Quantifying Selection on Codon Usage in Signal Peptides: Gene Expression and Amino Acid Usage Explain Apparent Selection for Inefficient Codons

Alexander L. Cope[1], Robert L. Hettich[1,2], and Michael A. Gilchrist[1,3,4]

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

[2]Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN

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

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

Last compiled on Thursday 14th June, 2018 at 18:44.

Abstract

Secreted proteins play central roles across all taxa. Although secretion mechanism can vary across taxa, all taxa share the Sec secretion pathway. A critical and distinct feature shared by Sec secreted proteins is the signal peptide. Researchers claim signal peptides contain a bias for translation inefficient codons in signal peptides, leading researchers to suggest selection favors translation inefficiency in this region. We investigate codon usage in the signal peptides of *E. coli* using the Codon Adaptation Index (CAI) and tRNA Adaptation Index (tAI), and the ribosomal overhead cost formulation of the stochastic evolutionary model of protein production rates (ROC-SEMPPR). Initial comparisons between signal peptides and 5'-ends of non-signal peptide genes using CAI and tAI are consistent with translationally inefficient codons being preferred in signal peptides. However, simulations reveal these differences are due to amino acid usage and gene expression – we find evidence for novel selection disappears when accounting for both of these factors. In contrast, ROC-SEMPPR, a mechanistic population genetics model capable of separating the effects of selection and mutation bias, shows codon usage bias (CUB) of the signal peptides is indistinguishable from the 5'-coding regions of cytoplasmic proteins. Additionally, we find CUB in the 5'-coding regions is weaker than later segments of the gene. Results illustrate the value in using models grounded in population genetics to interpret genetic data. In summary, we show failure to account for mutation bias and the effects of gene expression on the efficacy of selection against translation inefficiency can lead to a misinterpretation of codon usage patterns.

Introduction

A secreted protein can broadly be defined as any protein entering a secretory pathway for transport through a cellular membrane. These proteins serve important cellular functions, including metabolism and antibiotic resistance (Green and Mecsas, 2016; Saier, 2006). Secreted proteins also play essential roles in the virulence of pathogenic bacteria (Green

and Mecsas, 2016). Numerous secretion systems exists and vary between and within taxa (Bendtsen *et al.*, 2005; Green and Mecsas, 2016; Saier, 2006). Despite the diversity of secretion pathways, the general secretion pathway, also commonly referred to as the Sec pathway, is found across all domains of life (Green and Mecsas, 2016; Natale *et al.*, 2008). In brief, proteins are transported to the SecYEG translocon located in the membrane via the SecA/B chaperone-dependent (SecA/B and SRP) or chaperone-independent manner (Natale *et al.*, 2008; Tsirigotaki *et al.*, 2017). All SecA/B-dependent proteins and chaperone-independent, as well as some SRP-dependent proteins, contain a short peptide chain located at the N-terminus of the protein (Green and Mecsas, 2016; Natale *et al.*, 2008; Tsirigotaki *et al.*, 2017) This short peptide chain is called the signal peptide which is essential for the protein to enter into the SecYEG translocon. Although signal peptides do vary in their amino acid sequences, signal peptides have distinct physicochemical properties which constrain their amino acid usage. A signal peptide generally consists of 3 regions: a positively charged N-terminus, a hydrophobic core, and a polar C-terminus, where the signal peptide is cleaved from the rest of the protein (a.k.a. the mature peptide) (Natale *et al.*, 2008; Tsirigotaki *et al.*, 2017; Zalucki *et al.*, 2009).

The ability to accurately predict signal peptides is useful for identifying secreted proteins in non-model organisms; this has led to the development of machine learning approaches to predict signal peptides which take advantage of the distinct physicochemical properties of signal peptides, such as SignalP (Petersen *et al.*, 2011). Although the physicochemical properties of signal peptides are consistent, previous work found altering the N-terminus has a range of effects on protein secretion: from a decrease in secretion to no effect (Inouye *et al.*, 1982; Nesmeyanova *et al.*, 1997; Puziss *et al.*, 1989; Vlasuk *et al.*, 1983). These varying effects led some researchers to suspect other mechanisms also contribute to the efficacy of protein secretion (Zalucki *et al.*, 2009, 2011a).

Numerous studies suggests codon usage bias (CUB) – the non-uniform usage of synonymous codons – contributes to effective protein secretion in *E. coli* (Burns and Beachamn,

1985; Power *et al.*, 2004; Zalucki and Jennings, 2007; Zalucki *et al.*, 2008, 2010, 2011b). Power *et al.* (2004) found *E. coli* K12 MG1655 signal peptides are biased for translation inefficient codons, which are predicted to be translated slower than their synonymous counterparts. This is in stark contrast to the rest of the *E. coli* proteome, where *E. coli* is biased towards the most efficient codons (Ikemura, 1981; Power *et al.*, 2004). Li *et al.* (2009); Liu *et al.* (2017); Mahlab and Linial (2014) examined the usage of inefficient codons in signal peptides of *S. coelicolor*, *S. cerevisiae*, and various mutlicellular eukaroytes and came to similar conclusions when applying codon usage indices such as the Codon Adaptation Index (CAI, Sharp and Li, 1987) and tRNA Adaptation Index (tAI, dos Reis *et al.*, 2004). Consistent across this work is the interpretation that selection is driving the apparent increase in inefficient codon usage in signal peptides. Similar studies concluded an overabundance of the lysine codon AAA at the second position in the signal peptide promoted efficient translation initiation (Zalucki *et al.*, 2007).

Researchers proposed an adaptive role for inefficient codons in the protein secretion process in which the combination of efficient translation initiation and inefficient translation resulted in reduced distance between sequential ribosomes translating the mRNA of a protein containing a signal peptide (Zalucki *et al.*, 2009, 2011a). They argued this would lead to more efficient recycling of the chaperones involved in the secretion process. Other explanations for the observed increase in inefficient codons include the inability of *E. coli* SRP to induce a translational pause following signal peptide recognition(Powers and Walter, 1997; Zalucki *et al.*, 2009) and slowing down the co-translational folding of the protein, as a folded protein cannot be translocated through the SecYEG translocon (Power *et al.*, 2004; Zalucki and Jennings, 2007; Zalucki *et al.*, 2008, 2011a). If signal peptides have a different CUB relative to the rest of the genome, then codon-level information could be incorporated into signal peptide prediction tools.

In contrast, a recent analysis of ribosome profiling data found no difference in the ribosome densities of the signal peptides and the 5'-ends of nonsecretory genes in various

eukaryotes (Liu *et al.*, 2017). If selection were acting on codon usage in signal peptides to slow down translation, we would expect to see higher ribosome densities in these regions. Additionally, while both Mahlab and Linial (2014) and Liu *et al.* (2017) examined codon usage in relation to secretion in *H. sapiens* using a metric based on tAI, only Mahlab and Linial (2014) found results consistent with increased frequencies of inefficient codons in signal peptides. From a population genetics perspective, it is surprising statistically significant results were obtained in a mammal, which usually have little adaptive CUB due to their lower effective population sizes (Charlesworth, 2009; Lynch *et al.*, 2016). More recently, Samant *et al.* (2014) found codon optimization of a signal peptide improved localization of the protein to the periplasm of *E. coli*, seemingly contradicting a general role for inefficient codon usage in signal peptides. A potential reason for these contradictions is the previous analyses of signal peptide codon usage by Li *et al.* (2009); Liu *et al.* (2017); Mahlab and Linial (2014); Power *et al.* (2004) did not adequately account for the evolutionary forces shaping codon usage (Bulmer, 1990; Gilchrist *et al.*, 2015; Shah and Gilchrist, 2011; Wallace *et al.*, 2013).

We re-examined CUB in signal peptides of *E. coli* using CAI, tAI, and ROC-SEMPPR - a population genetics model which accounts for selection, mutation bias, and gene expression - to determine if selection on codon usage in signal peptides differs from the 5'-ends of genes. Although we find significant differences in codon usage using CAI and tAI, we present evidence these differences are due to signal peptide-specific amino acid biases and differences in the gene expression distributions of signal peptide and non-signal peptide genes. When comparing signal peptides and the 5'-ends of non-signal peptides genes with ROC-SEMPPR, we find signal peptide codon usage is consistent with the 5'-ends of genes not containing a signal peptide. We find selection on codon usage favors the efficient codons, but the strength of selection is weaker at the 5'-ends, corroborating previous analyses (Eyre-Walker, 1996; Gilchrist and Wagner, 2006; Gilchrist, 2007; Power *et al.*, 2004; Qin *et al.*, 2004).

Our work demonstrates the value of analyzing CUB from a formal population genetics

framework, as well as highlights potential limitations with using more common metrics such as CAI for analyzing codon usage on relatively small regions of the genome. Failure to account for variation in the strength of selection due to variation in gene expression can lead to conflating mutation bias with selection, resulting in a misinterpretation of observed codon usage patterns. Our work also illustrates the importance of considering non-adaptive forces in shaping biological phenomenon before invoking adaptive explanations (Gould and Lewontin, 1979). We believe this is particularly important in the modern genomic-age when the combination of large datasets, misinterpretation of p-values, and and inherent bias towards adaptationist interpretations could mislead researchers.

Materials and Methods

Signal Peptide Prediction

Signal peptides were predicted using SignalP 4.1 (Petersen *et al.*, 2011) using both the default cutoff D-score of 0.51 and a more conservative D-score of 0.75. In brief, SignalP consists of two neural networks, one for determining the amino acid sequence similarity to signal peptides and the other for identifying the most likely cleavage site. The results of both neural networks are combined into one value, called the D-score. The D-score ranges between 0 and 1. Setting the cutoff D-score closer to 1 results in a lower false positive rate. A set of confirmed signal peptides for *E. coli* K12 MG1655 was taken from The Signal Peptide Website. All analyses in the main text will focus on the set of signal peptides with $D \geq 0.51$ as this set provides us with the most data; analyses of the $D > 0.75$ and set of confirmed signal peptides give similar results (see Supplementary Material).

ROC-SEMPPR

Given a set of protein-coding genes, ROC-SEMPPR employs a Markov Chain Monte Carlo (MCMC) to estimate codon specific parameters for mutation bias ΔM and pausing times $\Delta\eta$

for each codon within a synonymous codon family (Table 1). In previous work, $\Delta\eta$ was scaled relative to the most efficient codon, which had $\Delta\eta$ and ΔM values fixed at 0. To avoid the choice of reference codon affecting our comparisons of CUB between regions, all $\Delta\eta$ values in this paper are re-scaled such that these values are centered around 0 for each amino acid. The $\Delta\eta$ values reflect the strength and direction of selection against translation inefficiency in a set of protein-coding regions (e.g. the signal peptides). A region with stronger selection against translation inefficiency will have higher $\Delta\eta$ values on average than a region with weaker selection. Similarly, a region which favors translation inefficiency would be expected to have $\Delta\eta$ values which negatively correlate with a region which favors translation efficiency.

ROC-SEMPPR also estimates an average protein production rate ϕ for each gene (Table 1). We find ROC-SEMPPR estimated ϕ values correlate well with empirical measurements of protein production rates for *E. coli* (see Supplementary Methods: Assessing ROC-SEMPPR Model Adequacy and Figures S1 - S2). If changes in synonymous codon usage alter the efficiency at which a protein is translated, then such a change will have the largest impact on the energetic costs of proteins with high production rates, making ϕ a more appropriate gene expression metric than say, mRNA abundance or protein abundance. Thus, we use protein production rates ϕ as our metric of gene expression. For more details on ROC-SEMPPR, see Gilchrist *et al.* (2015). Analysis of CUB with ROC-SEMPPR was performed using AnaCoDa (Landerer *et al.*, 2018).

Parameters	Description
$\Delta\eta_i$	Cost of translating codon i relative to reference codon
ΔM_i	Mutation bias towards codon i relative to the reference codon
ϕ_k	Average Protein Production Rate of gene k

Table 1: Description of ROC-SEMPPR parameters used in this paper.

CAI and tAI

Analysis of CUB was also performed using CAI (Sharp and Li, 1987) and tAI (dos Reis *et al.*, 2004). Both CAI and tAI quantify CUB by assigning weights to the 61 sense codons. For CAI, each codon is assigned a weight based on its relative frequency to its synonymous counterparts in a reference set of highly expressed genes, such as ribosomal protein coding genes. The key assumption of CAI is the most frequent codons in the reference set are the most efficient codons (Sharp and Li, 1987). In contrast, tAI assigns weights based on tRNA abundances corresponding to a codon, as well as accounting for codon-anticodon interactions. The key assumption of tAI is the most efficient codons are usually those with the most abundant tRNA (dos Reis *et al.*, 2004).

CAI and tAI both range between 0 and 1. A CAI score closer to 1 represents a sequence which more closely resembles the codon usage of the reference set of genes, while a tAI closer to 1 indicates a sequence is more closely adapted to the genomic tRNA pool (dos Reis *et al.*, 2004; Sharp and Li, 1987). Calculations for CAI were performed using the AnaCoDa (Landerer *et al.*, 2018), while tAI was calculated using the R package tAI (dos Reis, 2016).

Generating Datasets

Previous analysis of the *E. coli* genome found a set of genes with CAI values that had a negative correlation with their gene expression estimates (dos Reis *et al.*, 2003). It is expected many of these genes were the result of horizontal gene transfer and had not yet reached evolutionary equilibrium with respect to their CUB. We repeated the analysis described in dos Reis *et al.* (2003) on the current *E. coli* K12 MG1655 genome (version 3 ,NC_000913.3). Briefly, correspondence analysis was performed using CodonW (Peden, 1999), followed by clustering based on the principle axis scores using the CLARA algorithm (Maechler *et al.*, 2018) in R. Our analysis was consistent with the findings of (dos Reis *et al.*, 2003), revealing 782 genes with a CUB deviating significantly from the majority of the *E. coli* genome. We will refer to this set of 782 genes as the "exogenous" component of the genome and the rest of

the *E. coli* genome as the "endogenous" for simplicity. All analyses presented will consider only "endogenous" genes because the "exogenous" genes may violate the assumptions of ROC-SEMPPR, CAI, and tAI.

Proteins with a signal peptide were split into the signal peptide and the mature peptide – the segment of the peptide chain after the signal peptide. On average, the signal peptides were 23 codons long. For comparisons to the 5'-ends of nonsecretory genes – defined here as those lacking a signal peptide – the first 23 codons of the nonsecretory genes were used. We note the nonsecretory genes have an average protein production rate ϕ lower than that of the signal peptide genes ($\bar{\phi} = 0.992$ and $\bar{\phi} = 1.08$,respectively, Figure S3).

As the strength of selection on CUB scales with protein production rate ϕ, we created a control group that eliminates differences in the distribution of ϕ for the nonsecretory genes and signal peptide genes. Specifically, the nonsecretory genes were selected using acceptance-rejection sampling to create the "pseudo-secreted proteins". In brief, acceptance-rejection sampling is a procedure for sampling from a population such that its distribution of a metric for one population mirrors the distribution of the same metric for another population. In this case, the pseudo-secreted proteins were sampled such that the mean and variance of the $\log(\phi)$ values reflected those of the genes with a signal peptide. The CUB signature of a gene varies with protein production rate ϕ; thus we can be more confident any differences seen between genes with a signal peptide and pseudo-signal peptide genes are not due to differences in their respective ϕ distributions. All pseudo-secreted proteins were split into two regions we will refer to as the "pseudo-signal peptides" and the "pseudo-mature peptides" (the first 23 codons and the remainder of the gene, respectively).

To assess the performance of CAI and tAI when comparing regions with differences in the distributions of protein production rates ϕ and amino acid biases, simulated sequences were used. Sequences based on the 5'-ends of nonsecretory genes, pseudo-signal peptides, and signal peptides were simulated using the AnaCoDa package (Landerer *et al.*, 2018). To normalize for amino acid usage, sequences 23 amino acids in length were randomly generated to

match the amino acid frequencies of the signal peptides. The codon usage of these sequences was also simulated in AnaCoDa, assuming either the ϕ distribution of the nonsecretory genes or the pseudo-secreted proteins. All sequences were simulated using the pausing times $\Delta\eta$ and mutation bias ΔM parameters estimated from the 5'-end of endogenous nonsecretory genes.

CUB analyses

We estimated protein production rates ϕ by fitting ROC-SEMPPR to the complete protein-coding sequences in the *E. coli* K12 MG1655 genome. Analysis of intragenic (eg. signal vs. mature peptides) and intergenic (eg. pseudo-signal peptides vs. real signal peptides) CUB was carried out using the mixture distribution functionality available in the AnaCoDa implementation of ROC-SEMPPR (Landerer *et al.*, 2018). Each group of regions (eg. signal peptides, mature peptides, etc.) was assumed to have an independent CUB, allowing pausing time $\Delta\eta$ estimates to vary between them. We assumed mutation bias was consistent for the entire genome; thus, we forced mutation bias ΔM parameters to be equal across the groups of regions. ϕ was fixed for each region at the value estimated from the region's corresponding complete protein-coding sequence. This is done for two reasons: (a) shorter regions, such as the signal peptide, likely have insufficient information to accurately estimate ϕ and (b) this guarantees our gene expression metric has the same impact on the estimates of $\Delta\eta$ and ΔM for intragenic regions, such as a signal peptide and its corresponding mature peptide.

A Model-II regression was used to compare pausing times $\Delta\eta$ between regions. Unlike ordinary least squares, Model-II regression, or errors-in-variables regression, accounts for errors in both the x and y variables (Sokal and Rohlf, 1995). When both variables are subject to error, which is the case for the $\Delta\eta$ estimates, the use ordinary least squares leads to downwardly biased parameter estimates. A Model-II regression slope $\beta = 1$ (or the $y = x$ line) will serve as the null hypothesis, as this indicates both the strength and direction of selection between two regions are the same. The intercept parameter was fixed at $\alpha = 0$

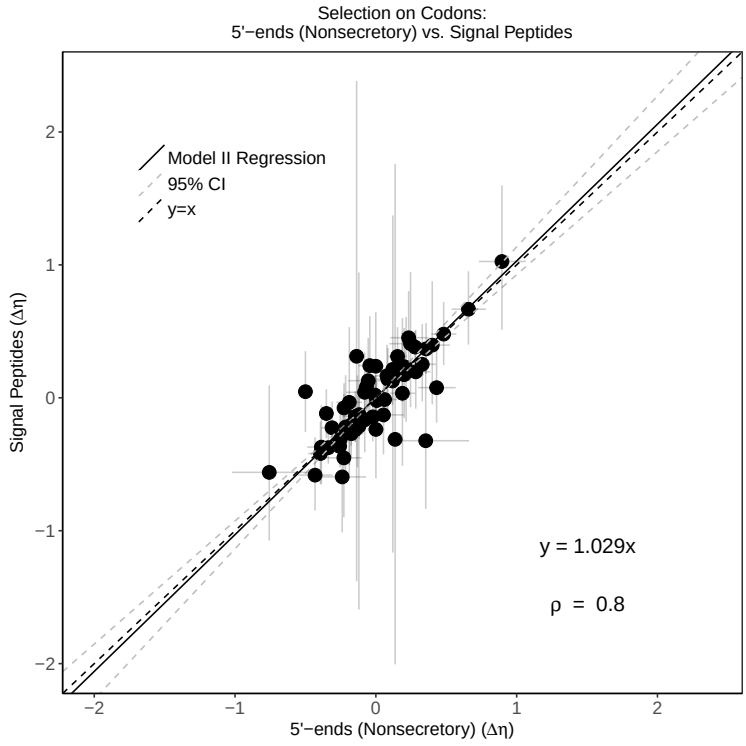
because the $\Delta\eta$ estimates are scaled such that the mean value of $\Delta\eta$ is 0. We note that when we allowed the α parameter to vary, it was as expected, approximately 0. For more details on our use of Model-II regression, see Supplementary Methods.

CAI and tAI were used to compare codon usage between signal peptides, 5'-ends, and pseudo-signal peptides (dos Reis *et al.*, 2003, 2004; Sharp and Li, 1987). As recommended by Sharp and Li (1987), methionine and tryptophan were not included when normalizing for the length of the gene in our calculations of CAI. Statistical significance was assessed using a one-tailed Welch's t-test in R (R Core Team, 2018). R and Python scripts used for this paper can be found at https://github.com/acope3/Signal_Peptide_Scripts.

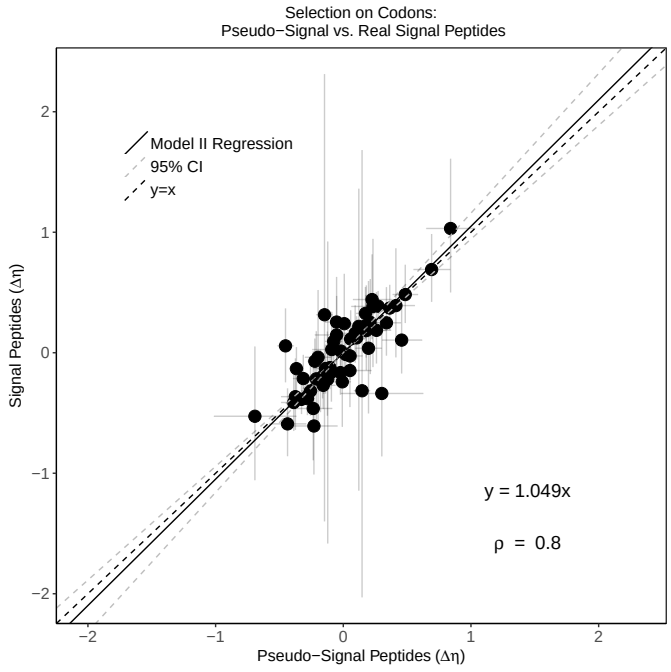
Results

Our analysis of CUB in signal peptides and the 5'-ends of nonsecretory genes using ROC-SEMPPR revealed these regions to be highly similar. Qualitatively, the expected codon frequencies for the 5'-ends of nonsecretory genes and the signal-peptides based on the pausing time $\Delta\eta$ and mutation bias ΔM values estimated from these regions are similar (Figure S4). Notable exceptions appear to be cysteine, aspartic acid, lysine, glutamine, and tyrosine; however, the 95% posterior probability intervals of cysteine and glutamine are the only ones which fail to overlap with $y = x$ line. When comparing the pausing times $\Delta\eta$ of signal peptides to the 5'-ends of nonsecretory genes using a Model-II regression, we find no significant difference from the $y = x$ line (slope β 95% confidence interval: 0.923 – 1.128, Figure 1a). To determine if differences were not detected due to underlying differences in the distributions of ϕ, we compared $\Delta\eta$ estimated from signal peptides and pseudo-signal peptides. Again, no statistically significant difference from the $y = x$ line was found and the expected codon frequencies are similar (β 95% confidence interval: 0.939 – 1.149, Figure 1b andS5). Similar results are obtained using the signal peptides with a D-score greater than 0.75 or the confirmed signal peptides (Figures S6 - S7). We also see no significant result when

using empirically estimated ϕ values ($\beta = 0.908$, 95% confidence interval: $0.67 - 1.16$, Figure S8), although these results show much more variability. The increased variability in the $\Delta\eta$ values and corresponding regression line is unsurprising given the empirically estimated ϕ values are subject to significant noise (Figure S2), but are, in this case, treated as fixed values representative of the true average protein production rate for a gene.



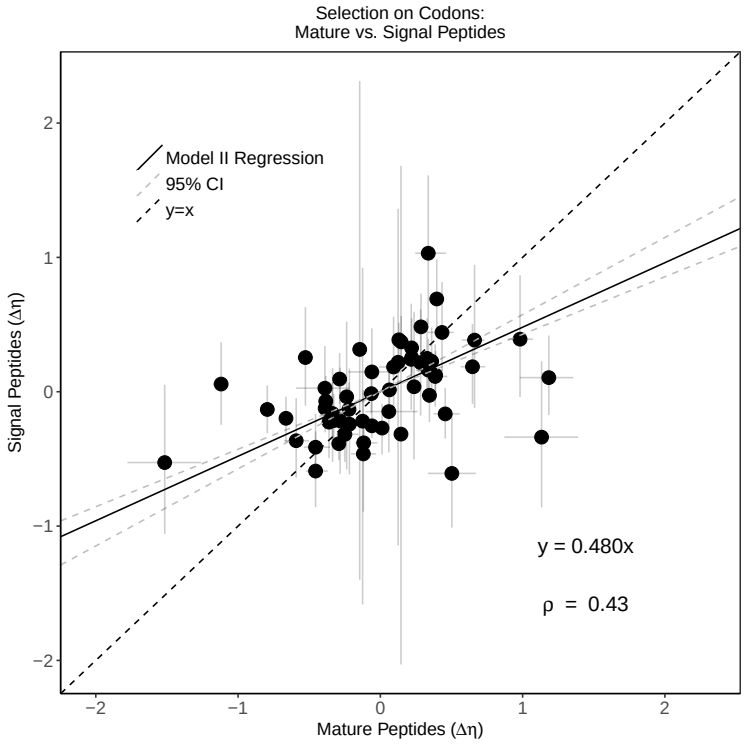
(a)



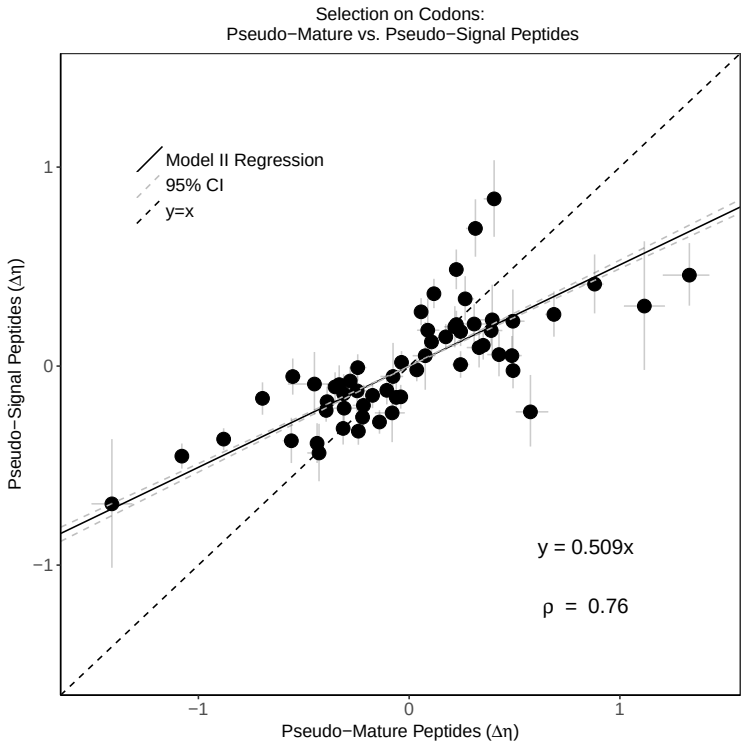
(b)

Figure 1: Comparing the pausing time estimates $\Delta\eta$ between (a) the 5%-ends of nonsecretory genes or (b) pseudo-signal peptides to signal peptides. Grey dashed lines represent the 95% confidence intervals of the regression line. Results clearly show a strong positive linear relationship ($\rho = 0.80$) between the regions and a regression line not significantly different from $y = x$.

The Model-II regression lines estimated from the mature vs. signal peptide comparison and the pseudo-mature vs. pseudo-signal peptide comparison are similar, which serves as further evidence the selection on codon usage in signal peptides and the 5'-ends of nonsecretory genes is the same (Figure 2). The mature vs. signal peptide comparison produces a regression line with slope $\beta = 0.480$ (95% confidence interval: 0.428 - 0.574), while the pseudo-mature vs. pseudo-signal peptide comparison produces a regression line with slope $\beta = 0.496$ (95% confidence interval: 0.490 - 0.533). If selection on codon usage differs in signal peptides from pseudo-signal peptides, we would not expect to see similar regression lines.



(a)



(b)

Figure 2: (a) Comparing the codon pausing time estimates $\Delta\eta$ between mature peptides and signal peptide regions. Grey dashed lines represent the 95% confidence intervals of the regression line. Results show a positive linear relationship ($\rho = 0.43$) between the $\Delta\eta$ estimates for the two regions. This indicates codons favored in one region tend to be favored in the other. (b) Same comparison for pseudo-signal peptide genes. Regression estimates are similar to those estimated for the mature and signal peptide comparison.

Noting CAI and tAI do not account for the effects of gene expression, mutation bias, drift, or amino acid biases, we found signal peptides have lower CAI and tAI values compared to the first 23 codons of nonsecretory genes (one-tailed Welch's t-test, $p < 10^{-5}$). This was also the case when looking at the pseudo-signal peptides, which normalizes for protein production rates ϕ. These results with CAI and tAI can potentially be explained by either the preferred use of inefficient codons in signal peptides *or* as artifacts of amino acid biases. Signal peptides have a different amino acid composition from the 5'-end due to the required physicochemical properties of this region (Figure S9). We examined the robustness of tAI and CAI as a means of quantifying differences in selection on codon usage when underlying differences between amino acid composition and ϕ exists using data simulated under the same mutation bias ΔM and pausing time $\Delta\eta$ parameters. When comparing simulated signal peptides to simulated 5'-end of nonsecretory genes and simulated pseudo-signal peptides using CAI, the simulated signal peptides are found to have a significantly lower mean CAI (Welch's t-test, $p < 0.05$) 100% of the time (Figure 3a-b), despite the fact the $\Delta\eta$ and ΔM parameters used to simulate these regions were the same. This suggests the amino acid usage is biasing the signal peptides towards a lower CAI.

When using simulated 5'-ends of nonsecretory genes which have amino acid composition consistent with the signal peptides, the p-values were heavily skewed towards 1. (Figure 3c). This odd behavior is due to the differences in the ϕ distribution differences of the signal peptide and nonsecretory genes. As the former has a higher mean ϕ, the signal peptides on average will have a stronger CUB after normalizing for the amino acid biases. A one-tailed Welch's t-test with the alternative hypothesis being signal peptides have a lower mean CAI, when in reality they likely have a larger mean CAI, would skew the p-value distribution towards 1. Importantly, ROC-SEMPPR did not detect significant differences between signal peptides and the 5'-ends of non-secretory genes, despite differences in the ϕ distributions (Figure 1a). When normalizing for both amino acid usage and ϕ, significant differences in CAI are found approximately 4% of the time, which is close to the expected number of false

positives at the 0.05 significance level (Figure 3d). Similar results are seen when using tAI (Figure S10). Our results indicate CAI and tAI are prone to inflating differences in CUB between two regions when differences in ϕ and amino acid usage are not accounted for.

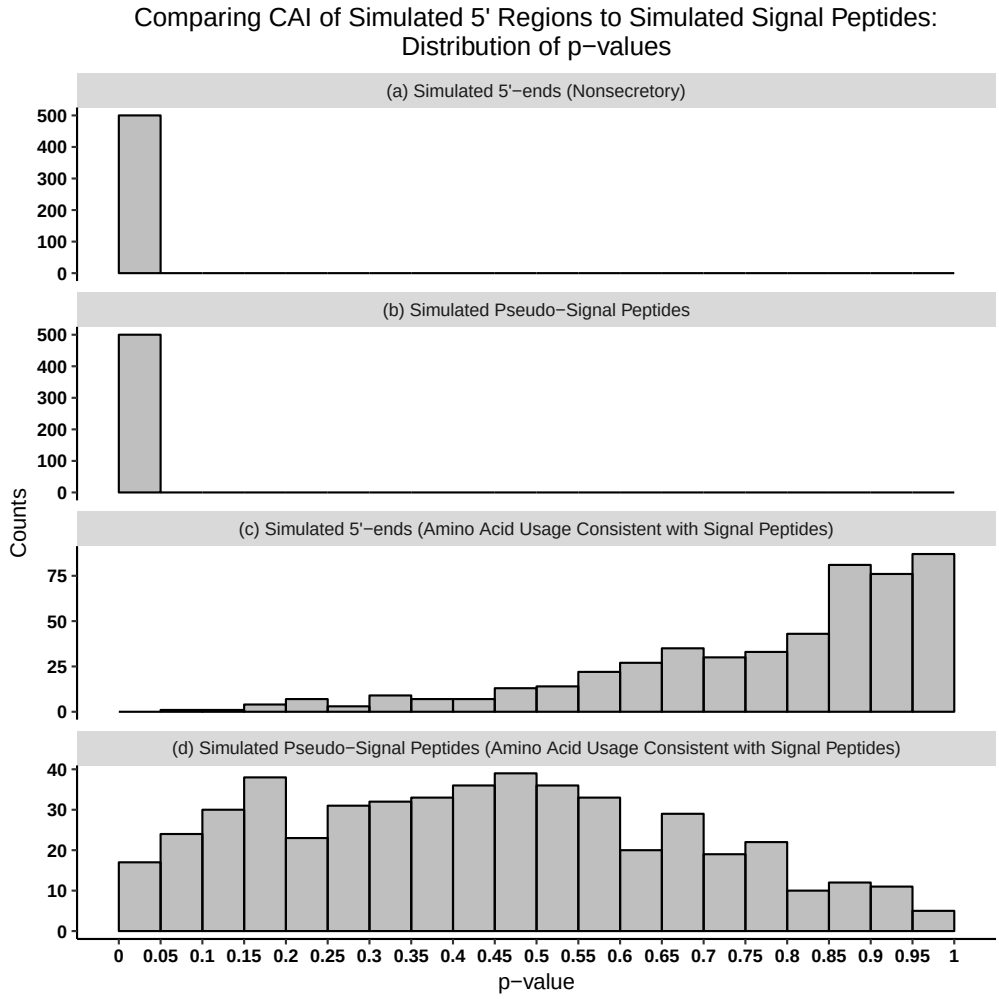


Figure 3: Distribution of p-values from a one-tailed Welch's t-test comparing CAI in simulated nonsecretory 5'-ends, pseudo-signal peptides, and signal peptides in which all regions were simulated using the same pausing time $\Delta\eta$ and ΔM parameters. (a-b) The CAI of simulated signal peptides was found to be significantly lower on average at a 100% false positive rate when compared to simulated 5'-ends of nonsecretory genes and simulated pseudo-signal peptides. (c) Adjusting the amino acid frequencies of the 5'-end of nonsecretory genes to match those of the signal peptides results in a heavily skewed distribution. (d) Adjusting the amino acid frequencies of the pseudo-signal peptides to match those of the signal peptides results in a more uniform distribution.

Notably, selection on codon usage near the N-terminus appears to be on average approx-

imately 50% weaker than the remainder of the gene based on the slope β. Previous analyses using a variety of codon usage metrics found CUB near the 5'-end to be weaker than middle sections of the gene (Eyre-Walker, 1996; Gilchrist and Wagner, 2006; Gilchrist, 2007; Hockenberry *et al.*, 2014; Qin *et al.*, 2004; Power *et al.*, 2004), with these differences being attributed to selection against nonsense errors and to maintain translation initiation efficiency by reducing mRNA secondary structure. We confirm this trend using ROC-SEMPPR (Figure S11).

It was also proposed selection for translation initiation efficiency was shaping signal peptide codon usage, particularly the use of lysine codon AAA, in signal peptides at position 2 of the peptide (Zalucki *et al.*, 2007). We do find AAA appears to be slightly favored in signal peptides, which is not the case in the pseudo-signal peptides, although the 95% posterior probability interval overlaps with the $y = x$ line (Figure S12). If the slight but statistically insignificant favored usage of AAA is due to an increased selection for translation initiation efficiency in signal peptides, then removing the first 3 codons when analyzing signal peptide codon usage should remove this effect. Doing so results in no change in the behavior of AAA, suggesting if there is any selection for increased AAA usage in signal peptides, it is not due to selection for increased translation initiation efficiency (Figure S13). Notably, AAA is both mutationally and selectively-favored for lysine by *E. coli*. Keeping in mind selection on CUB is weaker near the 5'-end of the genes in *E. coli*, the combination of weaker selection, mutational favorability, and a slight increase in the occurrence of lysine in signal peptides (Figure S9) likely drives up the frequency of codon AAA in signal peptides relative to the 5'-ends of nonsecretory genes.

Discussion

In summary, we found no evidence suggesting a general significant difference between selection on codon usage in signal peptides and the 5'-ends of nonsecretory genes in *E. coli* using

a mechanistic model of CUB which incorporates the effects of selection, mutation bias, gene expression, and amino acid usage. Instead, we find failures to account for amino acid usage and protein production rate ϕ resulted in the commonly used codon metrics CAI and tAI indicating significant differences between regions simulated under the same parameters, but these differences disappear when accounting for both amino acid usage and ϕ. Importantly, both amino acid usage and ϕ were significant confounding factors when analyzing CUB with CAI and tAI – only accounting for one of these factors still suggested significant differences between the simulated regions. Although we are not the first to note potential issues with metrics like CAI or tAI for intragenic CUB analysis (Hockenberry *et al.*, 2014), our results demonstrate these metrics are insufficient for intragenic CUB analysis when these regions have drastically different amino acid usage or ϕ distributions, resulting in incorrect biological interpretation.

This is not to say CUB plays no role in the secretion of specific proteins. For example, experimental evidence demonstrates codon optimization of the *E. coli* maltose binding protein's (MBP) signal peptide results in a decrease in protein abundance. Evidence suggests this is due to increased targeting of the codon optimized MBP by proteases due to improper folding (Zalucki and Jennings, 2007; Zalucki *et al.*, 2008). However, CUB as a means to guide proper co-translational folding is not a phenomenon unique to proteins with a signal peptide (Chaney and Clark, 2015; Pechmann and Frydman, 2013; Yu *et al.*, 2015). Although inefficient codons might be crucial to the fold of certain secreted proteins, our results do not indicate this is any more or less so than nonsecretory genes.

Although we found no general difference in selection on codon usage between signal peptides and the 5'-ends, it is possible CUB differences exist among the chaperone-dependent and chaperone-independent mechanisms of the Sec pathway. We are unaware of any CUB comparisons of these three groups, but researchers have noted a region of slower translation downstream from the signal peptide of transmembrane proteins, which are typically secreted via SRP in bacteria (Natale *et al.*, 2008). Using a modified form of the tAI, previous efforts

found a consistent trend of inefficient codons 35-40 codons downstream of the SRP-binding site in various yeasts species (Pechmann *et al.*, 2014). Ribosomal profiling data taken from *S. cerevisiae* provided experimental support for this hypothesis; however this analysis was limited to a small, closely-related phylogeny. Further work is needed to determine the generality of this observation to bacteria and other eukaryotes. Similarly, SRP-dependent transmembrane proteins in *E. coli* have a higher frequency of "programmed pause sites," areas of high ribosomal density downstream from Shine-Dalgarno-like sequences, at the beginning of the gene (Fluman *et al.*, 2014). A higher frequency of programmed pause sites was not observed in the region downstream from the signal peptides in periplasmic proteins. Notably, this region of higher ribosome density downstream from the signal peptides was not observed in periplasmic proteins, which are normally secreted via SecA/B (Natale *et al.*, 2008; Tsirigotaki *et al.*, 2017) However, recent work challenges the findings that Shine-Dalgarno-like sequences are largely responsible for translational pause (Mohammad *et al.*, 2016).

Notably, we do find selection on CUB is weaker at the 5'-ends relative to later portions of the gene, corroborating previous work (Eyre-Walker, 1996; Gilchrist and Wagner, 2006; Gilchrist, 2007; Hockenberry *et al.*, 2014; Power *et al.*, 2004; Qin *et al.*, 2004). Weaker selection at the 5'-ends is often attributed to selection against nonsense errors and selection against mRNA secondary structure. Importantly, the advent of ribosome profiling revealed the presence of high ribosomal density at the 5'-ends, often referred to as the "5'-ramp" (Tuller *et al.*, 2010). The 5'-ramp was originally thought to be the result of increased selection for slow translation at the 5'-end to reduce ribosomal interference further down the transcript, but simulations suggest the 5'-ramp is an artifact of short genes with high initiation rates (Shah *et al.*, 2013). Selection for co-translational folding is also thought to shape intragenic CUB (Chaney and Clark, 2015; Pechmann and Frydman, 2013; Yu *et al.*, 2015). Further work is needed to understand how these various selective forces are balanced to maintain translation efficiency and efficacious protein biogenesis.

Ultimately, our work further illustrates the value of population genetics models which

include nonadaptive evolutionary forces when analyzing genomic data. Biologists are often tempted to explain statistically significant results in the context of selection and adaptation, but researchers must first provide evidence these results cannot be explained by nonadaptive evolutionary forces (eg. mutation bias and genetic drift) and/or as an artifact of some other constraint on the trait of interest (eg. amino acid biases). We are certainly not the first to note the importance of considering nonadaptive explanations. Almost four decades ago, Gould and Lewontin (1979) critiqued the propensity of evolutionary biologists to invoke natural selection and adaptation without seriously considering possible nonadaptive explanations. The explosion of genomic data means now, more than ever, biologists should be hesitant to adopt adaptationists explanations to biological phenomenon without first investigating if such results could be shaped by nonadaptive forces. The embrace of "big data" by biological researchers is a double-edged sword: while we have the ability to investigate patterns and explore hypotheses which would not have been possible 20 years ago, the use of large datasets can lead to incredibly small p-values, which are often misinterpreted as both evidence of a strong effect and a small probability of the null hypothesis being true (Wasserstein and Lazar, 2016). The misinterpretation of p-values and a bias towards adaptationist explanations can be a dangerous combination, with researchers over-interpreting their results and misleading other researchers.

The development of models incorporating both adaptive and nonadaptive evolutionary forces will be important for understanding the selective forces shaping complex biological data. In the case of the studying CUB, codon indices like CAI have long been employed, but these metrics often are unable to disentangle the effects of amino acid biases, mutation, and selection. While often good proxies of gene expression, these indices do not directly incorporate gene expression information into the weights estimated for each codon. This could lead to further problems of conflating mutation bias with selection when comparing CUB across regions. In contrast, because ROC-SEMPPR is grounded in population genetics and thus, is able to decouple selection and mutation bias, it serves as a more accurate and

evolutionarily-grounded tool for researchers interested in studying CUB.

References

Bendtsen, J. D., Kiemer, L., Fausbøll, A., and Brunak, S. (2005). Non-classical protein secretion in bacteria. *BMC Microbiology*, **5**(1), 58.

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

Burns, D. and Beachamn, I. (1985). Rare codons in *E. coli* and *S. typhimurium* signal sequences. *FEBS Letters*, **189**, 318–324.

Chaney, J. and Clark, P. (2015). Roles for synonymous codon usage in protein biogenesis. *Annu. Rev. Biophysics*, **44**, 143–166.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, **10**, 195–205.

dos Reis, M. (2016). *tAI: The tRNA adaptation index*. R package version 0.2.

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

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

Eyre-Walker, A. (1996). Synonymous codon bias is related to gene length in *Escherichia coli*: Selection for translational accuracy? *Mol. Biol. Evol.*, **13**(6), 864–872.

Fluman, N., Navon, S., Bibi, E., and Pilpel, Y. (2014). mrna-programmed translation pauses in the targeting of e. coli membrane proteins. *eLife*, **3**, e03440.

Gilchrist, M. (2007). Combining models of protein translation and population genetics to predict protein production rates from codon usage patterns. *Mol. Biol. Evol.*, **24**(11), 2362–2372.

Gilchrist, M. and Wagner, A. (2006). A model of protein translation inducing codon bias, nonsense errors, and ribosome recyling. *Journal of Theoretical Biology*, **239**, 417–434.

Gilchrist, M., Chen, W., Shah, P., Landerer, C., and Zaretzki, R. (2015). Estimating gene expression and codon-specific translational efficiencies, mutation biases, and selection coefficients from genomic data alone. *Genome Biology and Evolution*, **7**, 1559–1579.

Gould, S. and Lewontin, R. (1979). The spandrels of san marco and the panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London*, **205**(1161), 581–598.

Green, E. and Mecsas, J. (2016). Bacterial secretion systems - an overview. *Microbiol Spectr.*, **4**(1).

Hockenberry, A., Sirer, M., Amaral, L., and Jewett, M. (2014). Quantifying position-dependent codon usage bias. *mol. Biol. Evol*, **31**(7), 1880–1893.

Ikemura, T. (1981). Correlation between the abundance of *Escherichia coli* transfer rnas and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *Journal of Molecular Biology*, **151**, 389–409.

Inouye, S., Soberon, X., Franceschini, T., Nakamura, K., Itakura, K., and Inouye, M. (1982). Role of positive charge on the amino-terminal region of the signal peptide in protein secretion across the membrane. *Proc. Natl. Acad. Sci. USA.*, **79**, 3438–3441.

Landerer, C., Cope, A., Zaretzki, R., and Gilchrist, M. (2018). Anacoda: analyzing codon data with bayesian mixture models. *Bioinformatics*, page bty138.

Li, Y., Xie, Z., Du, Y., Zhou, Z., Mao, X., Lv, L., and Li, Y. (2009). The rapid evolution of signal peptides is mainly caused by relaxed selection on non-synonynous and synonymous sites. *Gene*, **436**, 8–11.

Liu, H., Rahman, S., Mao, Y., Xu, X., and Tao, S. (2017). Codon usage bias in 5' terminal coding sequences reveals distinct enrichment of gene functions. *Genomics*, **109**, 506–513.

Lynch, M., Ackerman, M., Gout, J., Long, H., Sung, W., Thomas, W., and Foster, P. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, **17**, 704–714.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2018). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.0.7-1 — For new features, see the 'Changelog' file (in the package source).

Mahlab, S. and Linial, M. (2014). Speed controls in translating secretory proteins in eukaryotes - an evolutionary perspective. *PLoS Computational Biology*, **10**(1), e1003294.

Mohammad, F., Woolstenhulme, C., Green, R., and Buskirk, A. (2016). Clarifying the translational pausing landscape in bacteria by ribosome profiling. *Cell Reports*, **14**, 686–694.

Natale, P., Bruser, T., and Driessen, A. (2008). Sec- and tat-mediated protein secretion across the bacterial cytoplasmic membrane—distinct translocases and mechanisms. *Biochimica et Biophysica Acta*, **1778**, 1735–1756.

Nesmeyanova, M., Karamyshev, A., Karamysheva, Z., Kalinin, A., Ksenzenko, V., and Kajava, A. (1997). Positively charged lysine at the n-terminus of the signal peptide of the *Escherichia coli* alkaline phosphatase provides the secretion efficiency and is involved in the interaction with anionic phospholipids. *FEBS Letters*, **403**, 203–207.

Pechmann, S. and Frydman, J. (2013). Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature Structural and Molecular Biology*, **20**(2), 237–243.

Pechmann, S., Chartron, J., and Frydman, J. (2014). Local slowdown of translation by nonoptimal codons promotes nascent-chain recognition by srp *in vivo*. *Nature Structural and Molecular Biology*, **21**(12), 1100–1105.

Peden, J. (1999). *Analysis of Codon Usage*. Ph.D. thesis, University of Nottingham.

Petersen, T., Brunak, S., von Heijne, G., and Nielsen, H. (2011). Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods*, **8**(10), 785–786.

Power, P., Jones, R., Beacham, I., Bucholtz, C., and Jennings, M. (2004). Whole genome analysis reveals a high incidence of non-optimal codons in secretory signal sequences of *Escherichia coli*. *Biochemical and Biophysical Research Communications*, **322**, 1038–1044.

Powers, T. and Walter, P. (1997). Co-translational protein targeting catalyzed by the *Escherichia coli* signal recognition particle and its receptor. *The EMBO Journal*, **16**(16), 4880–4886.

Puziss, J., Fikes, J., and Bassford, P. (1989). Analysis of mutational alterations in the hydrophilic segment of the maltose-binding protein signal peptide. *Journal of Bacteriology*, **171**, 2303–2311.

Qin, H., Wu, W., Kreitman, J. C. M., and Li, W. (2004). Intragenic spatial patterns of codon usage bias in prokaryotic and eukaryotic genomes. *Genetics*, **168**, 2245–2260.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Saier, M. (2006). Protein secretion systems in gram-negative bacteria. *Microbe*, **1**(9), 414–419.

Samant, S., Gupta, G., Karthikeyan, S., amd A. Nair, S. H., Sambasivam, G., and Sukumaran, S. (2014). Effect of codon-optimized *E. coli* signal peptides on recombinant *Bacillus stearothermophilus* maltogenic amylase periplasmic localization, yield and activity. *J. Ind. Microbial Biotechnol*, **41**, 1435–1442.

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

Shah, P., Ding, Y., Niemczyk, M., Kudla, G., and Plotkin, J. (2013). Rate-limiting steps in yeast protein translation. *Cell*, **153**, 1589–1601.

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

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

Tsirigotaki, A., Geyter, J. D., Sostaric, N., Economou, A., and Karamanou, S. (2017). Protein export through the bacterial sec pathway. *Nature Reviews: Microbiology*, **15**, 21–36.

Tuller, T., Carmi, A., Vestsigian, K., Navon, S., Dorfan, Y., Zaborske, J., Pan, T., Dahan, O., Furman, I., and Pilpep, Y. (2010). An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell*, **141**, 344–354.

Vlasuk, G., Inouye, S., Ito, H., Itakura, K., and Inouye, M. (1983). Effects of the complete removal of basic amino acid residues from the signal peptide on secretion of lipoprotein in *Escherichia coli*. *J. Biol. Chem.*, **258**, 7141–7148.

Wallace, E., Airoldi, E., and Drummond, D. (2013). Estimating selection on synonymus

codon usage from noisy experimental data. *Molecular Biology and Evolution*, **30**(6), 1438–1453.

Wasserstein, R. and Lazar, N. (2016). The asa's statement on p-values: Context, process, and purpose. *The American Statistician*, **70**(2), 129–133.

Yu, C., Dang, Y., Zhou, Z., Wu, C., Zhao, F., Sachs, M., and Liu, Y. (2015). Codon usage influences the local rate of translation elongation to regulate co-translational protein folding. *Molecular Cell*, **59**, 744–754.

Zalucki, Y. and Jennings, M. (2007). Experimental confirmation of a key role for non-optimal codons in protein export. *Biochemical and Biophysical Research Communications*, **355**, 143–148.

Zalucki, Y., Power, P., and Jennings, M. (2007). Selection for efficient translation initiation biases codon usage at the second amino acid position in secretory proteins. *Nucleic Acids Research*, pages 1–7.

Zalucki, Y., Gittins, K., and Jennings, M. (2008). Secretory signal sequence non-optimal codons are required for expression and export of β-lactamase. *Biochemical and Biophysical Research Communications*, **366**, 135–141.

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

Zalucki, Y., Jones, C., Ng, P., Schulz, B., and Jennings, M. (2010). Signal sequence non-optimal codons are required for the correct folding of mature maltose binding protein. *Biochimica et Biophysica Acta*, **1798**, 1244–1249.

Zalucki, Y., Beacham, I., and Jennings, M. (2011a). Coupling between codon usage, translation and protein export in *Escherichia coli*. *Biotechnology Journal*, **6**, 660–667.

Zalucki, Y., Shafer, W., and Jennings, M. (2011b). Directed evolution of effeicient secretion in the srp-dependent export of tolb. *Biochemica et Biophysica Acta*, **1808**, 2544–2550.