Codon Usage Bias Regulates the Dynamics of Protein Translation

Frank Liu
Yun S. Song
Dasheng Bi
Daniel Erdmann-Pham
Sanjit Batra

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-109
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-109.html

May 11, 2023

Codon Usage Bias Regulates the Dynamics of Protein Translation

by

Frank Liu

A thesis submitted in partial satisfaction of the

requirements for the degree of

Masters of Science

in

Electrical Engineering & Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Yun S. Song, Chair
Professor Jennifer Listgarten

Spring 2023

Abstract

Codon Usage Bias Regulates the Dynamics of Protein Translation

by

Frank Liu

Masters of Science in Electrical Engineering & Computer Science

University of California, Berkeley

Professor Yun S. Song, Chair

Proteins are produced by a process called translation, wherein particles called ribosomes move along mRNAs and assemble together polypeptide chains, one amino acid at a time. The speed at which they do so (referred to as elongation rate) is a crucial determinant of how fast and how much protein is produced at a time, and therefore is of biological and medical interest. With the advent of new experimental technology, it has become possible to probe the positions of ribosomes across mRNAs (through a protocol called Riboseq), possibly allowing inference of and new insights into these elongation rates. In this project, we devised and implemented a novel algorithm that infers codon-specific and coding sequence (CDS) position-specific elongation rates accurately. The algorithm is capable of outperforming state-of-the-art machine learning methods, while still exhibiting favorable runtime. Using the inferred elongation rates, we quantitatively disentangled the role played by synonymous codon usage bias and amino acid choice in explaining variance in smoothed elongation rates. Furthermore, we demonstrated a prominent role played by codon usage bias in regulating the dynamics of protein translation, by optimizing translation efficiency in early regions of the CDS.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my mentors Dr. Dan D. Erdmann Pham, Dr. Sanjit S. Batra, and Professor Yun S. Song along with my collaborator Dasheng Bi for their gracious support and feedback.

Chapter 1

Introduction

1.1 Background

Proteins are produced by a process called translation, wherein particles called ribosomes move along mRNA molecules and assemble together polypeptide chains, one amino acid at a time. The speed at which they do so (referred to as elongation rate) is a crucial determinant of how fast and how much protein is produced, and therefore is of biological and medical interest. Elongation rate is known to vary depending on various factors such as the identity of the codon (and hence amino acid) being translated, the position of that codon along the length of the mRNA coding sequence (CDS), secondary structure within the mRNA molecule, and biophysical properties of the nascent polypeptide [20] [4].

It has also been widely observed that codon usage follows non-uniform trends throughout the coding portions of the genome (i.e. some codons are used more frequently than others) [16]. This phenomenon is termed codon usage bias. It has been postulated that differences in codon usage frequency are attributable to differences in tRNA abundance, such that the more frequently used codons are decoded by tRNA molecules whose concentrations are more abundant within the cell [7]. This would serve to optimize translation efficiency and reduce the quantity of ribosomes necessary to produce the same amount of protein product. This hypothesis is supported by previous studies which have shown that synonymous coding mutations, upregulation of tRNAs, and mutations within tRNAs can have dramatic effects of protein expression, folding, and stability [19] [12] [9] [13]. Codon usage bias has also been shown to exhibit broader impacts beyond just regulation of translation elongation speed. For example, codon usage has been demonstrated to regulate protein structure and function as a consequence of regulating translation elongation speed [21]. Slower translation may also destabilize mRNAs and thus decrease protein expression [17] [1].

With the advent of Riboseq [11], a new experimental technology, it has become possible to probe the positions of ribosomes across genes. Riboseq, also known as ribosome profiling, involves isolating and sequencing fragments of mRNA that are protected by ribosomes during translation. The resulting data (known as ribosome footprints) can be used

to determine which regions of mRNA are being translated. Meanwhile, RNA-seq is another high-throughput sequencing technique used to study the transcriptome of a cell or tissue. Data from RNA-seq experiments can be used to quantify gene expression levels. In particular, normalizing ribosome footprint counts by the quantity of mRNA sequences for that transcript (as determined by RNA-seq) allows one to determine ribosome density measurements at various loci within the translatome.

A frequently-used mathematical model for studying the dynamics of ribosome movement along mRNA molecules during protein translation is the Totally Asymmetric Simple Exclusion Process (TASEP) [14] [24] [25]. The model consists of a one-dimensional lattice of sites, each of which can be either empty or occupied by a particle. The particles move in a single direction (from left to right) and can only occupy empty neighboring sites. The name "totally asymmetric" refers to the fact that the particles can only move in one direction, and "simple exclusion" refers to the fact that each site can only be occupied by one particle at a time.

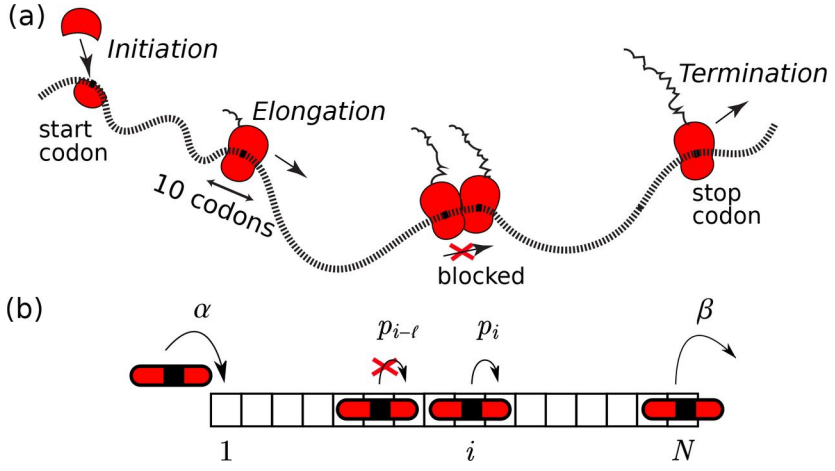

Figure 1.1: Visualization of the TASEP model. Figure borrowed from [5].

The TASEP model can be applied towards studying the process of translation. Ribosomes enter the mRNA lattice with rate α, progress forward with elongation rate λ_i (which varies depending on the position i), and exit the mRNA lattice with rate β. Furthermore, the variable ρ_i in the TASEP model represents the density of particles at any given locus i along the lattice. Here in our project, ρ_i also corresponds to the experimentally observed ribosome density measurement at locus i.

Using formulas derived from the TASEP model (see equation (2.1) through (2.3) in Methods) [5], smoothed ribosome density measurements ρ_i can be inverted to recover smoothed elongation rates λ_i along with initiation rate α and termination rate β. (Here, "smoothed" refers to taking the arithmetic mean of values within every consecutive window of 10 codons). In the current project, we use experimentally-gathered Riboseq and RNA-seq data from *S. cerevisae* [20] along with various vertebrate species (human, macaque, mouse, opossum,

chicken) [22] in order to calculate ρ_i and λ_i for all genes in each dataset. Afterwards, we deconvolve the smoothed elongation rates λ_i in order to recover codon-specific and CDS position-specific elongation rates. We may then use these novel inferred elongation rates to quantitatively assess the role played by codon usage bias in regulating the dynamics of translation.

Being able to infer codon-specific and CDS position-specific elongation rates has many practical applications in the realm of optimizing codon sequences for efficient protein synthesis. It is well-known that the biological language for mapping codons to amino acids is degenerate; in other words there often exist multiple synonymous codons that code for the same amino acid. Therefore, the same polypeptide sequence can be coded through mRNA in multiple different ways. Judicious choice of which synonymous codons to use can have large impacts on the rate at which a protein is synthesized, and its corresponding abundance within the cell. One example of a medical application for this idea relates to optimizing the sequence design for the COVID-19 mRNA vaccine.

1.2 Related Work

Multiple studies in the past have attempted to tackle a similar problem of detecting codon-specific elongation rates. Early studies in 2012 and 2013 aligned ribosome footprint reads to the reference genome to identify the 10 codons found within each footprint, tabulated the frequency of each codon appearing in each position, and did not detect codon-specific differences in decoding rates [6] [3]. However, this method of analysis was criticized for over-weighting highly-expressed genes and failing to define the right normalizations to compensate for differences in gene expression, gene length, sequence composition, etc [7].

In 2014, Gardin et al. [7] measured average decoding rates for each of the 61 sense codons in E. *coli* from ribosome profiling data. Their approach is as follows: for each of the 61 sense codons, the authors first identify all translated regions in the genome where that particular codon uniquely appears at the center of a 19 codon-wide window. For ribosome footprints that are 10 codons long, there are exactly 10 classes of footprints that can fit entirely within this window. The authors calculate the relative frequency with which each of these 10 classes are observed, and average these frequencies across thousands of relevant windows in the transcriptome. Finally, the authors invert these frequencies to compute an average "ribosome residence time" (RRT) metric for each of the 61 sense codons, which represents the average amount of time a ribosome spends decoding that codon. The authors used their inferred elongation rates to conclude that frequent codons are decoded more quickly than rare codons.

In 2018, Duc et al. [4] used probabilistic modeling to estimate initiation and local elongation rates from ribosome profiling data. Briefly, their procedure approximates position-specific elongation rates by taking the inverse of the observed footprint number, and then using simulation to search over the initiation rate that minimizes the difference between the experimental detected-ribosome density and the one obtained from simulation. This is

followed by a procedure for detecting and fixing "error sites" where the difference between absolute density and simulated density exceeds a certain threshold. Using these inferred elongation rates, the authros conclude that biophysical features, such as the amount of electric charges and the hydrophobic properties of the nascent polypeptide explain observed variation in translation elongation rates.

In 2018, Tunney et al. [19] worked on solving a similar task of predicting ribosome densities from codon sequence information in the Weinberg et al. dataset [20]. By training a simple feedforward neural network and incorporating RNA secondary structure information into their predictions, they were able to achieve a 0.57 Pearson correlation between predicted vs experimentally observed ribosome densities. While this approach does not return codon-specific elongation rates, it does demonstrate the capacity for accurate prediction of ribosome densities from codon sequence information alone.

Chapter 2

Methods

2.1 Elongation Inference Procedure

The elongation inference procedure begins with a set of G genes (a training dataset) from which we learn to infer codon-specific and CDS position-specific elongation rates. For each gene $g \in G$ having codon sequence C_g of length L_g, Riboseq and RNA-seq experimental data provide an empirically observed ribosome densities vector of length L_g. The ribosome densities vector is then smoothed (i.e. its values are averaged by taking the arithmetic mean within every consecutive window of 10 codons). Henceforth, ρ_g (having length $L_g - 9$) shall denote the smoothed ribosome densities vector for gene g.

The TASEP model provides us with equation (2.1), (2.2), and (2.3) for inverting ρ_g to recover $\alpha_g, \beta_g, \lambda_g$ [5], where α_g is a scalar denoting the gene-specific initiation rate, β_g is a scalar denoting the gene-specific termination rate, and λ_g is a vector (having the same length as ρ_g) denoting smoothed elongation rate at each window position along the CDS. The superscript i denotes the value of the vector at window position i. In equation (2.3), ρ_g^{1+} is a special symbol denoting the terminal entry in the non-smoothed ribosome densities vector.

$$\lambda_g^i = \frac{1 - (\ell - 1)\rho_g^i}{\rho_g^i(1 - \ell\rho_g^i)} \tag{2.1}$$

$$\alpha_g = \frac{1}{1 - \ell\rho_g^0} \tag{2.2}$$

$$\beta_g = \frac{1}{\rho_g^{1+}} \tag{2.3}$$

In our protocol, we also truncate all codon sequences past codon position 599 due to the sparsity of available data in regions of the CDS that are further downstream past this point.

Next, we introduce several foundational assumptions of our model. Based upon previous reports that different codons have different average ribosome densities (and thus different elongation rates) due to varying cognate tRNA abundances, and that mean elongation rates exhibit non-uniform trends across the length of the CDS [20] [7] [4], we model this system by introducing unique codon-specific and CDS position-specific elongation rates for each of the 64 codons (61 sense codons + 3 nonsense codons). Let r_c^i denote the elongation rate of codon c appearing at CDS position i. In defining r_c^i we assume that within a particular dataset, a given codon appearing at a given CDS position will always have the same elongation rate, regardless of the gene in which it appears. The goal of our model is to learn these unknown r_c^i values. However, to prevent overfitting, we choose to aggregate together certain regions of the CDS that are distal from the 5' end of the mRNA, such that all CDS positions within the specified regions share the same set of 64 codon-specific elongation rates. The choice to do so is consistent with prior observations that mean ribosome densities (and thus elongation rates) across genes exhibit lesser variance in latter regions of the CDS [20]. Assuming 0-indexing of codon positions, the aggregated regions we choose to set are: codon positions 100 to 149 inclusive, 150 to 349 inclusive, and 350 to 599 inclusive. The total number of unknown r_c^i values is thus $64 \times 100 + 64 \times 3 = 6592$. See Figure 2.1 below for a visual representation of the regions, where alternating blue/orange colors indicate successive regions of the CDS, each with their own unique set of 64 codon-specific elongation rates.

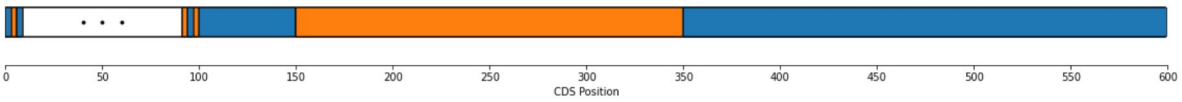

Figure 2.1: Visualization of region boundaries. The alternating blue and orange colors assist with distinguishing neighboring regions from one another.

The next key model assumption is that λ_g^i (the smoothed elongation rate for gene g at window position i) can be represented as:

$$\lambda_g^i = \tau_g \times \frac{1}{10}\left(\sum_{j=0}^{9} r_{C_g^{i+j}}^{i+j} + \epsilon\right), \tag{2.4}$$

where τ_g is a gene-specific time scaling factor, C_g^{i+j} denotes the identity of the codon at index $i+j$ in the codon sequence C_g, and ϵ is some zero-mean Gaussian random noise. Simply put, equation (2.4) states that the smoothed elongation rate for gene g at window position i can be approximated by the arithmetic mean of the ten codon and CDS position-specific elongation rates within the window, times a gene-specific scaling factor τ_g. Equation (2.4) also suggests that some modified version of linear regression can be used to jointly infer the unknown r_c^i and τ_g values.

To solve for the unknown r_c^i and τ_g values, we employ a coordinate-ascent algorithmic approach. We start by assuming (temporarily) that the τ_g values are known, in order to

recover the optimal values for r. In this case, we can rearrange equation (2.4) as shown:

$$\Lambda_g^i := \frac{\lambda_g^i}{\tau_g} = \frac{1}{10}\left(\sum_{j=0}^{9} r_{C_g^{i+j}}^{i+j} + \epsilon\right). \tag{2.5}$$

Here we define Λ_g^i as being equal to λ_g^i normalized by τ_g, or in other words the smoothed homogenized elongation rate. We then further simplify:

$$10\Lambda_g^i = \sum_{j=0}^{9} r_{C_g^{i+j}}^{i+j} + \epsilon. \tag{2.6}$$

By writing out equation (2.6) for every window within every gene from the transcriptome, we can create a system of linear equations from which one can solve for unknown r_c^i values. To simplify notation, we rewrite this system of equations in linear-algebraic form as

$$y = Xr + \epsilon, \tag{2.7}$$

where y is defined as a vector containing the concatenation of all $10\Lambda_g^i$ entries from the left hand side of equation (2.6), r is a vector of length 6592 containing the concatenation of all r_c^i elongation rates, and X is a design matrix having number of rows equal to $\sum_{g \in G} L_g - 9$ and number of columns equal to 6592. Each row of X corresponds to a window from the training dataset, and entries within that row enumerate the number of times each codon from each CDS region appears within the given window of interest. Thus, the sum of entries in each row of X is 10, since there are exactly 10 codons per window. Notably, the X matrix is sparse (i.e. contains many zeros), and can thus be stored in Scipy's CSR matrix format for space efficiency. Lastly, ϵ refers to some Gaussian random noise, distributed according to $\mathcal{N}(0, \sigma^2 I)$ with unknown variance term σ^2.

Although we could proceed to recover r through non-negative least squares regression applied on equation (2.7), we can reduce overfitting in the final result by using FUSED-LASSO regression instead. FUSED-LASSO augments the traditional mean squared error objective function with an additional L1 penalty term, as shown below:

$$\min_r \|Xr - y\|_2^2 + \gamma \sum_{k=0}^{98} \|r_k - r_{k+1}\|_1, \tag{2.8}$$

where $\gamma > 0$ is a hyperparameter that can be tuned via cross-validation, and r_k denotes the vector of 64 codon-specific elongation rates at CDS position k. Intuitively, the additional L_1 penalty term prevents the r_c^i values from varying excessively between successive CDS positions in the first 100 codons of the CDS. The entries of the r vector are then recovered using CVXPY optimization software.

Now that we have established the procedure for inferring r given τ_g, we address the other problem of inferring τ_g given r. To do so, we use the maximum likelihood estimation (MLE)

approach. Before we begin, we reparameterize the linear regression problem by modifying it slightly to simplify the process of calculating derivatives. We multiply both sides of equation (2.7) by the gene-specific τ scalars, resulting in the following form:

$$\tilde{y} = \tau X r + \tilde{\epsilon}, \tag{2.9}$$

where $\tilde{y}$ equals $\tau \times y$ and $\tilde{\epsilon} \sim \mathcal{N}(0, \tilde{\sigma}^2 I)$ for some adjusted variance term $\tilde{\sigma}^2$.

We wish to maximize the likelihood (or equivalently, minimize the negative log-likelihood) of the $\tilde{\epsilon_i} \sim \mathcal{N}(0, \tilde{\sigma}^2)$ terms. Under the assumption that $\tilde{\epsilon}_i$ values are independent across all windows, the likelihood function $\mathcal{L}_g$ (for each gene g) that we seek to maximize is

$$\mathcal{L}_g = \prod_{i=0}^{L_g - 9} \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \exp\left\{ \frac{(\tilde{y}_g^i - \tau_g (Xr)_g^i)^2}{2\tilde{\sigma}^2} \right\}. \tag{2.10}$$

Note that $\tilde{y}_g^i$ simply denotes the entry of the $\tilde{y}$ vector for gene g at index i, and $(Xr)_g^i$ is similarly defined. We can then rewrite the equation above as:

$$\mathcal{L}_g = \prod_{i=0}^{L_g - 9} \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \exp\left\{ \frac{(10\lambda_g^i - \tau_g (Xr)_g^i)^2}{2\tilde{\sigma}^2} \right\}. \tag{2.11}$$

We then take the negative natural log of $\mathcal{L}_g$ to get

$$-\log(\mathcal{L}_g) = \sum_{i=0}^{L_g - 9} -\log\left(\frac{1}{\sqrt{2\pi}\tilde{\sigma}} \right) - \frac{(10\lambda_g^i - \tau_g (Xr)_g^i)^2}{2\tilde{\sigma}^2}. \tag{2.12}$$

We then proceed to take the partial derivative of the function above with respect to τ_g and set it equal to 0 in order to recover its optimal value.

$$-\frac{\partial \log(\mathcal{L}_g)}{\partial \tau_g} = \sum_{i=0}^{L_g - 9} \frac{-1}{2\tilde{\sigma}^2} 2(10\lambda_g^i - \tau_g (Xr)_g^i)(-(Xr)_g^i) = 0. \tag{2.13}$$

Canceling out constants and simplifying yields

$$\sum_{i=0}^{L_g - 9} 10\lambda_g^i (Xr)_g^i - \tau_g ((Xr)_g^i)^2 = 0. \tag{2.14}$$

Lastly, solving for τ_g yields the final equation

$$\tau_g = \frac{\sum_{i=0}^{L_g - 9} 10\lambda_g^i (Xr)_g^i}{\sum_{i=0}^{L_g - 9} ((Xr)_g^i)^2}. \tag{2.15}$$

We have now accomplished a procedure for recovering τ_g given r, alongside a procedure for recovering τ_g given r. In order to initiate the coordinate ascent procedure, we start with a heuristic value for τ_g, set equal to $\frac{1}{L_g-9}\sum_{i=0}^{L_g-9}\lambda_g^i$ (in essence, the mean value of the λ_g vector). We can proceed with the coordinate ascent procedure until we reach convergence in estimates for r and τ_g. Empirically, we have found that only one iteration each of inferring r from τ_g and inferring τ_g from r is needed to reach convergence.

2.2 Variable Definitions

All variable definitions can be found in Table 2.1 below. All of the variable definitions (with the exception of ℓ, r, r_c^i) can be subscripted with the letter g to make them gene-specific. Capitalized Greek letters denote homogenized versions of their (heterogeneous) lowercase counterparts, and variables with a hat over them have been reconstructed using the inferred elongation rates.

Table 2.1: Variable Definitions

Variable	Definition
ℓ	constant that equals 10, denoting the width (in codons) of a typical ribosome footprint
C	codon sequence
C_g^i	codon located at CDS position i of gene g
L	length (in codons) of a particular gene
ρ	vector of smoothed ribosome densities, gathered from Ribo-seq and RNA-seq experimental data
ρ_0	first entry in ρ vector
ρ_1	last entry in ρ vector
λ	vector of smoothed elongation rates, recovered from applying the TASEP model onto ρ vectors
λ_g^i	smoothed elongation rate at window position i of gene g
λ_0	first entry in λ vector
λ_1	last entry in λ vector
$\lambda_{\min}$	minimum entry in λ vector
τ	time-scaling factor. Normalizing by this gene-specific constant brings translation rates into the homogeneous setting
Λ	defined as being equal to λ/τ, it is the homogenized version of λ
r	vector of codon and CDS position-specific elongation rates, returned by elongation inference procedure
r_c^i	elongation rate for codon c at CDS position i
$\hat{\Lambda}$	predicted values for Λ vector, where predictions are made by averaging codon and CDS position-specific elongation rates from the r vector within consecutive windows of 10 codons
$\hat{\rho}$	predicted values for ρ vector, where predictions are made by using the TASEP model to invert $\hat{\Lambda}$
α	heterogeneous initiation rate
A	homogenized initiation rate, defined as α/τ
β	heterogeneous termination rate
B	homogenized termination rate, defined as β/τ
X	design matrix in FUSED-LASSO regression problem
y	vector of labels in FUSED-LASSO regression problem. Defined as the concatenation of all Λ vectors, times ℓ.

Chapter 3

Results

3.1 Validating Accuracy of Inferred Elongation Rates

In order to assess the accuracy of our protocol, we used our inferred codon-specific and CDS position-specific elongation rates r to predict ρ vectors for each gene. Then we correlated predicted ribosome densities vs experimentally observed ribosome densities ρ in the *S. cerevisiae* dataset gathered from [4] in Figure 3.1. The model was trained on 80% of the genes from the dataset, and model performance was evaluated on the remaining held out 20% of genes.

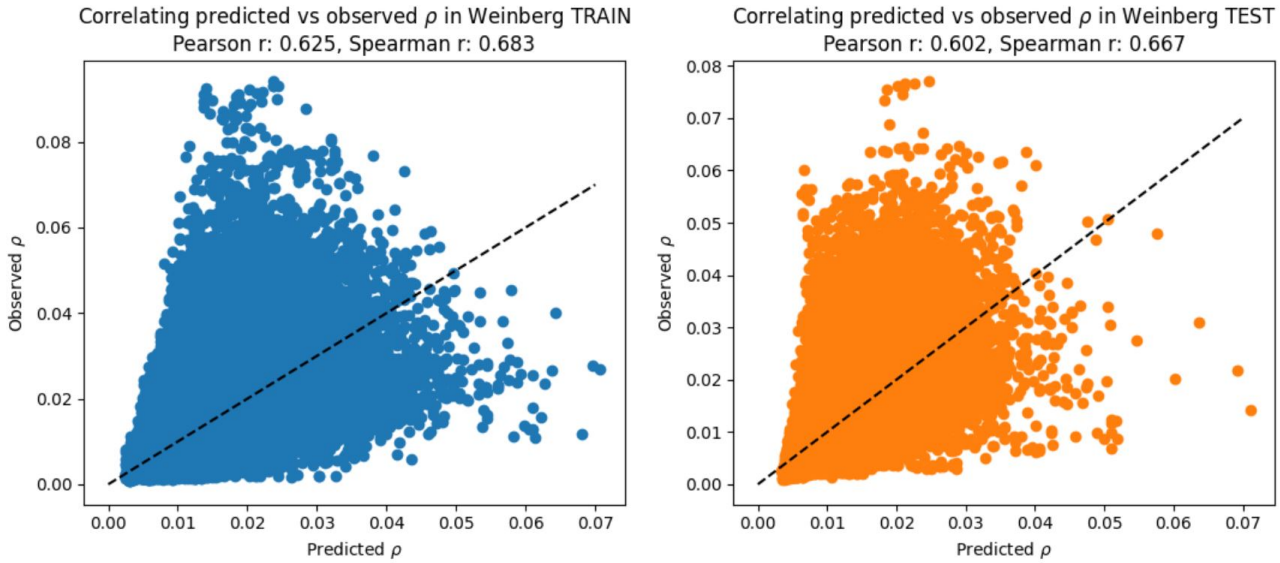

Figure 3.1: Correlating predicted vs ground truth ribosome densities ρ in a train (left) and test (right) subset of the Weinberg et al. dataset. Elongation rates were calculated using the FUSED-LASSO procedure detailed in the methods section with $\gamma = 1$

The black dashed lines in Figure 3.1 above indicates the $y = x$ diagonal. Furthermore, there exists a 0.625 Pearson correlation between predicted and experimentally observed ρ in the train dataset, and a 0.602 Pearson correlation in the test dataset. We find these correlations to be highly encouraging, as they supercede the performance reported by state-of-the-art neural network models on the same dataset [19].

While the previous plot demonstrates good correlation amongst predicted vs experimentally observed ribosome densities that have been aggregated together across genes, it leaves unclear whether or not there exists good correlation at a gene-by-gene level. Therefore, we sought to assess such correlation in Figure 3.2.

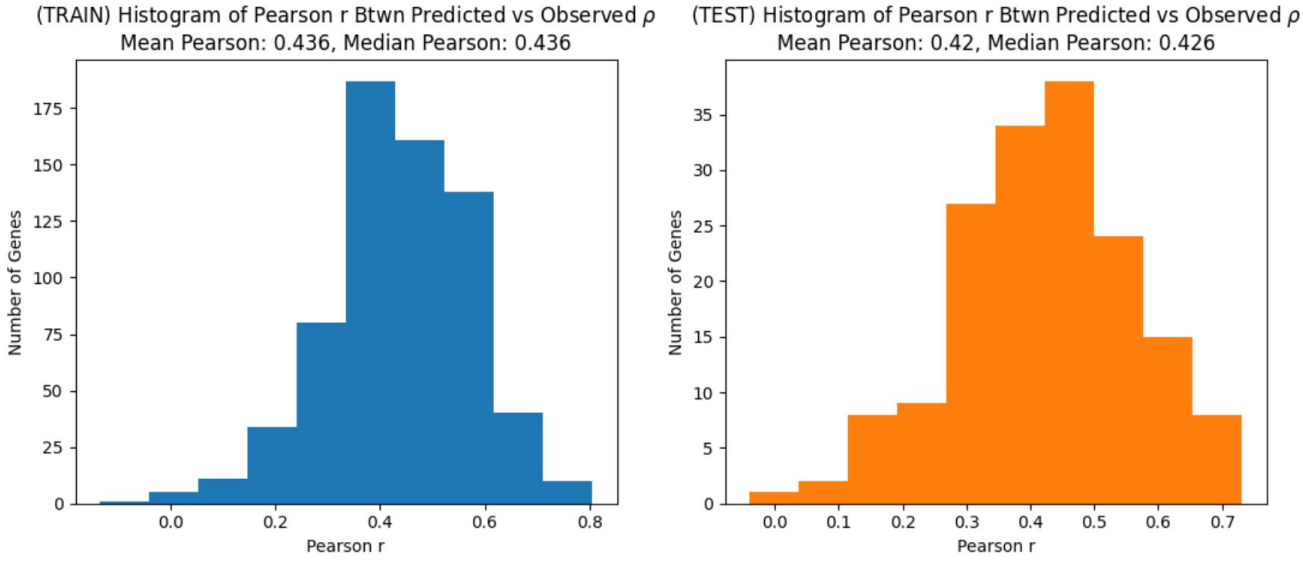

Figure 3.2: Histogram of predicted vs ground truth ribosome densities ρ, on a gene-by-gene basis, in a train subset (left) and test subset (right) of the Weinberg et. al dataset. Elongation rates were calculated using the FUSED-LASSO procedure detailed in the methods section with $\gamma = 1$

Each entry in the histograms above corresponds to the Pearson correlation between the predicted vs experimentally observed ρ vectors for one particular gene, in either the train subset (left) or test subset (right). In the train dataset, the mean and median Pearson correlations were both 0.436. In the test dataset, the mean Pearson correlation was 0.42 and the median Pearson correlation was 0.426. Once again, we assess this to be a fairly strong performance.

Another method for validating the accuracy of our method involves using the inferred elongation rates to predict $\hat{\Lambda}$ and $\hat{\rho}$ for each gene, and comparing these predictions against experimentally observed ρ (and correspondingly inferred Λ, see equation (2.1)). Residual vectors were calculated by subtracting $\hat{\Lambda}$ from Λ, and subtracting $\hat{\rho}$ from ρ. Mean residual values per window position were then calculated by taking the mean of residual values across

all genes, within each window position, in the test subset of the Weinberg et. al dataset [20]. Results are depicted in Figure 3.3 below.

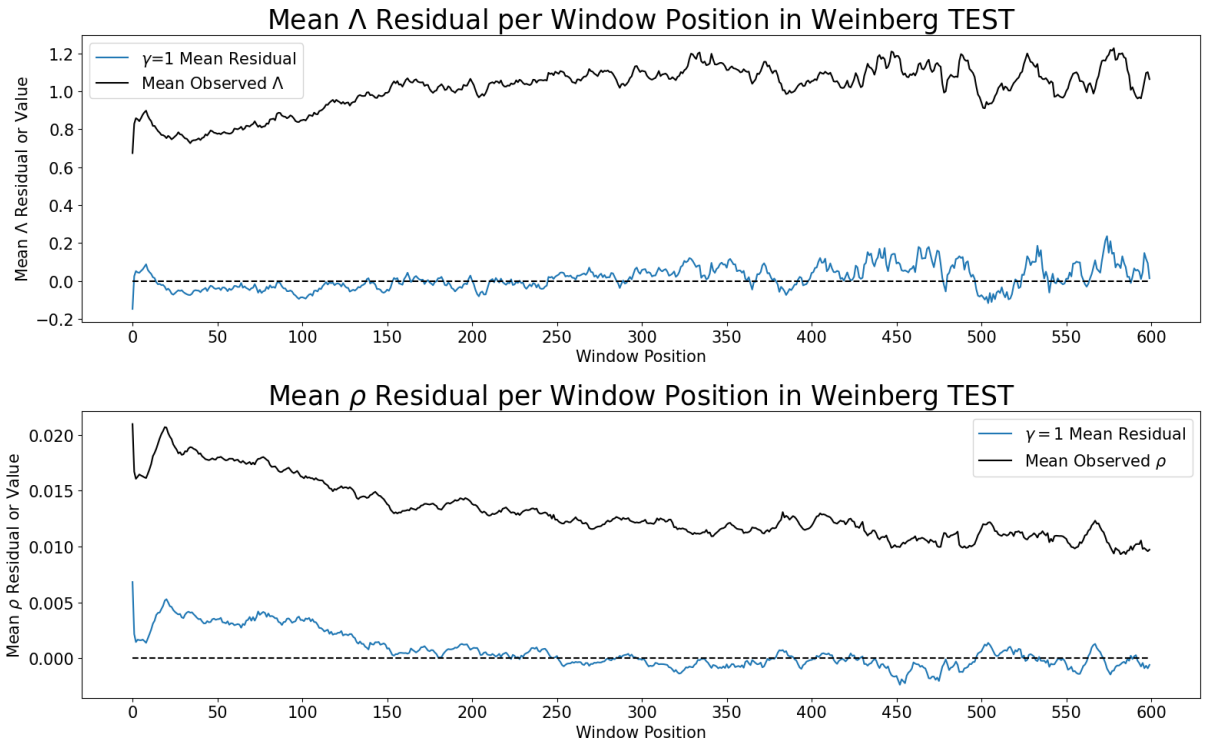

Figure 3.3: (Upper) Comparing mean Λ residual per window position (blue line) vs mean Λ value per window position (black line) in a test subset of the Weinberg et. al dataset. (Lower) Comparing mean ρ residual per window position (blue line) vs mean ρ value per window position (black line) in a test subset of the Weinberg et. al dataset. Elongation rates here were calculated using the FUSED-LASSO procedure detailed in the methods section with $\gamma = 1$.

Although the mean Λ residual is somewhat noisier within the first 10 windows of the CDS and the latter 3' regions of the CDS, it is fairly stable and unbiased throughout much of the CDS as desired. It should be noted that the early 5' region of a gene is typically where regulation of translation occurs [8] [10], so it may not be altogether surprising that codon identity and CDS position alone are insufficient to precisely predict elongation rates in this region. Furthermore, there are fewer genes and therefore less data in the latter regions of the CDS, which may contribute to the greater amount of noise observed in that region. Similarly, the mean ρ residuals appear to exhibit a slight positive bias within the first 150 window positions of the CDS, before stabilizing and becoming fairly unbiased in latter regions of the CDS that are further downstream. This may be attributable to the non-negativity constraint from our convex optimization procedure, which removes the statistical guarantee

of unbiasedness. The black lines in the figure above allow us to compare the magnitude of the mean residuals per window position with the magnitude of the mean experimentally observed ρ or Λ. In both cases, the former is comfortably smaller than the latter.

In addition to conducting this analysis in yeast, we sought to assess how the model performs with Riboseq and RNA-sequencing data gathered from vertebrate species. The results can be seen in Figure 3.4:

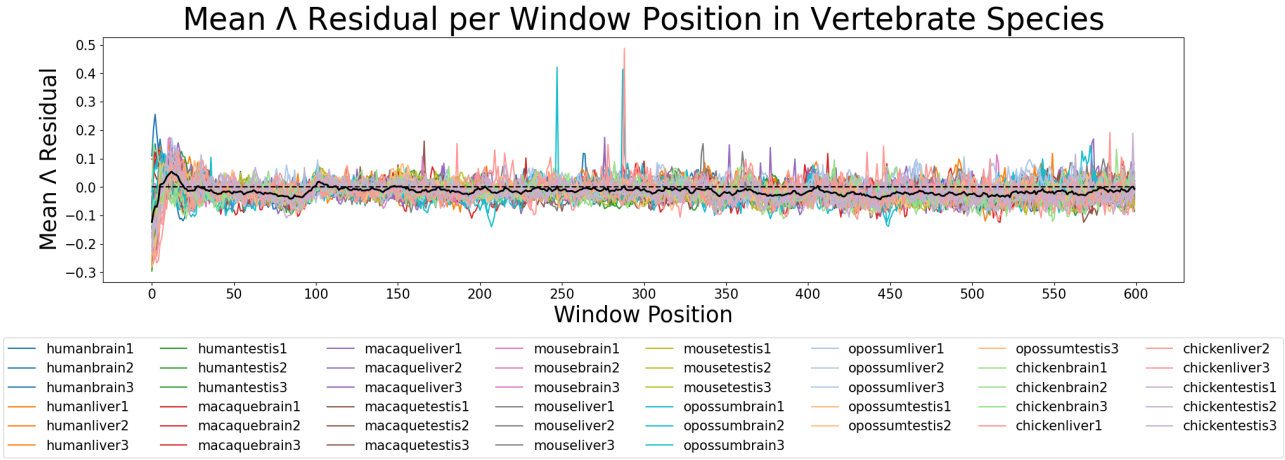

Figure 3.4: Mean Λ residual per window position across various vertebrate species datasets. Elongation rates here were calculated using the FUSED-LASSO procedure detailed in the methods section with $\gamma = 1$.

Each colored line in the plot above indicates the mean Λ residual vector per window position for a given vertebrate species/tissue pair. The central black line indicates the mean across all of the colored lines. With the slight exception of the first 10 window positions, the black line indicates that mean Λ residuals remain remarkably stable and unbiased throughout the bulk of the CDS.

While calculating mean Λ residual vectors for each vertebrate species/tissue pair, we observed an interesting trend, shown in Figure 3.5 below.

Once again, the central black lines in each plot indicate the mean across all of the colored lines. The trends observed in the mean Λ residual per window position vectors appear to segregate distinctly depending upon tissue of origin. The trends in liver and testis look similar to one another, whereas the trends in brain look distinct. Furthermore, in Figure 3.5 we confined the range of the x-axis to only examine window positions 0-40, since this is where the most noise typically appears in the Λ residual vectors. More future work is needed to understand why the trends above exhibit segregation by tissue type.

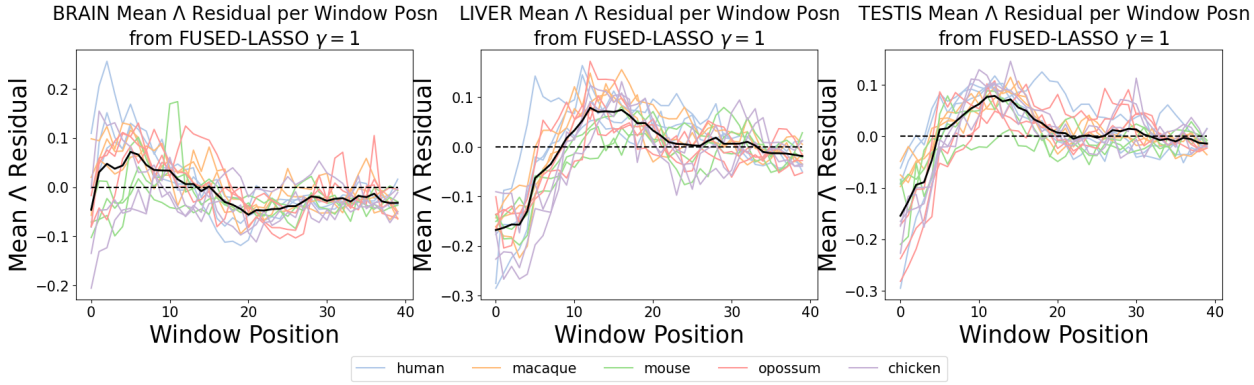

Figure 3.5: Mean Λ residual per window position across various vertebrate species, segregated by tissue type. Elongation rates here were calculated using the FUSED-LASSO procedure detailed in the methods section with $\gamma = 1$.

3.2 Variation in Λ Is Attributable to Both Synonymous Codon Choice and Amino Acid Choice

Due to the lack of precise elongation rate estimates that existed prior to our work, disentangling the relative contributions of synonymous codon choice vs amino acid choice in regulating variance in smoothed elongation rates Λ has remained elusive. The codon-specific elongation rates returned by our novel protocol have allowed us to probe this question in more depth by quantifying the fraction of variance in Λ that is explained by each. Figure 3.6 below visualizes the range of elongation rates achieved by various amino acids and synonymous codons in CDS position 50 of the Weinberg et. al dataset.

As indicated in Figure 3.6, there exists a considerable amount of variance in elongation rate across amino acids, and across synonymous codons that code for the same amino acid. It is also worth mentioning that the scatterplot in Figure 3.6 is merely a representative example; similar patterns can be easily observed in other CDS positions and within other vertebrate species/tissue pairs.

While the previous figure qualitatively confirmed that variation in elongation rate can be attributable to either synonymous codon choice or amino acid choice, we also sought to quantitatively disentangle relative contributions from the two. In order to do so, we devised a variance decomposition method based upon the Law of Total Variance. For reference, the Law of Total Variance for two arbitrary random variables X, Y is stated below:

$$\mathrm{Var}_Y(Y) = \mathbb{E}_X[\mathrm{Var}_Y(Y|X)] + \mathrm{Var}_X(\mathbb{E}_Y[Y|X])$$

. If we conceptualize X as a random variable representing amino acid identity at some codon position, and Y as representing the inferred elongation rate at that same codon position,

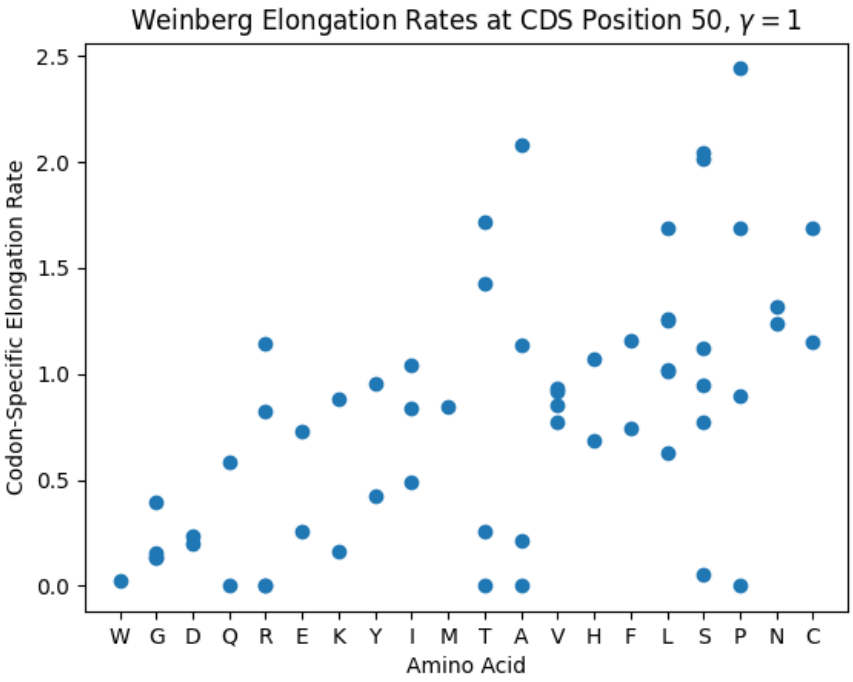

Figure 3.6: Weinberg codon-specific elongation rates at CDS position 50. The amino acids presented on the x-axis are sorted in order of increasing mean elongation rate. Elongation rates here were calculated using the FUSED-LASSO procedure detailed in the methods section with $\gamma = 1$.

then it becomes clear that the first term $\mathbb{E}_X[\mathrm{Var}_Y(Y|X)]$ represents the variance in inferred elongation rates that is attributable to synonymous codon choice, and the second term $\mathrm{Var}_X(\mathbb{E}_Y[Y|X])$ represents the variance in inferred elongation rates that is attributable to amino acid choice. Both of these quantities are plotted in Figure 3.7 below.

Figure 3.7 not only indicates that there is a greater magnitude of variance in inferred elongation rates within early regions of the CDS (roughly CDS positions 0-25), but it also indicates that synonymous codon choice plays a more prominent role in regulating elongation rates within this early region of the CDS, as compared to latter regions.

We next sought to extend this variance decomposition analysis to other vertebrate species/tissue pairs to see if similar trends persist outside of the human brain dataset. We also extended this analysis by decomposing variance in smoothed elongation rates $\hat{\Lambda}$, which are dependent on the identity of the 10 amino acids within a given window. Once again, starting from the Law of Total Variance:

$$\mathrm{Var}_Y(Y) = \mathbb{E}_X[\mathrm{Var}_Y(Y|X)] + \mathrm{Var}_X(\mathbb{E}_Y[Y|X])$$

. Setting X equal to the identity of the 10 amino acids within some window, and setting

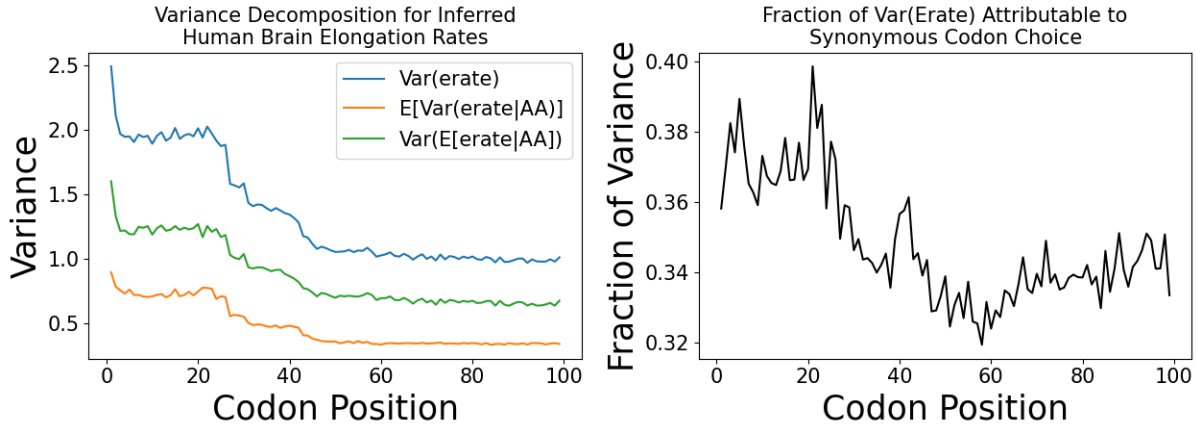

Figure 3.7: (Left) Decomposition of variance amongst inferred human brain elongation rates. The blue line indicates total variance in inferred elongation rate per codon position, the orange line indicates variance attributable to synonymous codon choice, and the green line indicates variance attributable to amino acid choice. (Right) Fraction of variance attributable to synonymous codon choice. This is calculated as the ratio of the values in the orange line over the values in the blue line.

Y equal to the predicted smoothed elongation rate $\hat{\Lambda}$ within that same window allows us to assess the fraction of variance in $\hat{\Lambda}$ that is attributable to synonymous codon choice. The results of this analysis are shown for each species/tissue pair in Figure 3.8.

With a few exceptions, the trends observed in Figure 3.8 generally corroborate the conclusions made from the previous figure – that synonymous codon choice plays a more prominent role in regulating elongation rates within early regions of the CDS.

3.3 Codon Usage Bias Optimizes Translation Efficiency in Early Regions of the CDS

Codon usage bias has been observed to occur across many different biological organisms. Furthermore, previous literature in the field has reported that amongst other purposes, codon bias serves the useful purpose of optimizing translation efficiency by tuning elongation rates, which has impacts on gene expression [16] [2] [15] [18] [23] [7]. Using the codon-specific elongation rates inferred from our procedure, we sought to test this hypothesis quantitatively.

In order to address this question, we started by computing absolute and synonymous codon usage frequencies for each of the 64 codons in each of the first 100 CDS positions. The absolute codon usage frequency of codon c at CDS position x is defined as the fraction of codons at position x in translated regions of the genome that are equal to c. Thus, the sum of absolute codon usage frequencies for all 64 codons c at a given CDS position x equals

1. Synonymous codon usage frequencies are simply a scaled version of the absolute codon usage frequencies. They are scaled in such a way that for any set C of synonymous codons coding for the same amino acid, the sum of synonymous codon usage frequencies across all elements of C equals 1.

If codon usage bias serves the purpose of optimizing overall translation efficiency, then it would make sense for faster codons to generally be used more frequently. Therefore, we produced scatterplots correlating codon elongation rate vs absolute codon usage frequency or synonymous codon usage frequency. The results are shown below in Figure 3.9.

We were inspired by the statistically significant positive Spearman correlation seen in the left panel of Figure 3.9, as it indicated that faster codons do tend to be used more frequently within the first 20 codon positions of the CDS. Similarly, in the right panel of Figure 3.9, there was a 0.256 Spearman correlation between elongation rate and synonymous codon usage frequency for codons that coded for 2-fold degenerate amino acids, 0.457 Spearman correlation for codons that coded for 4-fold degenerate amino acids, and a 0.527 Spearman correlation for codons that coded for 6-fold degenerate amino acids. Therefore, it seems that as the fold degeneracy of an amino acid increases, the tendency to use faster variants of the synonymous codons becomes more prominent.

We then wanted to gain a better understanding of the locality of this phenomenon – are faster codons used more frequently throughout the entire CDS, or is this trend only confined to certain regions of the CDS? To address this question, we calculated the Spearman correlation between inferred elongation rates r and absolute codon usage frequency at each CDS position x in the range 0 through 99 inclusive. The results of this analysis are shown in Figure 3.10 below:

In Figure 3.10, the blue line indicates Spearman correlation value, the orange line indicates the associated p-value of the correlation, and the red dashed line indicates the 0.05 threshold cutoff for significance. Clearly, the positive correlation between elongation rate and absolute codon frequency is strongest within the first 10 codon positions, before decaying in strength and becoming statistically insignificant past CDS position 20. This is consistent with biological expectation, as the elongation rates in the first 10 CDS positions determine the value for λ_0, which is known to play a critical role in regulating the current J of ribosomes trafficking through a given position in time [5].

While the previous figure makes clear that faster codons are used more frequently within the first 10-20 codon positions of the CDS, it remains unclear whether or not synonymous codon usage bias plays a role in manifesting this trend. In other words, amongst synonymous codons coding for the same amino acid, are faster synonymous codons used more frequently than slower synonymous codons to promote faster translation? To address this question, we first needed to calculate codon usage frequencies solely within the λ_0 windows across all genes. After computing these, we produced scatterplots of inferred elongation rates r vs codon usage frequencies in λ_0 for each amino acid separately. Points are color-coded based upon the identity of the synonymous codon, and results are shown in Figure 3.11.

As can be observed within each window panel of Figure 3.11, there typically exists a strong positive Spearman correlation between inferred elongation rates r and absolute codon

usage frequency, which confirms our suspicion that synonymous codon usage bias plays a prominent role in selecting for the usage of faster codons within the first 10 CDS positions. There are a few exceptions to this trend, however. Amino acids that are large and nonpolar (such as F, I, K, V, Y) tend to exhibit a negative correlation instead, and more work is needed to understand why these exceptions exist.

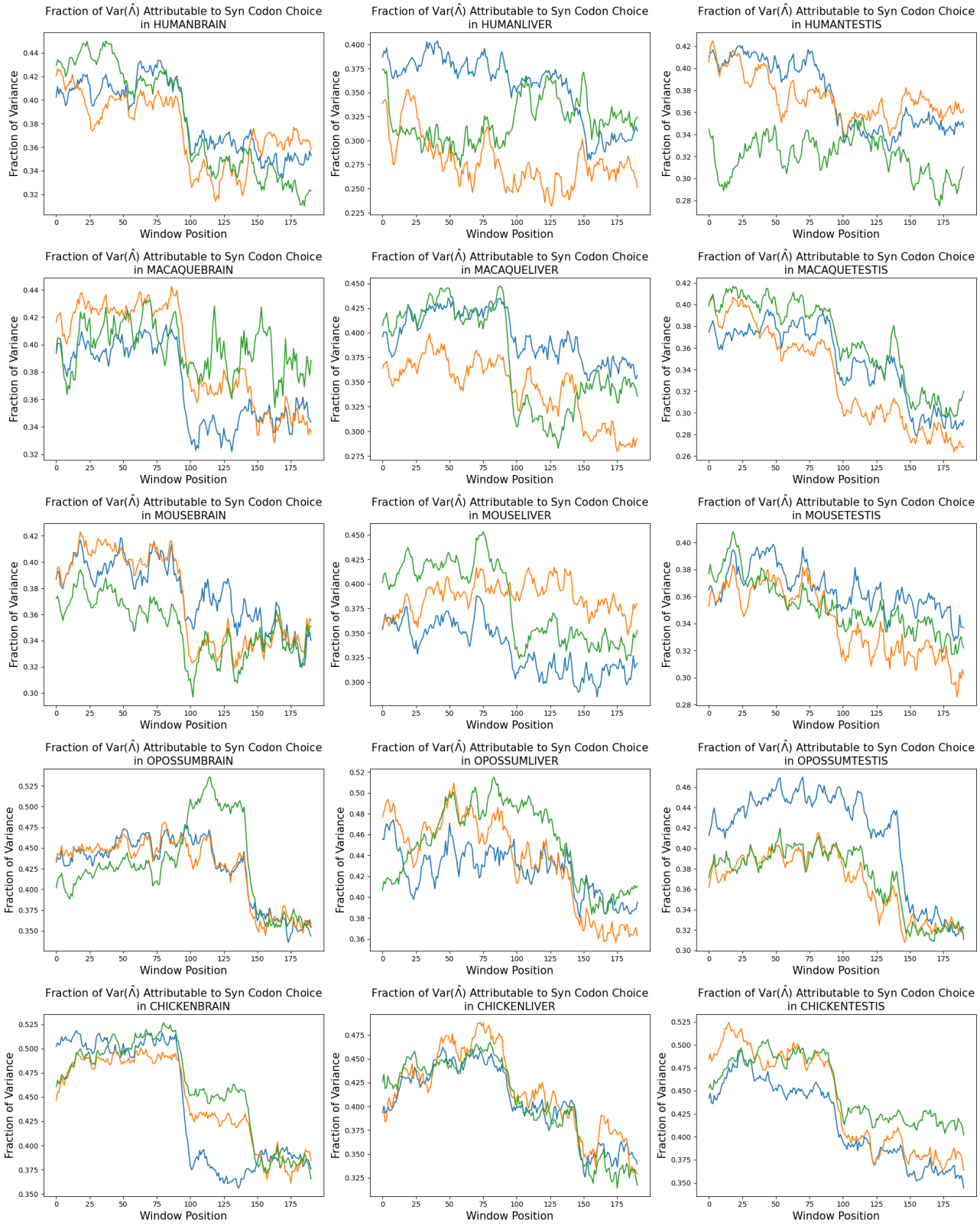

Figure 3.8: Var($\hat{\Lambda}$) Decomposition Across Various Vertebrate Species/Tissue Pairs. The blue, green, and orange lines represent different replicates of the same species/tissue pair.

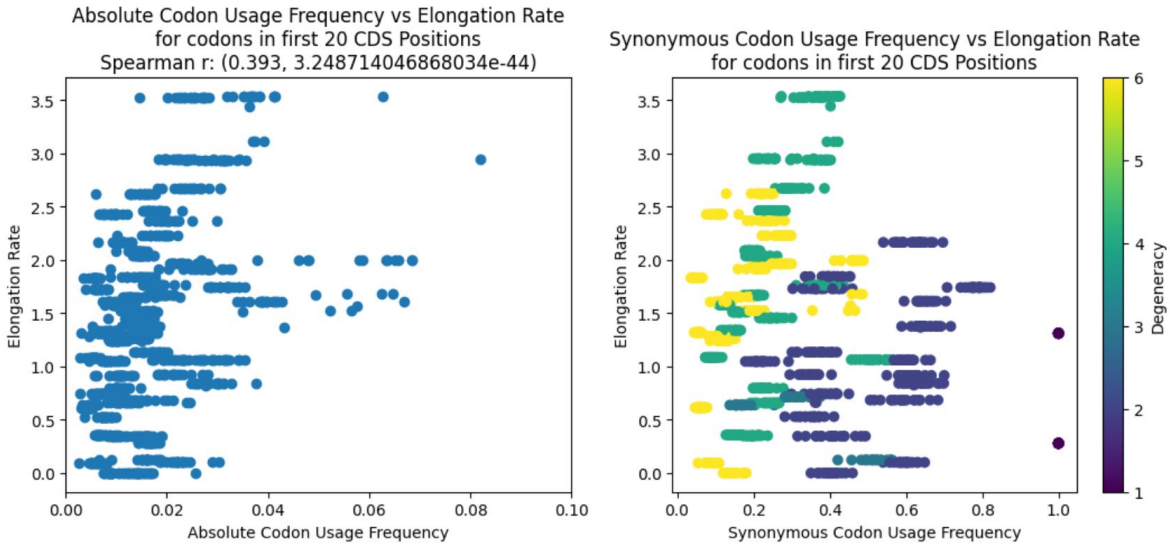

Figure 3.9: (Left) Scatterplot of absolute codon usage frequency vs elongation rate in the first 20 CDS positions. There are a total of 64 $\times$ 20 points in the scatterplot, one for each unique codon in each unique CDS position. (Right) Scatterplot of synonymous codon usage frequency vs elongation rate in the first 20 CDS positions, with point color corresponding to the degeneracy of the corresponding amino acid.

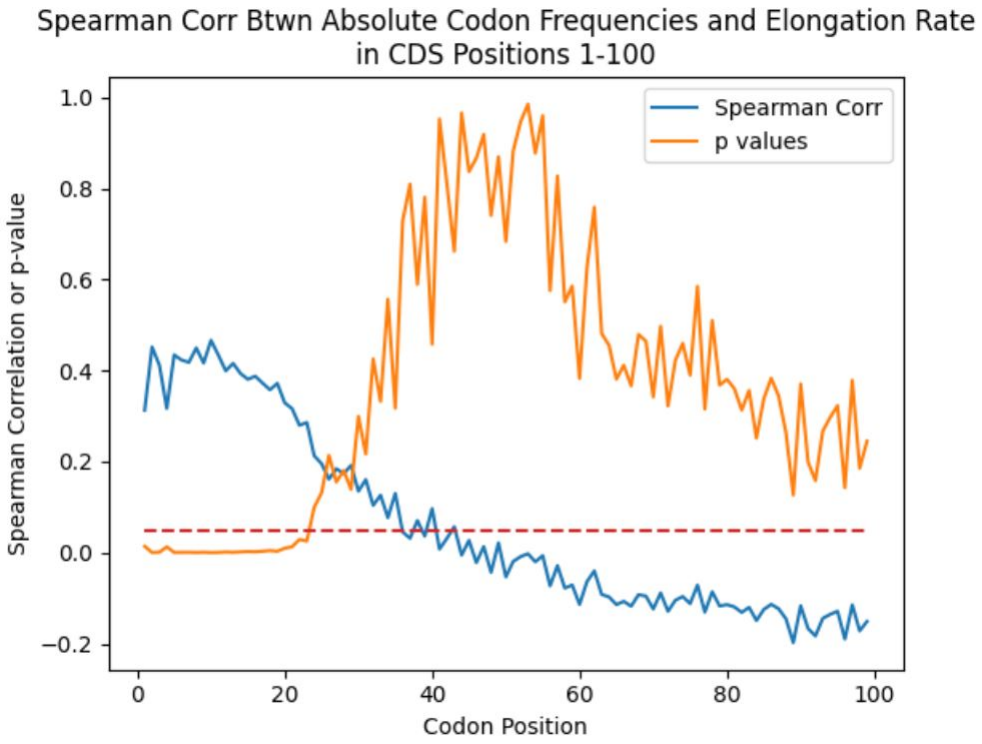

Figure 3.10: Spearman correlations and their corresponding p-values between inferred elongation rate and absolute codon usage frequency at each CDS position in the range 0 through 99 inclusive.

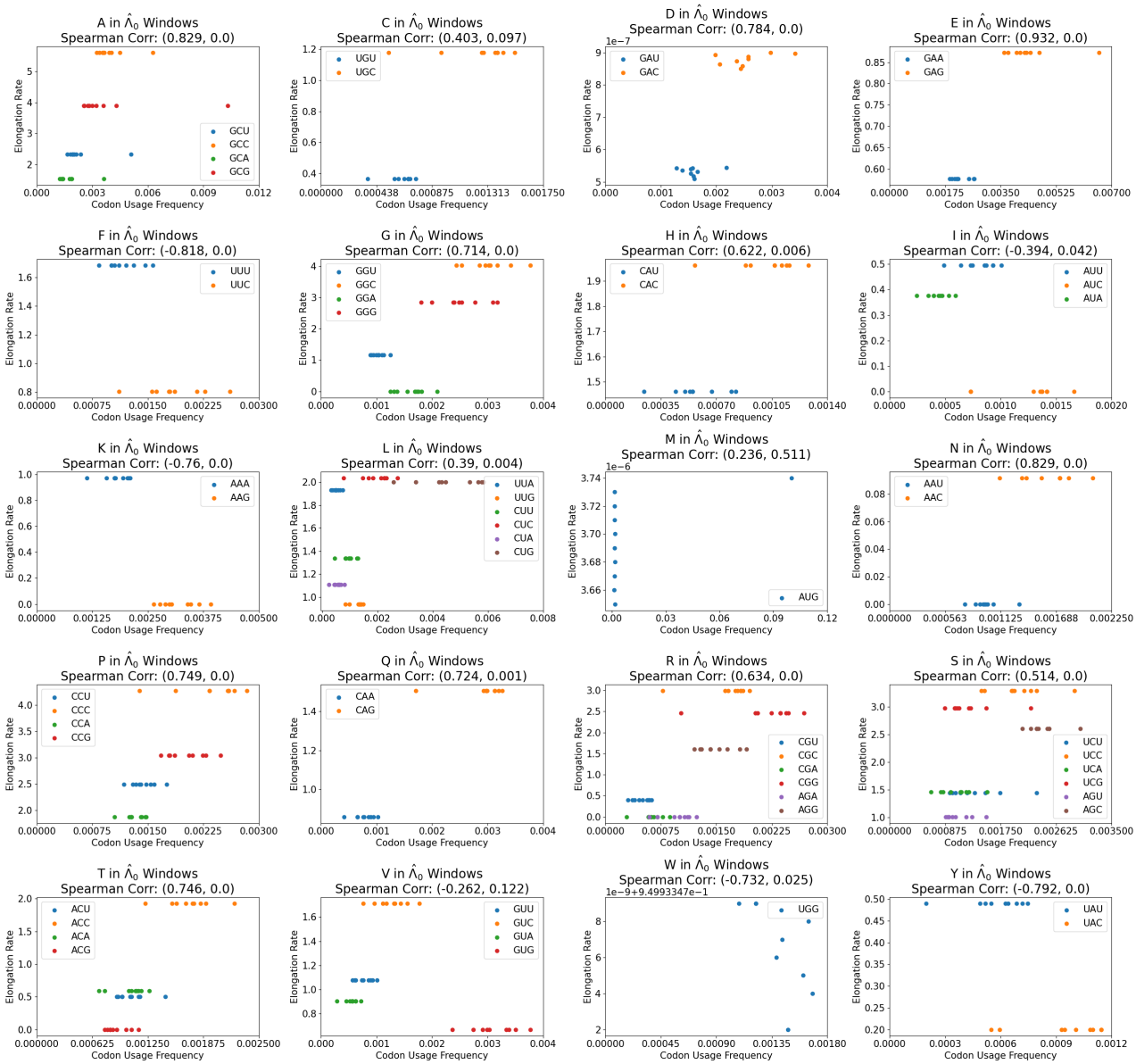

Figure 3.11: Scatterplots of inferred elongation rates r vs absolute codon usage frequency in the first 10 CDS positions, for each of the 20 amino acids. Points are color-coded based on synonymous codon identity. Trends for the M and W amino acids should be ignored as they are non-degenerate.

Chapter 4

Conclusion

4.1 Summary

We developed a novel algorithm for inferring codon-specific and CDS position-specific elongation rates r from Riboseq and RNA-sequencing data. By using these inferred elongation rates to predict λ and ρ vectors for genes, we found that predictions recovered from our inferred elongation rates exhibited strong correlations with their experimentally observed ground-truth counterparts, and that our method was capable of outperforming state-of-the-art machine learning methods on the task of predicting ribosome densities [19]. Using the inferred elongation rates, we were able to confirm qualitatively that elongation rates vary widely across different amino acids and different synonymous codons that code for the same amino acid. Furthermore, by employing a variance decomposition approach, we were able to quantitatively disentangle relative contributions from synonymous codon usage bias and amino acid choice in explaining observed variance amongst smoothed elongation rates $\hat{\Lambda}$. Using our inferred elongation rates, we also found that faster codons tend to be used more frequently in early regions of the CDS, particularly within the first 10 codon positions that regulate λ_0.

4.2 Future Work

Work is currently in progress on implementing a new version of the elongation rate protocol. This newest version models the Riboseq and RNAseq counts as independent Poisson random variables whose rates both depend on a common ground truth (but unobserved) quantity of mRNA transcripts that exist for a particular gene within the cell. Similar to before, the new protocol then uses statistical MLE inference to recover codon-specific and CDS position-specific elongation rates. Preliminary results from using this approach suggest that it is even more powerful than the original version presented in this project, in that it is capable of achieving an even stronger correlation between predicted and experimentally observed ρ, λ vectors and removes the occurrence of codons with zero elongation rate. We look forward to

running this new protocol on all vertebrate and yeast datasets in the near future and seeing whether the new results are capable of providing even stronger evidence to bolster biological conclusions we made.

In addition, we would like to correlate protein abundance levels (as determined by mass-spectrometry) with strength of codon usage bias in order to lend further support to our claim that codon usage bias optimizes translation efficiency. For example, we hypothesize that genes whose protein products are more highly expressed should exhibit a greater Spearman correlation between codon usage frequency and elongation rate (in other words, exhibit a greater tendency to use faster codons more frequently).

Future experiments will also study the role played by positional information in regulating elongation rates and therefore translation dynamics. For example, if one were to randomly sample codons in order to form a protein sequence, and use the inferred elongation rates to predict $\hat{\rho}$ and $\hat{\Lambda}$, would the associated ribosome densities and smoothed elongation rates for this synthetic gene still exhibit common features of translation such as the 5' ramp? Similarly, variance decompositions could be performed to assess the contributions of synonymous codon choice, amino acid choice, and positional information in explaining observed variance amongst smoothed elongation rates.

Bibliography

[1] Ariel A Bazzini et al. "Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition". In: *EMBO Journal* 35.19 (2016).

[2] Susanta K. Behura and David W. Severson. "Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes". In: *Biological Reviews* 88.1 (2012).

[3] Catherine A. Charneski and Laurence D. Hurst. "Positively Charged Residues Are the Major Determinants of Ribosomal Velocity". In: *PLOS Biology* 11 (2013).

[4] Khanh Dao Duc and Yun S. Song. "The impact of ribosomal interference, codon usage, and exit tunnel interactions on translation elongation rate variation". In: *PLos Genetics* 14.8 (2018), e1007620.

[5] Dan D. Erdmann-Pham, Khanh Dao Duc, and Yun S. Song. "The key parameters that govern translation efficiency". In: *Cell Systems* 10.2 (2020), pp. 183–192.

[6] Justin Gardin et al. "Balanced Codon Usage Optimizes Eukaryotic Translational Efficiency". In: *PLOS Genetics* 8 (2012).

[7] Justin Gardin et al. "Measurement of average decoding rates of the 61 sense codons in vivo". In: *eLife* 3 (2014).

[8] Fátima Gebauer and Matthias W. Hentze. "Molecular mechanisms of translational control". In: *Nature Reviews Molecular Cell Biology* 5 (2004).

[9] Hani Goodarzi et al. "Modulated Expression of Specific tRNAs Drives Gene Expression and Cancer Progression". In: *Cell* 165.6 (2016).

[10] John W.B. Hershey, Nahum Sonenberg, and Michael B. Mathews. "Principles of Translational Control: An Overview". In: *Cold Spring Harbor Perspectives in Biology* 4.12 (2012).

[11] Nicholas T. Ingolia et al. "Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling". In: *Science* 324.5924 (2009).

[12] Ryuta Ishimura et al. "Ribosome stalling induced by mutation of a CNS-specific tRNA causes neurodegeneration". In: *Science* 345.6195 (2015).

[13] Sebastian Kirchner et al. "Alteration of protein function by a silent polymorphism linked to tRNA abundance". In: *PLOS Biology* 15.5 (2017).

[14] Carolyn T. MacDonald, Julian H. Gibbs, and Allen C. Pipkin. "Kinetics of biopolymerization on nucleic acid templates". In: *Biopolymers* 6.1 (1968).

[15] Sujatha Thankeswaran Parvathy, Varatharajalu Udayasuriyan, and Vijaipal Bhadana1. "Codon usage bias". In: *Molecular Biology Reports* 49.1 (2022).

[16] Joshua B. Plotkin and Grzegorz Kudla. "Synonymous but not the same: the causes and consequences of codon bias". In: *Nature Reviews Genetics* 12 (2010).

[17] Vladimir Presnyak et al. "Codon optimality is a major determinant of mRNA stability". In: *Cell* 160.6 (2015).

[18] Tessa E.F. Quax et al. "Codon Bias as a Means to Fine-Tune Gene Expression". In: *Molecular Cell* 59.2 (2015).

[19] Robert Tunney et al. "Accurate design of translational output by a neural network model of ribosome distribution". In: *Nature Structural Molecular Biology* 25 (2018), pp. 577–582.

[20] David E. Weinberg et al. "Improved Ribosome-Footprint and mRNA Measurements Provide Insights into Dynamics and Regulation of Yeast Translation". In: *Cell Reports* 14.7 (2016), pp. 1787–1799.

[21] Fangzhou Zhao, Chien-Hung Yu, and Yi Liu. "Codon usage regulates protein structure and function by affecting translation elongation speed in Drosophila cells". In: *Nucleic Acids Research* 45.14 (2017).

[22] Evgeny Leushkin Zhong-Yi Wang et al. "Transcriptome and translatome co-evolution in mammals". In: *Nature* 588 (2020), pp. 642–647.

[23] Zhipeng Zhou et al. "Codon usage is an important determinant of gene expression levels largely through its effects on transcription". In: *Proceedings of the National Academy of Sciences* 113.41 (2016).

[24] R. K. P. Zia, J. J. Dong, and B. Schmittmann. "Modeling Translation in Protein Synthesis with TASEP: A Tutorial and Recent Developments". In: *Journal of Statistical Physics* 144 (2011).

[25] Hadas Zur and Tamir Tuller. "Predictive biophysical modeling and understanding of the dynamics of mRNA translation and its evolution". In: *Nucleic Acids Research* 44.19 (2016).