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]Department of Genetics, Rutgers University

[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

Last compiled on Friday 18th December, 2020 at 17:49.

Abstract

Empirical studies indicate changes to synonymous codon usage can impact protein folding. However, genome-wide computational analyses attempting to establish a more general relationship between codon usage and protein structure have often led to contradictory results. Using a population genetics model, we quantified codon-specific shifts in natural selection across and within protein structures, revealing a complex relationship between codon usage and protein structure. However, these shifts are small, suggesting differences in selection related to protein structure are either very weak or apply to relatively few codons. Using a

previously published result, we demonstrate how tests for selection on codon usage can be confounded when failing to account for amino acid biases and gene expression. This work demonstrates the value of using population genetics-based models to quantify and tests for shifts in selection on codon usage. Extensions to this approach are discussed.

Introduction

Patterns of codon usage bias (CUB), or the non-uniform usage of synonymous codons, vary both within and across species (Drummond and Wilke, 2009; Hershberg and Petrov, 2008; Plotkin and Kudla, 2011; Labella et al., 2019). 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, 1982) and stronger codon bias in highly expressed genes (Bulmer, 1990; Gilchrist et al., 2015; Shah and Gilchrist, 2011) suggests selection against translation inefficiency is a major factor shaping genome-wide codon patterns. Importantly, codon usage bias also varies within a gene (Clarke IV and Clark, 2008; Gilchrist and Wagner, 2006; Gilchrist et al., 2009), which is hypothesized to reflect various other forces of selection. For example, CUB is thought to be shaped by selection for translation accuracy, including to reduce missense errors at functionally-important sites and nonsense errors later in transcripts, both of which can result in non-functional proteins (Akashi, 1994; Drummond and Wilke, 2008; Kurland, 1992; Gilchrist and Wagner, 2006; Gilchrist, 2007; Gilchrist et al., 2009; Stoletzki and Eyre-Walker, 2007). Intragenic variation in synonymous codon usage has also been proposed to be shaped by selection to prevent ribosomal queuing (Tuller et al., 2011), selection against mRNA secondary structures near the 5'-end to allow for efficient translation initiation (Kudla et al., 2009; Hockenberry et al., 2014), efficient protein secretion (Liu et al., 2017; Mahlab and Linial, 2014; Power et al., 2004), efficient transmembrane insertion (Pechmann et al., 2014; Fluman et al., 2014), and accurate co-translational folding

(Chaney et al., 2017; Pechmann and Frydman, 2013).

Although adaptive codon usage bias is generally 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 evolutionary 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). For example, Cope et al. (2018) found the increased frequency of inefficient codons in signal peptides relative to 5'-ends on 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).

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 are unable to perform their function, can aggregate within the cell, and possibly disrupt key cellular processes (Bucciantini et al., 2002; Drummond and Wilke, 2008; Feyertag et al., 2017; Geiler-Samerotte et al., 2011; Gidalevitz et al., 2009; Yang et al., 2012). For example, 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, 2009). In addition, codon usage is thought to modulate

protein folding via changes in the elongation rates at key steps during the co-translational folding of the protein (Ciryam et al., 2013; O'Brien et al., 2014; Jacobs and Shakhnovich, 2017). Various lines of empirical work indicate changes in elongation rates (modulated via synonymous codon usage) alter co-translational protein folding (Buhr et al., 2016; Fu et al., 2016; Holtkamp et al., 2015; Kimchi-Sarfaty et al., 2007; Komar et al., 1999; Krasheninnikov et al., 1991; Walsh et al., 2020; Yu et al., 2015; Zhou et al., 2013, 2015).

Given that protein secondary structures generally differ in their physicochemical properties and ability to co-translationally fold, researchers have hypothesized that different structures will exhibit different codon usage patterns (Chaney and Clark, 2015; Pechmann and Frydman, 2013; Saunders and Deane, 2010). Despite numerous studies examining the relationship between codon usage and protein secondary structure, 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, 1996b; Chaney et al., 2017). Even at higher levels of proteins structure, analysis of codon usage patterns have been inconsistent. For example, some studies concluded that "non-optimal" codons are more frequent in IDRs (Homma et al., 2016; Zhou et al., 2015), while others found that clusters of inefficient codons occur more frequently in structured regions, often downstream from sub-domains (Chaney et al., 2017; Jacobs and Shakhnovich, 2017). Biological interpretations are also often inconsistent between studies. For example, Homma et al. (2016) argued the increased frequency of "non-optimal" codons was due to reduced selection against missense errors in IDRs. In contrast, Zhou et al. (2015) argued the increased frequency of "non-optimal" codons was due to increased selection for slow translation in IDRs to assist in the co-translational folding of upstream structured domains.

A likely cause for inconsistent results between studies are the differences in metrics for quantifying codon usage (see Chaney and Clark (2015) for a summary of many of the metrics). For example, Saunders and Deane (2010) found that *E. coli* secondary structures exhibited different codon usage patterns at the termini of secondary structures, but these

patterns only held if using tRNA abundances rather than the Codon Adaptation Index (CAI) (Sharp and Li, 1987) or MinMax% (Clarke IV and Clark, 2008). 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, such as glutamine in *S. cerevisiae* (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; Zhou et al., 2009). Although other studies have recognized the potential issues of these confounding factors and attempted to control for them, these approaches were somewhat *ad-hoc*. For example, Pechmann and Frydman (2013) attempted to control for gene expression by examining codon usage patterns in subsets of high and low expression genes.

Although some work has attempted a comparative approach to examining codon usage as it relates to protein folding, appropriately recognizing that selection on codon usage related to protein folding would lead to conserved patterns across species (Chaney et al., 2017; Pechmann and Frydman, 2013), this work does not actually model 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; Cope et al., 2018; Gilchrist, 2007; Gilchrist

et al., 2009; Shah and Gilchrist, 2011; Gilchrist et al., 2015; Landerer et al., 2020; 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.

Although originally developed for estimating selection and mutation biases based on genome-wide codon frequencies, recent work 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, we tested for differences in selection on codon usage across related to protein secondary structures and IDRs in *S. cerevisiae*. Although model comparisons indicate selection on codon usage differs across protein structures, 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 no evidence for general differences in selection on codon usage at the termini of sheets and coils, but do find some evidence suggesting natural selection differs at the termini of helical structures. We also present evidence previous conclusions regarding enrichment of slow translating codons near the start of helices was biased by amino acid usage and/or gene expression. 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

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). We used both empirical and computationally-determined secondary structure predictions for comparing selection on codon usage across secondary structures, as done previously (Pechmann and Frydman, 2013). However, we used only empirically-determined secondary structures for analysis of codon usage at termini of secondary structures.

Empirically-determined protein secondary structures and corresponding protein sequences were obtained from the Protein Data Bank (PDB). Residues were grouped into four overarching structural groups based on their DSSP classification: helix (DSSP H, G, I), sheet (DSSP E,B), and coil (DSSP S, T, .). This classification system is consistent with secondary structure prediction algorithms (Jones, 1999) and other analyses of codon usage patterns using empirically-determined secondary structures (Adzhubei et al., 1996; Thanaraj and Argos, 1996a; 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 were aligned to the *S. cerevisiae* proteome using BLAST. Sequences were considered mapped 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 protein sequences with empirically-determined secondary structures. 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* proteins using the PsiPred software (Jones, 1999), resulting in 5,983 secondary structure predictions. 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 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 evolutionary 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 $N_e s$, 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.

As a first step, ROC-SEMPPR was fit to all 5,983 nuclear protein-coding sequences in *S.*

cerevisiae 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.

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 to length of the secondary structures, we varied the minimum number of amino acids for a secondary structure between 4 and 10. We also checked for robustness of our results to termini length by repeating the analysis with the termini defined as either 2 or 3 amino acids. . To test the hypothesis presented by Pechmann and Frydman (2013), helices of minimum length 6 amino acidus 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. To help distinguish between model comparisons based on structural information, we will denote all models exclusively using secondary structure information with "A," all models exclusively using structured/IDR information with "B," and all models combining the information with "C." 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.

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. As the choice of reference codons for $\Delta\eta$ can change the results of any Deming regression, we rescaled $\Delta\eta$ of each synonymous codon relative to the mean $\Delta\eta$, such that each synonymous codon family had a mean $\Delta\eta$ of 0 (Cope et al., 2018). 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 these codons show the strongest differences in natural selection between structural groupings.

Analysis of position-specific codon usage patterns in helices

Previous work found that positions 2 and 3 (relative to the start) of helices in yeast were enriched in non-optimal codons (Pechmann and Frydman, 2013). In order 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 aligned by position. For each position, enrichment of non-optimal codons (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 allowed selection on codon usage to differ between helices, sheets, and coils (Table 1, Model A1 vs. A2, ΔDIC = -68). 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, ΔDIC = -1.16).

Table 1: Comparison of model fits examining variation in codon usage between predicted protein secondary structures. Models with a DIC score greater than the null model (no difference in codon usage between secondary structures) were excluded from the table. H: helix. E: sheet. C: coil.

Structure Model	Groupings I	II	III	ΔDIC
A1	H	E	C	0
A2	HE	–	C	-68
A5	H	EC	–	-285
A3	HC	E	–	-425
A4	HEC	–	–	-621

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 $\Delta\eta$ estimates with a Deming regression revealed only a significant difference between coils and sheets (Deming regression $\hat{\beta} = 1.054$, 95% CI: 1.015 – 1.093, Figure 1B). This indicates selection on codon usage is approximately 5% greater on sheets compared to coils. However, we do not find a significant differences when comparing helices to coils (Deming regression $\hat{\beta} = 1.026$, 95% CI: 0.996 – 1.056, Figure 1A) or sheets (Deming regression $\hat{\beta} = 1.011$, 95% CI: 0.985 – 1.037, Figure 1C).

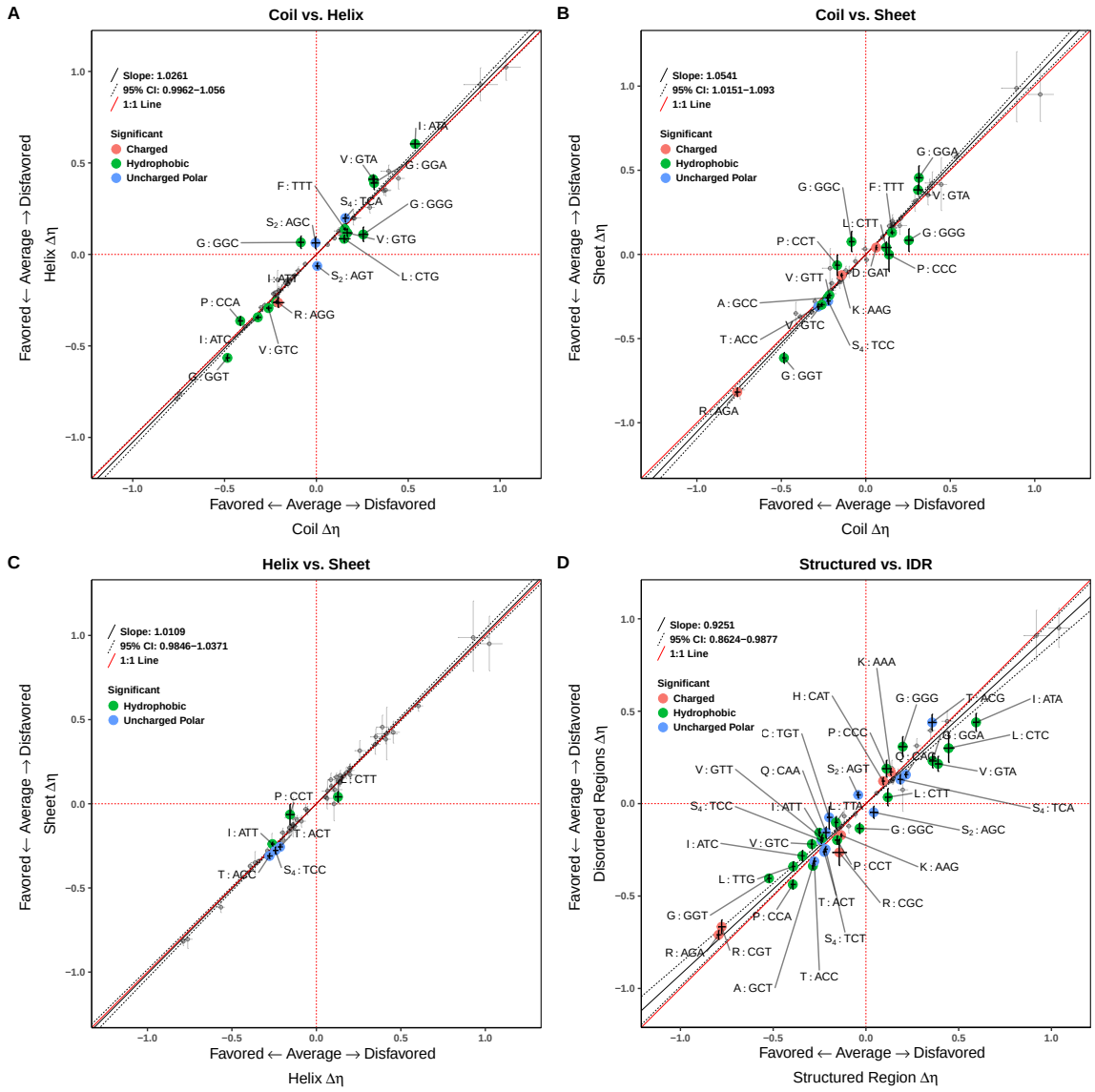

Figure 1: Comparison of mean-centered $\Delta\eta$ estimates between different protein secondary structures. Black dots represent $\Delta\eta$ values for each codon, while error bars represent the 95% posterior probability intervals. The Deming regression slope and 95% confidence intervals are represented by solid and dashed black lines, respectively. **(A)** Coil vs. Helices. **(B)** Coil vs. Sheet. **(C)** Helices vs. Sheet. **(D)** Structured regions vs. IDRs.

.

Despite 2 of the 3 Deming Regressions being non-significant, examination of the 95% posterior probability intervals for $\Delta\eta$ reveals significant differences in selection on individual codons. We find that 12 of the 18 amino acids with more than one synonymous codon have at least one codon that differs between secondary structures: alanine (A), glycine

(G), aspartic acid (D), serine (S_4 and S_2), phenylalanine (F), leucine (L), valine (V), lysine (K), arginine (R), proline (P), isoleucine (I), and threonine (T) (Figure 1A-C). Most of the significantly different codons code for non-polar, hydrophobic amino acids. Importantly, these differences generally reflect changes in the average strength of selection and not a switch in the preferred codon. We also note that most of the codon-specific differences in selection occurred between coils and the other two secondary structures. There appear to be very few differences in selection between helices and sheets, suggesting a large dataset is necessary to adequately detect these differences. This likely explains the ambiguousness of the results using empirically-determined secondary structures that represents approximately a sixth of the *S. cerevisiae* protein-coding sequences.

Intrinsically-disordered regions better explain codon usage patterns than secondary structure.

Consistent with Zhou et al. (2015), we found that whether a region was structured or disordered better explained intragenic codon usage patterns than protein secondary structures (Table 2, Models B1 vs. A1). Consistent with this, selection $\Delta\eta$ was 8% weaker, on average, in IDRs compared to structured regions (Deming Regression $\hat{\beta} = 0.925$, 95% CI: 0.862 – 0.988, Figure 1D). Although selection on codon usage was weaker, on average, in IDRs, we note a subset of amino acids demonstrate the opposite pattern in which selection appears to be stronger in the IDRs: alanine (A), lysine (K), proline (P), serine (S), and threonine (T). In particular, we note an apparent shift in the favored codon of serine between AGC and AGT, with the latter being favored in structured regions but disfavored in IDRs. These amino acids generally have a higher propensity for forming disordered regions and/or are phosphorylated in these regions (Pal Singh, 2015; Uversky, 2015).

We also found that categorizing the predicted structured regions based on the corresponding secondary structure predictions further improved the overall model fit (Model C5 vs. B1). This suggests that even though structured vs. disordered regions better explain

codon usage patterns than secondary structures, there are still differences in selection on codon usage between secondary structures after accounting for the disordered regions, contrary to the finding of Zhou et al. (2015). Comparing selection estimates $\Delta\eta$ estimated from the secondary structures (with IDRs removed) to $\Delta\eta$ estimated from the disordered regions suggests selection on codon usage is, on average, stronger in all 3 secondary structures compared to IDRs (Figure 2). When comparing $\Delta\eta$ between secondary structures after removing disordered regions, we found that the previously observed significant difference between sheets and coils disappeared (Figure S1B), although we note the previously observed significant difference was small. However, examination of the 95% posterior probability intervals revealed many codons to still be significantly different between the secondary structures after removal of the IDRs, with 9 of 18 amino acids having at least one codon that significantly differs between protein secondary structures (Figure S1). Amino acids no longer exhibiting significant differences in codon usage between secondary structures after removal of the IDRs include lysine (K), alanine (A), and arginine (R).

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 treating structured and disordered regions as separate categories are shown. 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						
Structure Model	I	II	III	IV	V	VI	ΔDIC
C1	Struct H	IDR H	E	–	Struct C	IDR C	0
C2	Struct H	IDR H	Struct E	IDR E	Struct C	IDR C	-0.46
C3	H	–	E	–	Struct C	IDR C	-83
C4	H	–	Struct E	IDR E	Struct C	IDR C	-84
C5	Struct H	–	Struct E	–	Struct C	IDR	-117
B1	Structured	–	–	–	–	IDR	-466
A1	H	–	E	–	C	–	-686

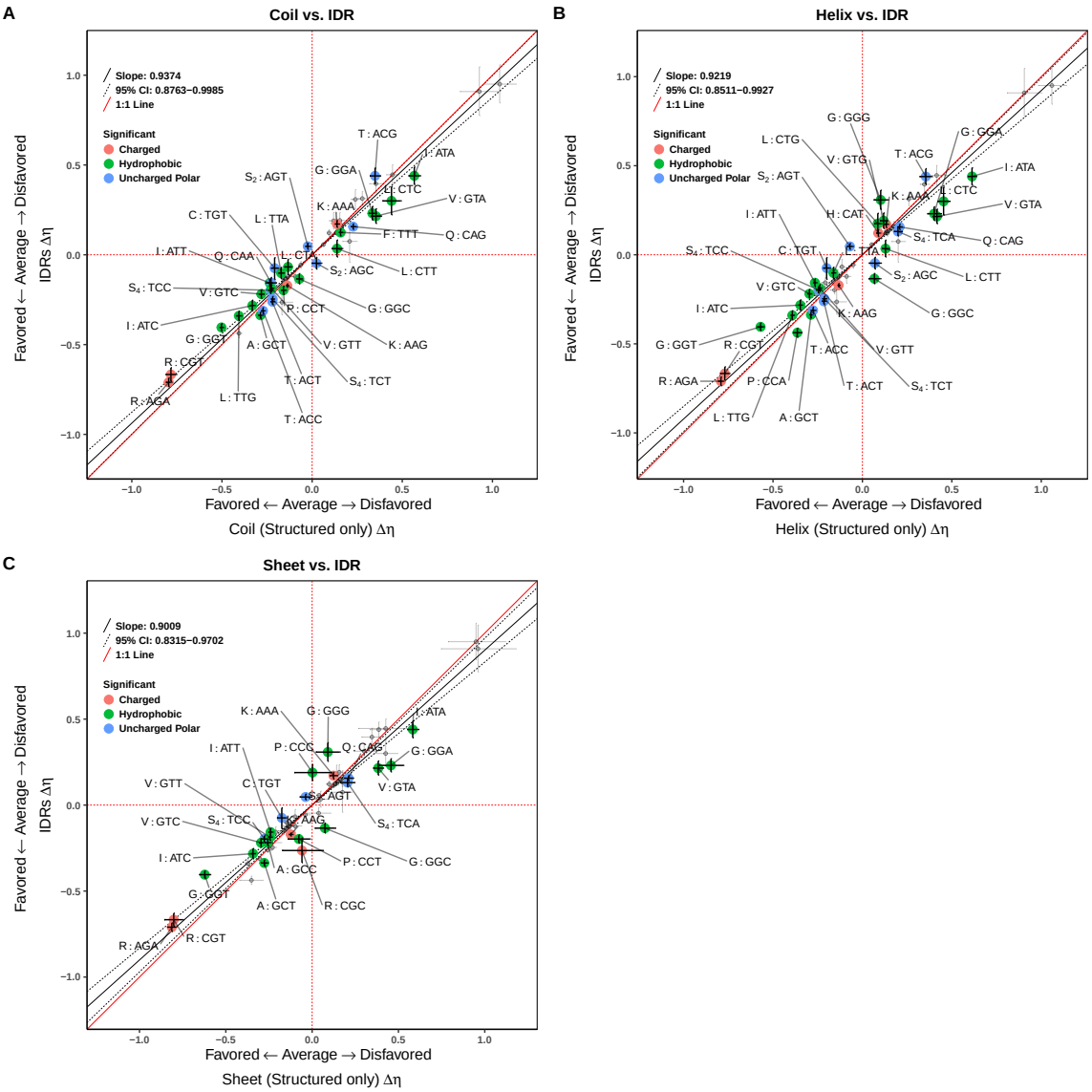

Figure 2: Comparison of selection $\Delta\eta$ from secondary structures (with intrinsically disordered regions removed) and disordered regions. In all cases, selection is on average stronger in the secondary structures relative to the disordered regions, but this effect is small.

Given that all secondary structures were predicted to fall into structured or disordered regions, we tested how further dividing up the secondary structures by whether they were classified as structured or disordered explained codon usage patterns. Unsurprisingly, most disordered regions predicted by IUPRED2 were also predicted to be coils by PsiPred (Table S13). Splitting up only the coils into structured and disordered regions improved upon the model fit where disordered regions were taken from all three secondary structure classifi-

cations (Model C3 vs. C5). Surprisingly, the best overall model fit divided both coils and helices into structured and disordered regions (Model C1). Splitting up sheets into structured and disordered regions was a slightly worse model, but essentially made no difference in the ability to explain codon usage patterns (Model C1 vs. C2).

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

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 S2 – S11). 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 S2). 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 S4 – S5). 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 S12); 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 only two codons are significantly different: proline (P) codon CCC and glycine (G) codon GGT, with the former showing a flip in preference, being more favored in the core relative to the termini (Figure

S2).

Although we found some support for differences in selection at the termini of helices, we did not find support for the specific pattern that the second and third positions of helices are enriched in inefficient codons (ΔDIC = -29.75) (Pechmann and Frydman, 2013). Using simulated data that assumes $\Delta\eta$ is the same for a given codon, regardless of location in a gene, we found the odds ratios reported by Pechmann and Frydman (2013) were within the range of odds ratios generated using the simulated data (Figure S3). This suggests the results reported by Pechmann and Frydman (2013) regarding position-specific codon usage in helices are an artifact of amino acid biases and/or gene expression.

Discussion

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. Using ROC-SEMPPR, we detected small, codon-specific shifts in selection on codon usage 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 (1023 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. translation inefficiency instead of translation efficiency (Figure S4). 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 also implicitly assumes that selection against translation inefficiency and selection against missense errors, both of which are thought to impact codon usage patterns related to protein folding in real data, are perfectly correlated,

a common assumption. However, there is evidence that the most efficient and most accurate codons are not necessarily the same (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 overall differences in natural selection between protein structures using ROC-SEMPPR. Using all nuclear protein-coding genes (approximately 6,000) genes in *S. cerevisiae*, many of our regression slopes were not significantly different from 1, suggesting that the strength and direction of natural selection does not significantly vary 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.

Consistent with Zhou et al. (2015), we found that intragenic codon usage patterns were better explained by (IDRs) than protein secondary structure (Model B1 vs. Model A1), although splitting structured regions into their corresponding secondary structures further improved the overall model fit (Model C5 vs. Model B1). Selection on codon usage in IDRs is approximately 8% weaker compared to structured regions, with small variations depending on protein secondary structure (6% weaker relative to coils, 8% weaker relative to helices, and 10% weaker relative to sheets). However, a subset of amino acids (alanine, lysine, proline, serine, and threonine) showed stronger selection between synonymous codons in IDRs compared to structured regions. As one might expect, these amino acids have higher propensities for forming IDRs or play common functional roles in IDRs (such as serving as sites for phosphorylation) (Pal Singh, 2015; Uversky, 2015). This 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. We speculate this pattern is due to stronger selection against missense errors at these amino acids

in IDRs. Furthermore, we found that the best overall model categorized IDRs separately by those regions predicted to be either helices or coils (Model C1). This was surprising given that the concept of a disordered helix seems contradictory. We speculate that this reflects 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.

Of the three major secondary structures, sheets demonstrate the clearest difference in selection on codon usage based on the Deming Regression, with selection on codon usage approximately 5% stronger relative to coils (if not excluding IDRs) and 10% stronger relative to IDRs. Previous work in *S. cerevisiae* observed similar patterns, with sheets having the highest frequency of conserved "optimal" codons and depletion of conserved "non-optimal" codons (Pechmann and Frydman, 2013). Unlike previous work, our approach identifies the specific amino acids and codons on which selection appears to differ, rather than broadly classifying codons as "optimal" or "non-optimal." Pechmann and Frydman (2013) concluded this 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. Empirical work suggests missense substitutions in sheets tend to be more destabilizing than in helices (Guo et al., 2004). Consistent with this, most of the 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 differential 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.

Previous work concluded selection on codon usage varies at the termini of secondary

structures Pechmann and Frydman (2013); Saunders and Deane (2010). Although we find no support for differences in natural selection on codon usage at the termini of sheets and coils, we find some support for differences in natural selection at the termini of helical structures. 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). Although model comparisons suggested possible differences between the core and the termini of helices, we only observed significant differences for two codons: CCC (P) and GGT(G). As before, this suggests that the effect is overall very weak or applies to only specific cases. A possible weak effect is also supported by the fact that model comparisons were not robust to minimum length of the helical structures included in the analysis.

We also tested for differences at the second and third positions of helical structures, as previously found by Pechmann and Frydman (2013). Specifically, Pechmann and Frydman (2013) found the second and third positions were enriched in non-optimal codons, but we found no support for this hypothesis. Furthermore, 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 often 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 parameter estimates are more clearly rooted in evolutionary processes, parameter interpretation becomes easier. It also is able to provide 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 evo-

lutionary 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 what patterns are expected if natural selection on codon usage was the same between two regions. 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, this distribution may be quite different from the null distribution assumed in many standard hypothesis tests.

Although we find that methods like ROC-SEMPPR can be very useful for the analysis of codon usage patterns, many challenges remain. 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, our current approach cannot 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. A possible solution 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, 2007; 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. Empirical secondary structure information can be downloaded directly from PDB. Code for analysis and generating figures available at 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.

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.

Rainer Bomblies, Manuel Patrick Luitz, Sandra Scanu, Tobias Madl, and Mar-

tin 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.

Monica Bucciantini, Elisa Giannoni, Fabrizio Chiti, Fabiana Baroni, Niccolò Taddei, Giampietro Ramponi, Christopher M. Dobson, and Massimo Stefani. Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature*, 416 (6880):507–511, apr 2002. ISSN 00280836. doi: 10.1038/416507a.

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 effect of context on synonymous codon usage in genes with low codon usage bias. *Nucleic Acids Res.*, 18(10):2869–2873, 1990.

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`.

Prajwal Ciryam, Richard I. Morimoto, Michele Vendruscolo, Christopher M. Dobson, and Edward P. O'Brien. In vivo translation rates can substantially delay the cotranslational folding of the Escherichia coli cytosolic proteome. *Proc. Natl. Acad. Sci. U. S. A.*, 110(2): E132–E140, jan 2013. ISSN 00278424. doi: 10.1073/pnas.1213624110.

Thomas F. Clarke IV and Patricia L. Clark. Rare codons cluster. *PLoS One*, 3(10), oct 2008. ISSN 19326203. doi: 10.1371/journal.pone.0003412.

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, 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.

D A Drummond and C O Wilke. The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.*, 10:715–724, 2009.

F Feyertag, P M Berninsone, and D Alvarez-Ponce. Secreted Proteins Defy the Expression Level-Evolutionary Rate Anticorrelation. *Mol. Biol. Evol.*, 34(3):692–706, 2017. doi: 10.1093/molbev/msw268.

N Fluman, S Navon, E Bibi, and Y Pilpel. mRNA-programmed translation pauses in the targeting of E. coli membrane proteins. *Elife*, 3:e03440, 2014. doi: 10.7554/eLife.03440.

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.

Ruth Hershberg and Dmitri A. Petrov. Selection on Codon Bias. *Annu. Rev. Genet.*, 42(1): 287–299, dec 2008. ISSN 0066-4197. doi: 10.1146/annurev.genet.42.110807.091442. URL `http://www.annualreviews.org/doi/10.1146/annurev.genet.42.110807.091442`.

A J Hockenberry, M I Sirer, L A N Amaral, and M C Jewett. Quantifying Position-Dependent Codon Usage Bias. *mol. Biol. Evol*, 31(7):1880–1893, 2014. doi: 10.1093/molbev/msu126.

W Holtkamp, G Kokie, M Jager, Joerg Mittelstaet, A A Komar, and M V Rodnina. Co-translational 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.

Toshimichi Ikemura. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and Escherichia coli with reference to the abundance of isoaccepting transfer R. *J. Mol. Biol.*, 158(4):573–597, jul 1982. ISSN 00222836. doi: 10.1016/0022-2836(82)90250-9.

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.

Anton A. Komar, Thierry Lesnik, and Claude Reiss. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett.*, 462(3):387–391, dec 1999. ISSN 00145793. doi: 10.1016/S0014-5793(99)01566-5.

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.

Igor A Krasheninnikov, Anton A Komar, and Ivan A Adzhubei. Nonuniform size distribution of nascent globin peptides, evidence for pause localization sites, and a cotranslational protein-folding model. *Artic. J. Protein Chem.*, 10(5), 1991. doi: 10.1007/BF01025472. URL https://www.researchgate.net/publication/21357366.

Grzegorz Kudla, Andrew W. Murray, David Tollervey, and Joshua B. Plotkin. Coding-sequence determinants of expression in escherichia coli. *Science (80-.).*, 324(5924):255–258, apr 2009. ISSN 00368075. doi: 10.1126/science.1170160.

C. G. Kurland. Translational Accuracy and the Fitness of Bacteria. *Annu. Rev. Genet.*, 26 (1):29–50, dec 1992. ISSN 0066-4197. doi: 10.1146/annurev.ge.26.120192.000333. URL http://www.annualreviews.org/doi/10.1146/annurev.ge.26.120192.000333.

Abigail L. Labella, Dana A. Opulente, Jacob L. Steenwyk, Chris Todd Hittinger, and Antonis Rokas. Variation and selection on codon usage bias across an entire subphylum. *PLoS Genet.*, 15(7):e1008304, jul 2019. ISSN 15537404. doi: 10.1371/journal.pgen.1008304. URL https://doi.org/10.1371/journal.pgen.1008304.

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`.

H Liu, S U Rahman, Y Mao, X Xu, and S Tao. Codon usage bias in 5' terminal coding sequences reveals distinct enrichment of gene functions. *Genomics*, 109:506–513, 2017. doi: 10.1016/j.ygeno.2017.07.008.

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.

S Mahlab and M Linial. Speed Controls in Translating Secretory Proteins in Eukaryotes - an Evolutionary Perspective. *PLoS Comput. Biol.*, 10(1):e1003294, 2014. doi: doi:10.1371/journal.pcbi.1003294.

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`.

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.

S Pechmann, J W Chartron, and J Frydman. Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by SRP \textit{in vivo}. *Nat. Struct. Mol. Biol.*, 21(12):1100–1105, 2014.

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.

P M Power, R A Jones, I R Beacham, C Bucholtz, and M P Jennings. Whole genome analysis reveals a high incidence of non-optimal codons in secretory signal sequences of \textit{Escherichia coli}. *Biochem. Biophys. Res. Commun.*, 322:1038–1044, 2004.

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.

P M Sharp and W Li. The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucl. Acids Res.*, 15(3):1281–1295, 1987.

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/10.1111/1467-9868.00353 https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9868.00353 https://rss.onlinelibrary.wiley.com/doi/10.1111/1467-9868.00353.

N Stoletzki and A Eyre-Walker. Synonymous Codon Usage in Escherichia coli: Selection for Translational Accuracy. *Mol. Biol. Evol.*, 24(2):374–381, 2007.

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 1996a. ISSN 09618368. doi: 10.1002/pro.5560051003. URL http://doi.wiley.com/10.1002/pro.5560051003.

T.A. Thanaraj and Patrick Argos. Ribosome-mediated translational pause and protein domain organization. *Protein Sci.*, 5(8):1594–1612, aug 1996b. ISSN 09618368. doi: 10.1002/pro.5560050814. URL http://www.ncbi.nlm.nih.gov/pubmed/8844849 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2143486 http://doi.wiley.com/10.1002/pro.5560050814.

Tamir Tuller, Isana Veksler-Lublinsky, Nir Gazit, Martin Kupiec, Eytan Ruppin, and Michal Ziv-Ukelson. Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol.*, 12(11): R110, nov 2011. ISSN 14747596. doi: 10.1186/gb-2011-12-11-r110. URL http://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-11-r110.

Vladimir N. Uversky. The intrinsic disorder alphabet. III. Dual personality of serine, 2015. ISSN 21690707.

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.

Jian-Rong Yang, Ben-Yang Liao, Shi-Mei Zhuang, and Jianzhi Zhang. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc. Natl. Acad. Sci. U. S. A.*, 109(14):E831–40, apr 2012. ISSN 1091-6490. doi: 10.1073/pnas.1117408109. URL `http://www.ncbi.nlm.nih.gov/pubmed/22416125` `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3325723`.

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.

Mian Zhou, Jinhu Guo, Joonseok Cha, Michael Chae, She Chen, Jose M. Barral, Matthew S. Sachs, and Yi Liu. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature*, 495(7439):111–115, mar 2013. ISSN 0028-0836. doi: 10.1038/nature11833. URL `http://www.nature.com/articles/nature11833`.

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`.