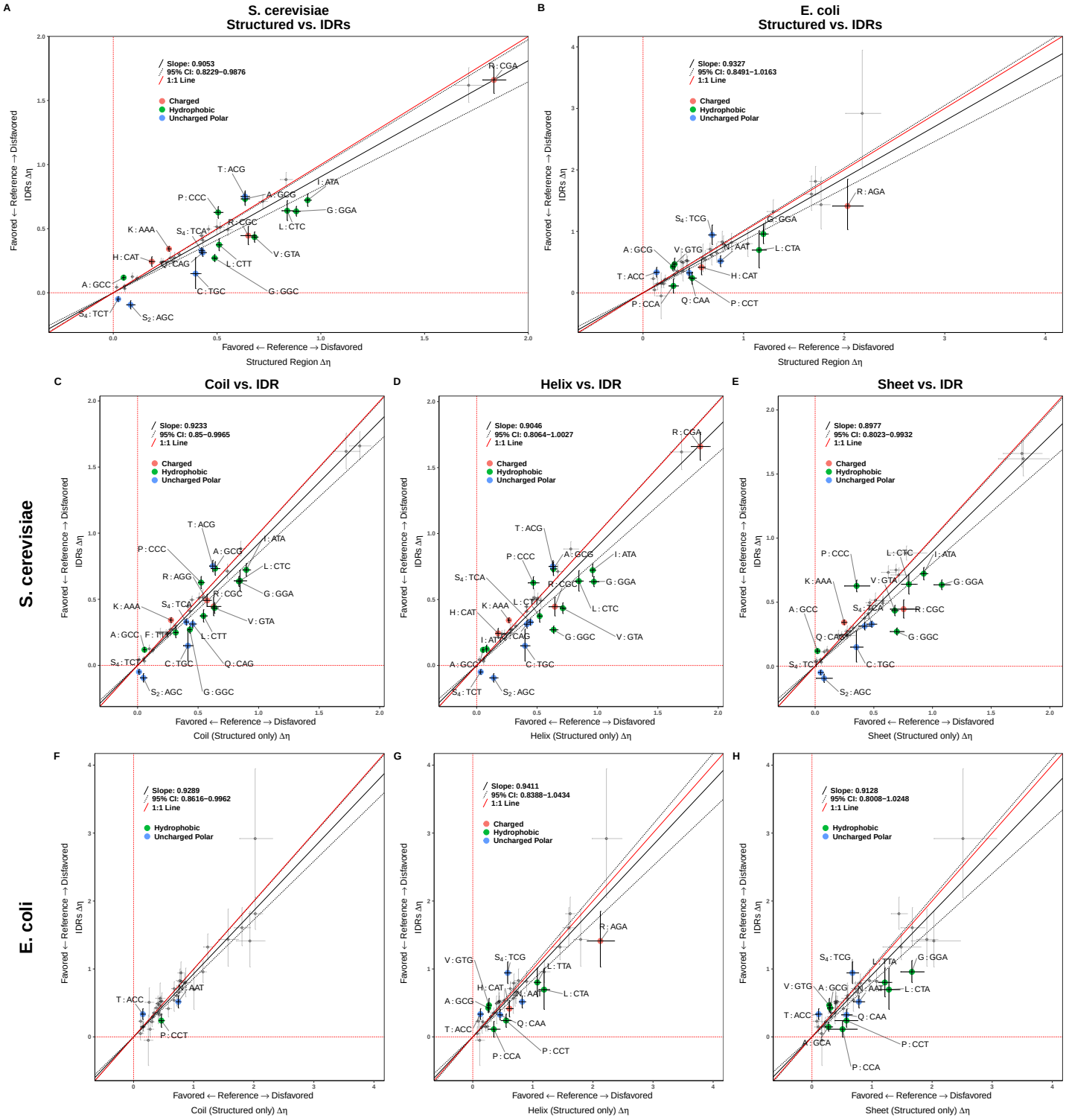

Figure 2: Comparison of selection estimates $\Delta\eta$ between structured regions and IDRs. Points represent $\Delta\eta$ values for each codon. Error bars represent the 95% posterior probability intervals. Codons showing significant selective shifts are colored by amino acid property. Negative values indicate a change in the selectively-favored codon relative to the genome-wide optimal. **(A,B)** Structured vs. IDRs. **(C,F)** Coils vs. IDRs. **(D,G)** Helices vs. IDRs. **(E,H)** Sheets vs. IDRs.

Given that all secondary structures were predicted to fall into structured or disordered regions, we tested whether further dividing up the secondary structures as structured or IDR explained codon usage patterns improved our model fits. Splitting up only the coils into structured and IDRs improved upon the model where IDRs were taken from all three secondary structure classifications in *S. cerevisiae* (Model Y_8 vs. Y_6), but was a worse fit in *E. coli* (Model E_8 vs. E_6). Surprisingly, the best overall model fit in *S. cerevisiae* divided both coils and helices into structured and IDRs (Model Y_{10}).

Examining effects of missense errors and elongation rates on codon-specific selective shifts

A possible explanation for the selective shifts between protein structures is increased selection against missense errors. Empirical evidence suggests some structures, such as sheets, are generally more prone to missense substitutions (Guo et al., 2004). Although high-throughput empirical estimates of missense errors are limited, mechanistic models of translation can provide us with expected missense error rates (Fluitt et al., 2007; Shah and Gilchrist, 2010). Using estimates of missense error rates based on a mechanistic model described by Shah and Gilchrist (2010), we find that absolute selective shifts are correlated with expected missense error rates in *E. coli* (Figure 3). Comparing the magnitude of selective shifts (regardless of the direction of the shift) between helices, sheets, and coils (all with IDRs removed) reveals weak to moderate, positive correlations with absolute missense error rates (Figure 3A–C), but only the shifts between coils and sheets, and helices and sheets are statistically significant. A similar observation is observed in the shifts between structured and IDR regions (Figure 3D).

Differences in codon usage between structures may also be due to selective shifts on elongation rate, such as in the case of slowing down translation to facilitate proper co-translational folding. The estimates of elongation rates based on the above model (Shah and Gilchrist, 2010) for *E. coli* are negatively correlated with the absolute shifts in $\Delta\eta$, suggesting the larger shifts are also occurring for slower codons Overall, model estimates of missense error rates and elongation rates are anti-correlated (Spearman Rank correlation $\rho = -0.56$, $p = 0.0001747$). To control for the relation-

ship between elongation rate and missense error rate, we computed partial Spearman correlation coefficients between missense error rates and absolute shifts while conditioning on the elongation rates. After controlling for the effects of elongation rates, the correlation of missense error rates with selective shifts between sheets and coils (partial Spearman correlation coefficient $\rho = 0.33$, $p = 0.0434$) and structured regions and IDRs (partial Spearman correlation coefficient $\rho = 0.36$, $p = 0.0256$) are reduced, but remain statistically significant.

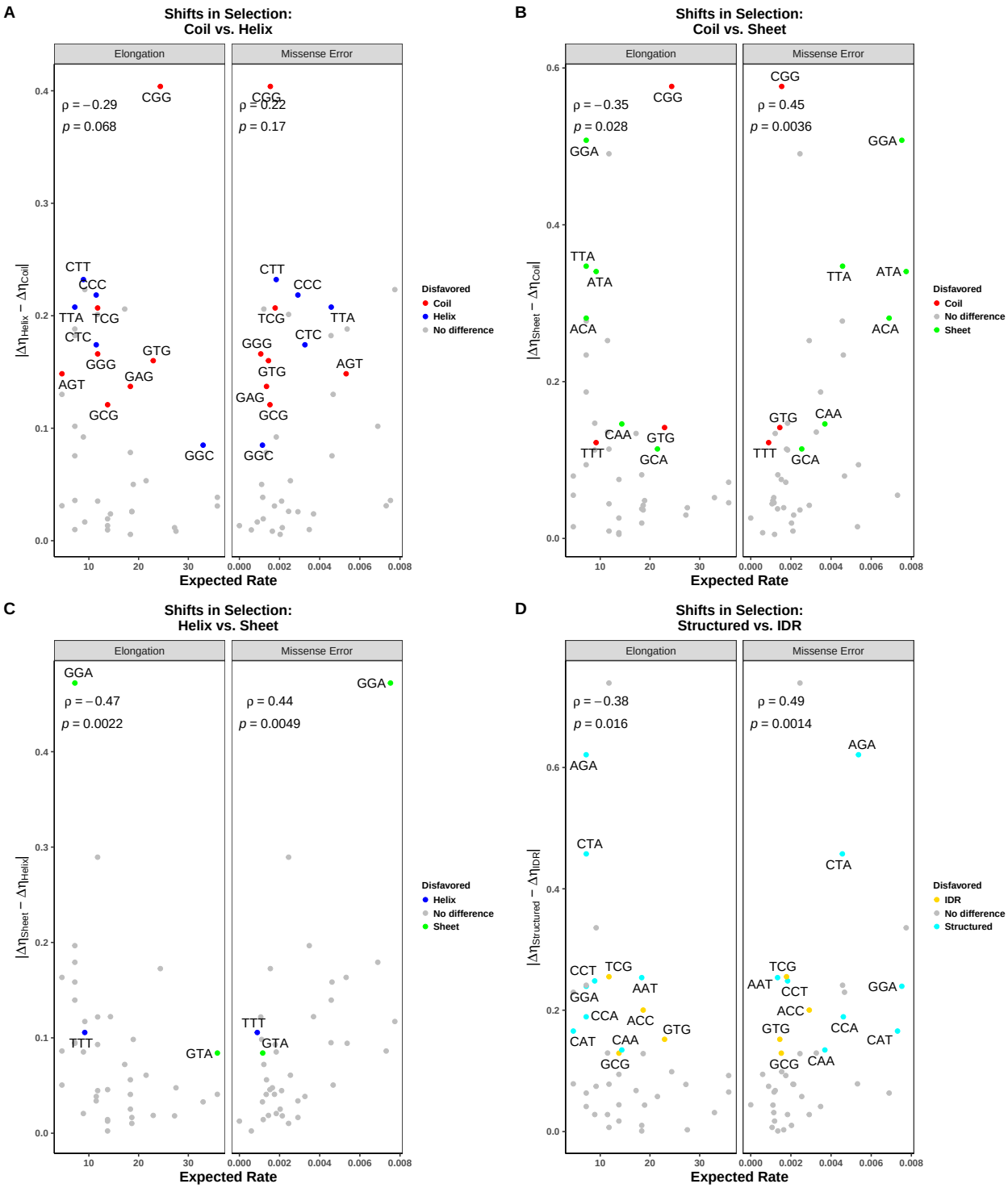

Figure 3: Comparison between expected missense error/elongation rates and absolute selective shifts $\Delta\eta$ between different protein structures in *E. coli*. Secondary structures include only those codons also predicted to be structured, removing possible effects of IDRs. Codons that showed significant differences between structures are labeled and the color indicates in which structure the codon was more disfavored. **(A)** Coils vs. Helices. **(B)** Coils vs. Sheets. **(C)** Helices vs. Sheets. **(D)** Structured vs. IDRs.

Interestingly, estimates of missense error and elongation rates from the same model (Shah and Gilchrist, 2010) for *S. cerevisiae* show no statistically significant correlations with absolute selective shifts (Figure S2). As an alternative to model-estimated elongation rates, we performed the same analysis using ribosome density estimates, which represent the average time a ribosome spends waiting to elongate a codon (i.e. the inverse of elongation rate) from a ribosome profiling experiment (Weinberg et al., 2016). We observe a strong correlation between ribosome densities and the absolute selective shifts, suggesting larger selective shifts are occurring for slower codons (Figure S3).

Selection on codon usage does not generally vary at the termini of secondary structures

In *S. cerevisiae*, we found no evidence that natural selection varies at the termini of sheets or coils based on model comparisons. This was regardless of our choice of minimum length of the structure or size of the termini (Table S3 – S12). However, model comparisons did not allow us to rule out the possibility of differences in natural selection at the termini of helices. When restricting our analysis to helices with a minimum of 4 codons and defining the N and C-terminus to be the first and last two codons, respectively, we found the model allowing for differences at the termini improved upon the model assuming no differences at the termini by 13.65 DIC units (Table S3). When helices were limited to those of length 6, the favored model did not allow for differences in selection between the core and the termini (Table S5 – S6). We tested if there was a general difference in selection between helices of lengths 4 or 5 vs those with a length 6 or greater, but we found no support for this model over grouping all helices together (ΔDIC = 27.06). Considering helices of length 4 vs those with a length 5 or greater yielded similar results (ΔDIC = 29.69). We note that the effect weakened if excluding 3_{10}-helices (DSSP G) and π-helices (DSSP I), but the overall patterns still suggests possible differences in selection at the termini of helices (Table S13); however, we found no evidence to support general differences in selection on codon usage between these two helical structures and the traditional α-helices (ΔDIC = 17.46). Regardless, when comparing the $\Delta\eta$ estimates from the termini and the core of the helices, we find that 4 codons are significantly

different: proline (P) codon CCC, glycine (G) codons GGA and GGC, and valine (V) codon GTT with the former (Figure S4).

Unlike in *S. cerevisiae*, for which differences in selection at the termini of helices is unclear, model comparisons clearly indicate differences at termini relative to the core in *E. coli* (Table 3 and Tables S14 – S16). We note that this pattern is consistent if restricting the minimum length of helices to 4 – 7 codons. In this case, we find that 8 codons are significantly different between the termini and core (Figure S5). On the other hand, sheets demonstrated variable patterns depending on the length. When restricting the length to a minimum of 4 or 5 codons, the model fits indicate there is a difference between the core and termini of sheets, but this disappears when restricting the analysis to only sheets of minimum length 6 or 7 codons (Table 3 and Tables S14 – S16). Although the same analysis in coils generally favored no differences in at the termini, we note that the ΔDIC scores were often less than 10 DIC units, meaning we could not confidently rule out differences in selection at the termini of coils. Interestingly, this effect disappeared when restricting the analysis to coils with at least 7 codons (Table S16).

Table 3: Comparing models with termini (first and last 2 codons) of secondary structures separated from the core of the structure in *E. coli*. Results are for secondary structures of minimum length 6. H: helix. E: sheet. C: coil. $\Delta\text{DIC} = \text{DIC}_i - \text{DIC}_{\text{Best}}$.

Structure Model	Secondary Structure	Groupings I	II	III	ΔDIC
E_{1b}	H	Termini	Core	–	0.0
E_{1c}		N-terminus	Core	C-terminus	12.55
E_{1a}		Whole Structure	–	–	63.71
E_{1e}	E	Whole Structure	–	–	0
E_{1g}		N-terminus	Core	C-terminus	37.08
E_{1f}		Termini	Core	–	42.74
E_{1h}	C	Whole Structure	–	–	0
E_{1i}		Termini	Core	–	6.21
E_{1j}		N-terminus	Core	C-terminus	13.75

Positions 2 and 3 of Helices are not Under Selection for Inefficient Codons

Although we found some support for differences in selection at the termini of helices in *S. cerevisiae*, we found no support for the model allowing for differences in selection at the second and third positions of helices (Pechmann and Frydman, 2013) relative to the model assuming no differences in selection within helical structures (ΔDIC = 29.75) (Pechmann and Frydman, 2013). Using simulated data that assumes the strength and direction of selection for a codon is constant across the entire genome, we found the odds ratios reported by Pechmann and Frydman (2013) were within the range of odds ratios generated using the simulated data (Figure S6). This suggests the enrichment of non-optimal codons at positions 2 and 3 of helices observed by Pechmann and Frydman (2013) are an artifact of amino acid biases and/or gene expression. Similarly, patterns consistent with the supposed enrichment of optimal codons at positions 1 and 4 are also observed in our simulated data.

Discussion

Although adaptive codon usage bias (CUB) is thought to be primarily shaped by selection for translation efficiency and accuracy (Plotkin and Kudla, 2011), selection to maintain correct protein folding (Drummond and Wilke, 2008; Chaney et al., 2017; Pechmann and Frydman, 2013) is thought to play a prominent role in intragenic variation in CUB. Synonymous changes have been empirically observed to alter the folded state of a protein, leading to reduced protein functionality and organismal fitness (Kimchi-Sarfaty et al., 2007; Walsh et al., 2020). However, establishing a general role for synonymous codon usage in protein folding has proven challenging, with various studies providing inconsistent results and interpretations (Nieuwkoop et al., 2020). A contributing factor to the inconsistencies are the various metrics for quantifying codon usage and the various potential confounding factors (effects of gene expression, amino acid biases, etc.) which are often ignored. Importantly, even though many previous approaches do not explicitly measure selection on codon usage, differences in codon usage are often interpreted in the context of natural selection.

This can be particularly problematic when confounding factors are not appropriately controlled.

In order to account for the effects of amino acid biases and gene expression, we used a population genetics-based model ROC-SEMPPR that explicitly includes the effects of mutation bias, selection, and genetic drift on synonymous codon usage patterns to test for differences in selection on codon usage related to protein structure in *S. cerevisiae* and *E. coli*. Using ROC-SEMPPR, we detected small, codon-specific selective shifts across protein secondary structures and IDRs. However, the direction and strength of selection between structures is overall the same across all structures. Using a simulated dataset based on the empirically-determined protein secondary structures (1,097 genes), we found we were able to detect overall selective differences between two regions even if only 1% of sites in one of the regions was under a different selective pattern, e.g. selection for folding related translation inefficiency instead of background translation efficiency (Figure S7). This assumes that (1) codon usage across all sites conforms to the assumptions of our model and (2) the nature of selection at the 1% of sites is perfectly anti-correlated with the other 99% of sites. Regarding (2), this assumes selection against translation inefficiency and selection against missense errors, both of which are thought to impact codon usage patterns related to protein folding, are perfectly correlated. However, if the abundance of near cognate tRNAs differs between synonyms, the most efficient codon may not necessarily be the most accurate one (Shah and Gilchrist, 2010). Our simulated results likely represent an approximate lower bound on the number of sites under differing selective pressures necessary to detect <u>overall</u> differences in natural selection between protein structures using ROC-SEMPPR. Using all nuclear protein-coding genes (approximately 6,000) genes in *S. cerevisiae* and a subset of the *E. coli* protein-coding genes (approximately 1,700), many of our regression slopes were not significantly different from 1, suggesting that the overall strength and direction of natural selection does not vary substantially between protein structures. We note that our selection estimates of $\Delta\eta$ represent sN_e for a gene with average expression. This means that differences in the strength of selective forces related to folding of different protein structures, such as the slowing of translation to facilitate co-translational folding, is rare, weak, or both on a gene with average expression relative to the strength of genetic drift.

We found that inclusion of IDR information with protein secondary structure information better explained intragenic codon usage patterns than either structure categorization method alone in

both species (Table 2). In *S. cerevisiae*, selection on codon usage in IDRs is approximately 9% weaker compared to structured regions, with small variations depending on protein secondary structure (Figure 2C–E). However, a subset of amino acids (alanine, glycine, histidine, lysine, proline, serine, and threonine) showed stronger selection between synonymous codons in IDRs compared to structured regions. Many of these amino acids have higher propensities for forming IDRs (e.g. proline, serine, lysine, alanine, glycine) or play common functional roles in IDRs such as serving as sites for phosphorylation (serine and threonine) (Pal Singh, 2015; Uversky, 2015). This detection of fine scale differences illustrates the value of population genetics-based models such as ROC-SEMPPR that provide codon-specific estimates of selection: although our results are overall consistent with Zhou et al. (2015), our approach detects codons in which selection actually appears stronger in IDRs. Although selective shifts between protein secondary structures better explained intragenic codon usage variation than IDRs in *E. coli* and the Deming regression comparing these regions was not significant, we still observed codon specific, selective shifts between these regions. Consistent with results in *S. cerevisiae*, selection on codons for alanine, serine, and threonine appeared to be stronger in *E. coli*. We speculate the apparent increased selection against certain codons in IDRs is due to stronger selection against missense errors for these amino acids in IDRs. Furthermore, we found that the best overall model in *S. cerevisiae* categorized IDRs separately by those regions predicted to be either helices or coils (Table 2). This was surprising given that classifying a region as a disordered helix seems self-contradictory. We speculate this could reflect stronger selection on codon usage in IDRs which form transient helices under certain conditions (Bomblies et al., 2017; Mizuguchi et al., 2014), but further work is needed.

Previous empirical and structural work concluded certain regions or structures of a protein may be more sensitive to missense errors, such as sheets or buried regions of a protein (Bloom et al., 2006; Guo et al., 2004; Zhou et al., 2009). Pechmann and Frydman (2013) concluded an increased frequency of optimal codons was due to increased selection for translation accuracy in sheets, which have a greater frequency of hydrophobic amino acids and propensity to aggregate compared to other secondary structures. Deming regression slopes were biased towards stronger selection in sheets when compared to coils and helices in both *S. cerevisiae* and *E. coli* (Figures 1), which would be consistent with enrichment of optimal codons in sheets, but these regressions

were not significant. In both species, we identify many codons for which selection against them is stronger in sheets compared to the other two secondary structures, although we find many codons for which the opposite is true. Many of these codons with significant changes in selection between protein secondary structures coded for hydrophobic amino acids. Previous work by Zhou et al. (2009) concluded that selection on codon usage varied significantly between the buried residues and exposed residues, following the logic that missense errors would be more deleterious in the former compared to the latter. Buried residues in a protein are likely to be more hydrophobic, so our results may partially reflect differences in natural selection between buried and exposed residues. This is also true of the structured regions and IDR comparison due to the the hydrophilic nature of IDRs, meaning they likely fall into more exposed sites in the folded state of the protein.

A major question regarding the role of codon usage in protein folding are the relative contributions of selection for inefficient translation to allow for co-translational folding and selection against missense errors. Based on estimates of absolute elongation rates and absolute missense error rates from a mechanistic model of translation (Shah and Gilchrist, 2010), we find that the magnitude (i.e. absolute value) of the selective shifts between protein structures is positively correlated with missense error rates and negatively correlated with elongation rates in *E. coli* (Figure 3). This suggests that larger selective shifts between protein structures were for the more inaccurate and slower codons. Notably, using partial correlation coefficients to control for the effects of elongation rates when comparing revealed missense error rates remained positively correlated with selective shifts between sheets and coils and structured regions and IDRs. This suggesting selection against missense errors plays a role in the observed selective shifts independent of elongation rates. Sheets and structured regions are hypothesized to be more sensitive to missense errors than their counterparts, consistent with the results observed here (Guo et al., 2004; Homma et al., 2016; Pechmann and Frydman, 2013). In contrast, *S. cerevisiae* showed no clear relationship between selective shifts and model estimates of elongation rates or missense error rates (Figure S2). However, empirical estimates of codon waiting times (i.e. inverse of elongation rates) from ribosome profiling experiments (Weinberg et al., 2016) showed a positive correlation with selection shifts, indicating larger selective shifts for slower codons. Taken together, results indicate a possible role for both elongation rates and missense errors in shaping codon usage patterns of protein structures in *E. coi*, but only find a

significant role for elongation rates in *S. cerevisiae*. Previous work found missense error rates were significantly lower in *S. cerevisiae* compared to *E. coli* (Kramer et al., 2010). Although our results in *E. coli* are consistent with expectations, the missense error rate estimates should be taken with caution given that they are based on a mechanistic model rather than an empirical measurement. Improved high-throughput, proteome-wide estimates of missense error rates could help elucidate the relationship between missense errors, codon usage, and protein structure (Mordret et al., 2019).

Previous work concluded selection on codon usage varies at the termini of secondary structures in *S. cerevisiae* and *E. coli* (Pechmann and Frydman, 2013; Saunders and Deane, 2010). Helical structures are able to begin folding within the ribosome tunnel, a common justification for any significant differences in codon usage for helices (Pechmann and Frydman, 2013; Zhou et al., 2015). For *S. cerevisiae*, we find no support for differences in natural selection on codon usage at the termini of sheets and coils, but find some support for differences in natural selection at the termini of helical structures. However, detection of these differences was sensitive to the minimum length of the helices included in the dataset. In contrast, we find some evidence for differences at the termini of both helices and sheets in *E. coli*, with the former being robust to the minimum length. We note that similar model comparisons for coils generally favored no differences between the core and termini, but most ΔDIC values were less than 10, indicating that we should not disregard the possibility of differences in selection at the termini of coils in *E. coli*. Unlike Pechmann and Frydman (2013), we found no support for differences at the second and third positions of helices in *S. cerevisiae*. Using simulated data in which selection did not vary by position in helical structures, we obtained results consistent with the findings of Pechmann and Frydman (2013), suggesting these specific results were likely an artifact due to amino acid biases and/or gene expression. As previously mentioned, metrics based on tRNA abundances sometimes reflect absolute differences across all codons rather than within a synonymous family of codons, making them particularly susceptible to amino acid biases (Chaney and Clark, 2015).

Ultimately, we find that methods rooted in population genetics can be used to test for shifts in natural selection on codon usage. As the parameter estimates of ROC-SEMPPR are clearly rooted in evolutionary processes, codon-specific parameter estimates are clearly interpretable in the context of natural selection and mutation bias. ROC-SEMPPR also provides estimates of selection for

individual codons, unlike other approaches based on heuristic measures of codon adaptation, such as CAI. Although *ad-hoc* approaches can be useful for analyzing codon usage patterns, a clear challenge is the development of an appropriate evolutionary null (Koonin, 2016). An evolutionary null is often meant as the expected patterns if a trait were evolving exclusively in a neutral manner (i.e. genetic drift), but in our case, it refers to the patterns expected if the strength and direction of natural selection on codon usage were the same between two regions. For example, Cope et al. (2018) found the increased frequency of inefficient codons in signal peptides relative to 5'-ends of non-secreted proteins in *E. coli* was due to signal peptide-specific amino acid biases and differences in average gene expression between the two classes of proteins, rather than the adaptationist hypotheses that the increased frequency of inefficient codons was selectively-driven to improve efficacy and efficiency of protein secretion (Zalucki et al., 2009). At the very least, population genetics approaches such as ROC-SEMPPR can be used for generating evolutionary null distributions for hypothesis testing with *ad-hoc* approaches. For example, genomes simulated with ROC-SEMPPR could be used to create an evolutionary null distribution of odds ratios. As shown here (Figure S6), this distribution may be quite different from the null distribution assumed in many standard hypothesis tests. Additionally, the selectively-favored codons identified by ROC-SEMPPR may be different from those identified by other methods which fail to account for the effects of mutational biases. The optimal codons for each amino acid identified by ROC-SEMPPR and the nTE metric (Pechmann and Frydman, 2013) are in agreement for 11 of the 19 amino acids (Figure S8). For the remaining 8 amino acids, a common pattern emerges: the selectively-favored codon is also mutationally-favored based on ROC-SEMPPR's parameter estimates. This suggests that the nTE metric misidentifies the selectively-favored codon by not separating out the effects of mutational biases and natural selection.

Our work demonstrates that ROC-SEMPPR and similar population genetics-based approaches are powerful tools for the analysis of codon usage patterns, providing clear estimates of natural selection and mutation bias across codons. A key advantage of ROC-SEMPPR is it can be applied to any organism with a sequenced genome, requiring no other input, such as empirical estimates of gene expression (Landerer et al., 2018; Gilchrist et al., 2015). However, ROC-SEMPPR is not the ultimate solution to this problem. Like all models, biological realism is sacrificed for the sake of

tractability. ROC-SEMPPR does not explicitly model all of the potential selective forces shaping codon usage, instead aggregating into a single parameter for each codon. For example, although we attempt to tie codon-specific selective shifts to elongation and missense error rates, our current estimates of selection do not separate out selection for translation efficiency from selection against missense errors. It is often assumed the most efficient codon is also the most accurate, but this is not necessarily the case (Shah and Gilchrist, 2010). Further work is needed to distinguish between these two evolutionary forces in order to better understand how they shape codon usage patterns related to protein structure and folding. For example, both Zhou et al. (2015) and Homma et al. (2016) found IDRs were enriched in non-optimal codons, but the former argued this was due to selection for co-translational folding of upstream domains while the latter argued this was due to reduced selection against missense errors. A possible solution for separating the effects of selection for translation efficiency and accuracy is to take a comparative approach similar to the Akashi test (Akashi, 1994) in which codon usage patterns are compared at evolutionarily conserved and variable sites. If differences in selection on codon usage between protein structures are driven primarily by selection against missense errors, then we expect that conserved vs variable sites will show different patterns between protein structures. If the differences are primarily due to differences in selection for translation efficiency, we do not expect to see a difference between evolutionarily conserved and variable sites.

Other notable exceptions excluded from ROC-SEMPPR include the effects of selection against nonsense errors (Qin et al., 2004; Gilchrist et al., 2009), GC-biased gene conversion, and hitchhiking effects. We note that GC-biased gene conversion has previously shown to be present in yeast, but its impact is relatively small (Lesecque et al., 2013) and the effects of hitchhiking on codon usage in yeasts have been somewhat controversial (Kliman et al., 2003; Qin et al., 2004; Pál et al., 2001; Zhou et al., 2005). Fortunately, the framework of models like ROC-SEMPPR allows for relatively easy incorporation of parameters reflecting these evolutionary processes. Future work investigating the effects of differences in natural selection on codon usage related to protein folding, protein secretion, and other processes will benefit from use of such models.

Data Availability

The *S. cerevisiae* CDS file and protein sequences can be downloaded from NCBI Refseq (GCF_000146045.2). The *E. coli* CDS file and protein sequences can be downloaded from NCBI Refseq (GCF_000005845.2). Empirical secondary structure information can be downloaded directly from PDB. Code for analysis and generating figures available on `https://github.com/acope3/CUB_Protein_Structure_Analysis`.

Acknowledgments

The authors would like to thank John Favate for his helpful comments in constructing this manuscript. A.L. Cope would like to acknowledge current financial support at Rutgers University (NIH: R35-GM124976, NSF:DBI-1936046, P.I.: P.R. Shah). A.L. Cope would like to acknowledge Genome Science and Technology (University of Tennessee), the National Institute for Mathematical and Biological Synthesis (NSF:DBI-1300426) and the Dept. of Eco/Evol. Biology (University of Tennessee) for previous financial support. The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

References

Alexei A. Adzhubei, Ivan A. Adzhubeib, Igor A. Krasheninnikov, and Stephen Neidle. Non-random usage of 'degenerate' codons is related to protein three-dimensional structure. *FEBS Lett.*, 399(1-2):78–82, dec 1996. ISSN 00145793. doi: 10.1016/S0014-5793(96)01287-2. URL `http://doi.wiley.com/10.1016/S0014-5793{\%}2896{\%}2901287-2`.

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

Walter Basile, Marco Salvatore, Claudio Bassot, and Arne Elofsson. Why do eukaryotic proteins contain more intrinsically disordered regions? *PLoS Comput. Biol.*, 15(7), jul 2019. ISSN

15537358. doi: 10.1371/journal.pcbi.1007186. URL /pmc/articles/PMC6675126/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6675126/.

Jesse D. Bloom, D. Allan Drummond, Frances H. Arnold, and Claus O. Wilke. Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.*, 23(9):1751–1761, sep 2006. ISSN 07374038. doi: 10.1093/molbev/msl040.

Rainer Bomblies, Manuel Patrick Luitz, Sandra Scanu, Tobias Madl, and Martin Zacharias. Transient helicity in intrinsically disordered Axin-1 studied by NMR spectroscopy and molecular dynamics simulations. *PLoS One*, 12(3):e0174337, mar 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0174337. URL https://dx.plos.org/10.1371/journal.pone.0174337.

Søren Brunak and Jacob Engelbrecht. Protein structure and the sequential structure of mRNA: α-Helix and β-sheet signals at the nucleotide level. *Proteins Struct. Funct. Bioinforma.*, 25(2):237–252, jun 1996. ISSN 1097-0134. doi: 10.1002/(SICI)1097-0134(199606)25:2⟨237::AID-PROT9⟩3.0.CO;2-E.

Daniel W A Buchan and David T Jones. The PSIPRED Protein Analysis Workbench: 20 years on. *Nucleic Acids Res.*, 47(Web Server issue), 2019. doi: 10.1093/nar/gkz297. URL https://academic.oup.com/nar/article-abstract/47/W1/W402/5480136.

Florian Buhr, Sujata Jha, Michael Thommen, Joerg Mittelstaet, Felicitas Kutz, Harald Schwalbe, Marina V. Rodnina, and Anton A. Komar. Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Mol. Cell*, 61(3):341–351, feb 2016. ISSN 1097-2765. doi: 10.1016/J.MOLCEL.2016.01.008. URL https://www.sciencedirect.com/science/article/pii/S1097276516000095?via{\%}3Dihub.

M Bulmer. The selection-mutation-drift theory of synonymous codon usage. *Genetics*, 129(3), 1991.

Kenneth P. Burnham and David R. Anderson. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociol. Methods Res.*, 33(2):261–304, nov 2004. ISSN 0049-1241. doi: 10.1177/0049124104268644. URL http://journals.sagepub.com/doi/10.1177/0049124104268644.

J L Chaney and P L Clark. Roles for Synonymous Codon Usage in Protein Biogenesis. *Annu. Rev. Biophys.*, 44:143–166, 2015. doi: 10.1146/annurev-biophys-060414-034333.

Julie L. Chaney, Aaron Steele, Rory Carmichael, Anabel Rodriguez, Alicia T. Specht, Kim Ngo, Jun Li, Scott Emrich, and Patricia L. Clark. Widespread position-specific conservation of synonymous rare codons within coding sequences. *PLOS Comput. Biol.*, 13(5):e1005531, may 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005531. URL `http://dx.plos.org/10.1371/journal.pcbi.1005531`.

A.L. Cope, R.L. Hettich, and M.A. Gilchrist. Quantifying codon usage in signal peptides: Gene expression and amino acid usage explain apparent selection for inefficient codons. *Biochim. Biophys. Acta - Biomembr.*, 1860(12), 2018. ISSN 18792642. doi: 10.1016/j.bbamem.2018.09.010.

M dos Reis, L Wernisch, and R Savva. Unexpected correlations between gene expression and codon usage bias from microarray data for the whole \textit{Escherichia coli} K-12 genome. *Nucleic Acids Res.*, 31(23):6976–6985, 2003.

M dos Reis, R Savva, and L Wernisch. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.*, 32(17):5036–5044, 2004.

D A Drummond and C O Wilke. Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution. *Cell*, 134:341–352, 2008.

A Fluitt, E Pienaar, and H Viljoen. Ribosome Kinetics and aa-tRNA competition determine rate and fidelity of peptide synthesis. *Comput. Biol. Chem.*, 31:335–346, 2007.

Jingjing Fu, Katherine A. Murphy, Mian Zhou, Ying H. Li, Vu H. Lam, Christine A. Tabuloc, Joanna C. Chiu, and Yi Liu. Codon usage affects the structure and function of the Drosophila circadian clock protein PERIOD. *Genes Dev.*, 30(15):1761–1775, aug 2016. ISSN 15495477. doi: 10.1101/gad.281030.116.

Kerry A. Geiler-Samerotte, Michael F. Dion, Bogdan A. Budnik, Stephanie M. Wang, Daniel L. Hartl, and D. Allan Drummond. Misfolded proteins impose a dosage-dependent fitness cost and

trigger a cytosolic unfolded protein response in yeast. *Proc. Natl. Acad. Sci. U. S. A.*, 108(2): 680–685, jan 2011. ISSN 00278424. doi: 10.1073/pnas.1017570108.

Tali Gidalevitz, Thomas Krupinski, Susana Garcia, and Richard I. Morimoto. Destabilizing Protein Polymorphisms in the Genetic Background Direct Phenotypic Expression of Mutant SOD1 Toxicity. *PLoS Genet.*, 5(3):e1000399, mar 2009. ISSN 1553-7404. doi: 10.1371/journal.pgen.1000399. URL `https://dx.plos.org/10.1371/journal.pgen.1000399`.

M A Gilchrist. Combining Models of Protein Translation and Population Genetics to Predict Protein Production Rates from Codon Usage Patterns. *Mol. Biol. Evol.*, 24(11):2362–2372, 2007. doi: 10.1093/molbev/msm169.

M A Gilchrist and A Wagner. A model of protein translation including codon bias, nonsense errors, and ribosome recylcing. *J. Theor. Biol.*, 239:417–434, 2006.

M A Gilchrist, P Shah, and R Zaretzki. Measuring and Detecting Molecular Adaptation in Codon Usage Against Nonsense Errors During Protein Translation. *Genetics*, 183:1493–1505, 2009. doi: 10.1534/genetics.109.108209.

M A Gilchrist, W C Chen, P Shah, C L Landerer, and R Zaretzki. Estimating Gene Expression and Codon-Specific Translational Efficiencies, Mutation Biases, and Selection Coefficients from Genomic Data Alone. *Genome Biol. Evol.*, 7:1559–1579, 2015.

S J Gould and R C Lewontin. The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme. *Proc. R. Soc. London*, 205(1161):581–598, 1979.

Haiwei H. Guo, Juno Choe, and Lawrence A. Loeb. Protein tolerance to random amino acid change. *Proc. Natl. Acad. Sci. U. S. A.*, 101(25):9205–9210, jun 2004. ISSN 00278424. doi: 10.1073/pnas.0403255101.

S. K. Gupta, S. Majumdar, T. K. Bhattacharya, and T. C. Ghosh. Studies on the relationships between the synonymous codon usage and protein secondary structural units. *Biochem. Biophys. Res. Commun.*, 269(3):692–696, mar 2000. ISSN 0006291X. doi: 10.1006/bbrc.2000.2351.

W Holtkamp, G Kokie, M Jager, Joerg Mittelstaet, A A Komar, and M V Rodnina. Cotranslational protein folding on the ribosome monitored in real time. *Science (80-.).*, 350(6264):1104–1107, 2015.

Keiichi Homma, Tamotsu Noguchi, and Satoshi Fukuchi. Codon usage is less optimized in eukaryotic gene segments encoding intrinsically disordered regions than in those encoding structural domains. *Nucleic Acids Res.*, page gkw899, oct 2016. ISSN 0305-1048. doi: 10.1093/nar/gkw899. URL `https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw899`.

T Ikemura. 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.

William M. Jacobs and Eugene I. Shakhnovich. Evidence of evolutionary selection for cotranslational folding. *Proc. Natl. Acad. Sci. U. S. A.*, 114(43):11434–11439, oct 2017. ISSN 10916490. doi: 10.1073/pnas.1705772114.

David T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, 292(2):195–202, sep 1999. ISSN 00222836. doi: 10.1006/jmbi.1999.3091.

Chava Kimchi-Sarfaty, Jung Mi Oh, In Wha Kim, Zuben E. Sauna, Anna Maria Calcagno, Suresh V. Ambudkar, and Michael M. Gottesman. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science (80-.).*, 315(5811):525–528, jan 2007. ISSN 00368075. doi: 10.1126/science.1135308.

Richard M. Kliman, Naheelah Irving, and Maria Santiago. Selection conflicts, gene expression, and codon usage trends in yeast. *J. Mol. Evol.*, 57(1):98–109, jul 2003. ISSN 00222844. doi: 10.1007/s00239-003-2459-9. URL `https://link.springer.com/article/10.1007/s00239-003-2459-9`.

Eugene V. Koonin. Splendor and misery of adaptation, or the importance of neutral null for understanding evolution. *BMC Biol.*, 14(1):114, dec 2016. ISSN 17417007. doi:

10.1186/s12915-016-0338-2. URL http://bmcbiol.biomedcentral.com/articles/10.1186/s12915-016-0338-2.

Emily B. Kramer, Haritha Vallabhaneni, Lauren M. Mayer, and Philip J. Farabaugh. A comprehensive analysis of translational missense errors in the yeast Saccharomyces cerevisiae. *RNA*, 16 (9):1797–1808, 2010. ISSN 14699001. doi: 10.1261/rna.2201210.

C Landerer, A Cope, R Zaretzki, and M A Gilchrist. AnaCoDa: analyzing codon data with Bayesian mixture models. *Bioinformatics*, page bty138, 2018. doi: 10.1093/bioinformatics/bty138. URL http://dx.doi.org/10.1093/bioinformatics/bty138.

Cedric Landerer, Brian C. O'Meara, Russell Zaretzki, and Michael A. Gilchrist. Unlocking a signal of introgression from codons in Lachancea kluyveri using a mutation-selection model. *BMC Evol. Biol.*, 20(1):109, aug 2020. ISSN 14712148. doi: 10.1186/s12862-020-01649-w. URL https://bmcevolbiol.biomedcentral.com/articles/10.1186/s12862-020-01649-w.

David S. Lawrie, Philipp W. Messer, Ruth Hershberg, and Dmitri A. Petrov. Strong Purifying Selection at Synonymous Sites in D. melanogaster. *PLoS Genet.*, 9(5):e1003527, may 2013. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003527. URL https://dx.plos.org/10.1371/journal.pgen.1003527.

Yann Lesecque, Dominique Mouchiroud, and Laurent Duret. GC-biased gene conversion in yeast is specifically associated with crossovers: Molecular mechanisms and evolutionary significance. *Mol. Biol. Evol.*, 30(6):1409–1419, jun 2013. ISSN 07374038. doi: 10.1093/molbev/mst056. URL https://academic.oup.com/mbe/article/30/6/1409/1145455.

Michael Lynch. The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc. Natl. Acad. Sci. U. S. A.*, 104(SUPPL. 1):8597–8604, may 2007. ISSN 00278424. doi: 10.1073/pnas.0702207104. URL www.nasonline.org/adaptationandcomplexdesign.

Heather E. Machado, David S. Lawrie, and Dmitri A. Petrov. Pervasive Strong Selection at the Level of Codon Usage Bias in Drosophila melanogaster. *Genetics*, 214(2):511–528, feb 2020. ISSN 19432631. doi: 10.1534/genetics.119.302542.

Bálint Mészáros, Gábor Erdős, and Zsuzsanna Dosztányi. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.*, 46(W1): W329–W337, jul 2018. ISSN 0305-1048. doi: 10.1093/nar/gky384. URL `https://academic.oup.com/nar/article/46/W1/W329/5026265`.

Mineyuki Mizuguchi, Takahiro Fuju, Takayuki Obita, Mitsuru Ishikawa, Masaaki Tsuda, and Akiko Tabuchi. Transient α-helices in the disordered RPEL motifs of the serum response factor coactivator MKL1. *Sci. Rep.*, 4(1):1–6, jun 2014. ISSN 20452322. doi: 10.1038/srep05224. URL `www.nature.com/scientificreports`.

Ernest Mordret, Orna Dahan, Omer Asraf, Roni Rak, Avia Yehonadav, Georgina D. Barnabas, Jürgen Cox, Tamar Geiger, Ariel B. Lindner, and Yitzhak Pilpel. Systematic Detection of Amino Acid Substitutions in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity. *Mol. Cell*, 75(3):427–441.e5, aug 2019. ISSN 10972765. doi: 10.1016/j.molcel.2019.06.041.

Thijs Nieuwkoop, Max Finger-Bou, John van der Oost, and Nico J. Claassens. The Ongoing Quest to Crack the Genetic Code for Protein Production. *Mol. Cell*, 80(2):193–209, oct 2020. ISSN 10972765. doi: 10.1016/j.molcel.2020.09.014. URL `https://doi.org/10.1016/j.molcel.2020.09.014`.

Edward P. O'Brien, Prajwal Ciryam, Michele Vendruscolo, and Christopher M. Dobson. Understanding the influence of codon translation rates on cotranslational protein folding. *Acc. Chem. Res.*, 47(5):1536–1544, may 2014. ISSN 15204898. doi: 10.1021/ar5000117.

Csaba Pál, Balázs Papp, and Laurence D. Hurst. Does the Recombination Rate Affect the Efficiency of Purifying Selection? The Yeast Genome Provides a Partial Answer. *Mol. Biol. Evol.*, 18 (12):2323–2326, dec 2001. ISSN 1537-1719. doi: 10.1093/oxfordjournals.molbev.a003779. URL `http://academic.oup.com/mbe/article/18/12/2323/1074380`.

Gajinder Pal Singh. Association between intrinsic disorder and serine/threonine phosphorylation in Mycobacterium tuberculosis. *PeerJ*, 2015(1), 2015. ISSN 21678359. doi: 10.7717/peerj.724.

S Pechmann and J Frydman. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.*, 20(2):237–243, 2013.

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

H Qin, W B Wu, J M Comeron M Kreitman, and W Li. Intragenic Spatial Patterns of Codon Usage Bias in Prokaryotic and Eukaryotic Genomes. *Genetics*, 168:2245–2260, 2004.

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

P Shah and M Gilchrist. Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *PNAS*, 108(25):10231–10236, 2011.

P Shah and M A Gilchrist. Effect of Correlated tRNA Abundances on Translation Errors and Evolution of Codon Usage Bias. *PLoS Genet.*, 6(9):1–9, 2010.

Vijaya L. Simhadri, Nobuko Hamasaki-Katagiri, Brian C. Lin, Ryan Hunt, Sujata Jha, Sandra C. Tseng, Andrew Wu, Amber A. Bentley, Ran Zichel, Qi Lu, Lily Zhu, Darón I. Freedberg, Dougald M. Monroe, Zuben E. Sauna, Robert Peters, Anton A. Komar, and Chava Kimchi-Sarfaty. Single synonymous mutation in factor IX alters protein properties and underlies haemophilia B. *J. Med. Genet.*, 54(5):338–345, may 2017. ISSN 14686244. doi: 10.1136/jmedgenet-2016-104072. URL `http://www.nupack.`

R R Sokal and F J Rohlf. *Biometry - The Principles and Practices of Statistics in Biological Research.* W.H. Freeman, New York, 3rd edition, 1995.

David J. Spiegelhalter, Nicola G. Best, Bradley P. Carlin, and Angelika Van Der Linde. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(4):583–616, oct 2002. ISSN 13697412. doi: 10.1111/1467-9868.00353. URL `https://rss.onlinelibrary.wiley.com/doi/full/10.1111/1467-9868.00353https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00353https://rss.onlinelibrary.wiley.com/doi/10.1111/1467-9868.00353`.

Xie Tao and Ding Dafu. The relationship between synonymous codon usage and protein structure. *FEBS Lett.*, 434(1-2):93–96, aug 1998. ISSN 00145793. doi: 10.1016/S0014-5793(98)00955-7. URL `http://doi.wiley.com/10.1016/S0014-5793{\%}2898{\%}2900955-7`.

T. A. Thanaraj and Patrick Argos. Protein secondary structural types are differentially coded on messenger RNA. *Protein Sci.*, 5(10):1973–1983, oct 1996. ISSN 09618368. doi: 10.1002/pro.5560051003. URL `http://doi.wiley.com/10.1002/pro.5560051003`.

Vladimir N. Uversky. The intrinsic disorder alphabet. III. Dual personality of serine. *Intrinsically Disord. Proteins*, 3(1), 2015. ISSN 21690707. doi: 10.1080/21690707.2015.1027032. URL `/pmc/articles/PMC5314895/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5314895/`.

E W J Wallace, E M Airoldi, and D A Drummond. Estimating Selection on Synonymus Codon Usage from Noisy Experimental Data. *Mol. Biol. Evol.*, 30(6):1438–1453, 2013.

Ian M. Walsh, Micayla A. Bowman, Iker F. Soto Santarriaga, Anabel Rodriguez, and Patricia L. Clark. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. *Proc. Natl. Acad. Sci. U. S. A.*, 117(7):3528–3534, feb 2020. ISSN 10916490. doi: 10.1073/pnas.1907126117.

David E. Weinberg, Premal Shah, Stephen W. Eichhorn, Jeffrey A. Hussmann, Joshua B. Plotkin, and David P. Bartel. Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation. *Cell Rep.*, 14(7):1787–1799, feb 2016. ISSN 22111247. doi: 10.1016/j.celrep.2016.01.043.

C Yu, Y Dang, Z Zhou, C Wu, F Zhao, M S Sachs, and Y Liu. Codon Usage Influences the Local Rate of Translation Elongation to Regulate Co-translational Protein Folding. *Mol. Cell*, 59: 744–754, 2015.

Y M Zalucki, I R Beacham, and M P Jennings. Biased codon usage in signal peptides: a role in protein export. *Trends Microbiol.*, 17(4):146–150, 2009.

Fangzhou Zhao, Chien Hung Yu, and Yi Liu. Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells. *Nucleic Acids Res.*, 45(14):8484–8492, 2017. ISSN 13624962. doi: 10.1093/nar/gkx501.

Mian Zhou, Tao Wang, Jingjing Fu, Guanghua Xiao, and Yi Liu. Nonoptimal codon usage influences protein structure in intrinsically disordered regions. *Mol. Microbiol.*, 97(5):974–987, sep 2015. ISSN 13652958. doi: 10.1111/mmi.13079.

T. Zhou, Z. H. Lu, and X. Sun. The correlation between recombination rate and codon bias in yeast mainly results from mutational bias associated with recombination rather than hill-robertson interference. In *Annu. Int. Conf. IEEE Eng. Med. Biol. - Proc.*, volume 7 VOLS, pages 4787–4790. Institute of Electrical and Electronics Engineers Inc., 2005. ISBN 0780387406. doi: 10.1109/iembs.2005.1615542.

T. Zhou, M. Weems, and C. O. Wilke. Translationally Optimal Codons Associate with Structurally Sensitive Sites in Proteins. *Mol. Biol. Evol.*, 26(7):1571–1580, jul 2009. ISSN 0737-4038. doi: 10.1093/molbev/msp070. URL `https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msp070`.