

# A Computational Model for Translational Efficiency and Frameshifts in *Escherichia coli* Using a Genetic Signal Processing Approach

Hao Lian, Vivek Bhattacharya, and Daniel R. Vitek

In genetics, *E. coli* is used as an expression system to commercially produce proteins. However, sequence-dependent features, such as rare codons and codon biases, affect translational efficiency. To tackle this problem, we proposed a stochastic model to computationally estimate translational efficiency and predict frameshifting, uniting ideas from biological literature and developing two predictive metrics. We ran our model on 4364 sequences from the *E. coli* genome and found over 90% of them to have predicted high yields; moreover, the model predicts ribosomal proteins to translate at even higher rates—both results that concur with experimental evidence. We investigated a set of eight sequences of bovine growth hormone, and the model correctly determined the translational efficiency for seven. We then examined variations of *prfB*, a gene with a programmed frameshift; the model grouped these sequences into two general categories of high and low yield, consistent with experimental results. Successful development of a computational model for translational efficiency implicates itself in optimizing recombinant protein yield in multiple fields, including commercial protein synthesis.

## Acknowledgments

Foremost, we thank our parents for their wise judgment in creating us. We would like to thank our mentors, Dr. Donald L. Bizter; Dr. Mladen A. Vouk; and Dr. Anne-Marie Stomp for just about everything. We are grateful to North Carolina State University for their office space and labs. Thank you to Dr. Fred Breidt and Robert Snyder for working with us to design and run the wet lab experiments. Thanks to Dr. Lalit Ponnala for the research that formed the basis of our work. Thanks to Scott Vu kicking our model's tires.

## 1 Introduction

We are part of an ongoing research project investigating the application of bioinformatics and genetic signal processing to better understand how information is encoded to and decoded from nucleic acids. The particular focus of the current research on ribosomal translation is within bacteria. Previous researchers [8] have developed a deterministic model of translational reading frame by studying the programmed frameshift present in the *prfB* gene of *E. coli*. The focus of our studies was to improve the model and apply its vatic powers.

Kozak [6] and Kane [5] studied the impact of sequence-dependent features, especially with regard to codon bias and rare codon usage, on translational efficiency. The importance of these features is especially pronounced in the synthesis of recombinant proteins [14]. In addition, secondary structure problems during protein folding can decrease transla-

tional efficiency [6]. However, scientists do not fully understand the specific connections between efficiency and these sequence-dependent factors, as even eliminating troublesome factors does not always increase efficiency. The goal of our studies is then to advance the development of a computational method that identifies mRNA sequence changes to improve protein yield. If successful, this work streamlines gene sequence optimization in the production of recombinant proteins. The model also addresses the significant challenge of creating cell lines that synthesize proteins at commercial yields in fields from medicine to agriculture. In addition, a successful model implies a mechanism by which molecular biologists test the stable maintenance of elongation.

Prior to our work, Ponnala et al. [9] created a deterministic model of frameshifting based on the hybridization between the 16S rRNA tail and mRNA nucleotides. The periodicity of this signal [9] suggests that a force emerges from the free energy of hybridization and acts as a mechanism to stabilize reading frame. We discuss limitations of a deterministic model in Section 3.1.

We assume that, at each cycle of ribosome translation, environmental noise results in the ribosome's imprecise alignment with the next codon to be translated. To capture this idea, we created a new metric of efficiency: the total deviation from the intended reading frame. We found this metric roughly correlate with translational efficiency (Section 4.1). Throughout this paper, we use this and other metrics to optimize tRNA availabilities in order to distinguish between high and low efficiency sequences. We also use our new metric

to optimize the performance of a translationally regulated gene and to test the robustness of our model by running ribosomal proteins and the *E. coli* genome.

## 2 Prior Model

### 2.1 Free Energy

Hybridization between two RNA sequences changes the free energy in a cell, occurring between the rRNAs, the mRNA, and the tRNAs [15]. Shine and Dalgarno [13] observed that the 3' end of the 16S rRNA is complementary to a sequence found directly upstream of the start codon of many prokaryotic mRNAs; they hypothesized that RNA hybridization plays an important role in translation initiation, later experimentally confirmed [2, 4].

Weiss et al. [16] subsequently observed that changes in the 16S tail can significantly change the frequency of frameshifting in *prfB*, an *E. coli* gene known to frameshift at the 25<sup>th</sup> codon. These results suggest that the 16S tail is positioned to interact with the mRNA during translation and elongation. Yusupova et al. [17] supported the spatial accessibility of the 16S tail with the mRNA with X-ray crystallography data.

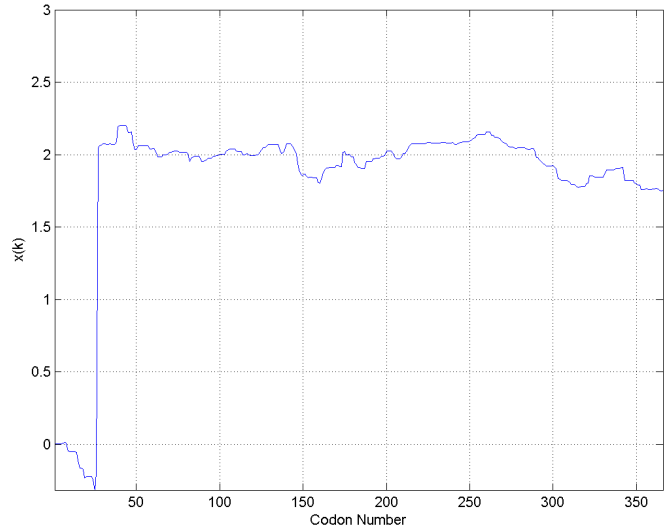
Freier et al. [1] proposed a thermodynamic model for calculation of free energy values, modeling the hybridization of permutations of pairs of consecutive RNA nucleotides.

### 2.2 Deterministic Model

Ponnala et al. [8] assumed a sinusoidal model of free energy against base-pair position because a Fourier transform of free energy indicated a strong component with a period of one codon. In their model, free energy projects onto magnitude and phase through a memory model that stores three values in a phasor, a concept from physics. Ponnala et al. represented this cumulative energy phasor at codon  $k$  as  $\mathbf{V} = M e^{i\theta}$ , where  $i$  is the imaginary constant, from which they calculate the magnitude and phase, modeled on a polar plot, via trigonometry. Differentiating the energy phasor with respect to distance along the mRNA strand gives vector  $\mathbf{D}$ , the force assumed by the model [8] to act on the ribosome to keep the mRNA in the reading frame.

The time that the force acts on the ribosome depends upon the tRNA availability associated with the codon at the A-site and interactions between said tRNA and the ribosome [7].

FIGURE 1. Plots of *prfB*: Deterministic displacement



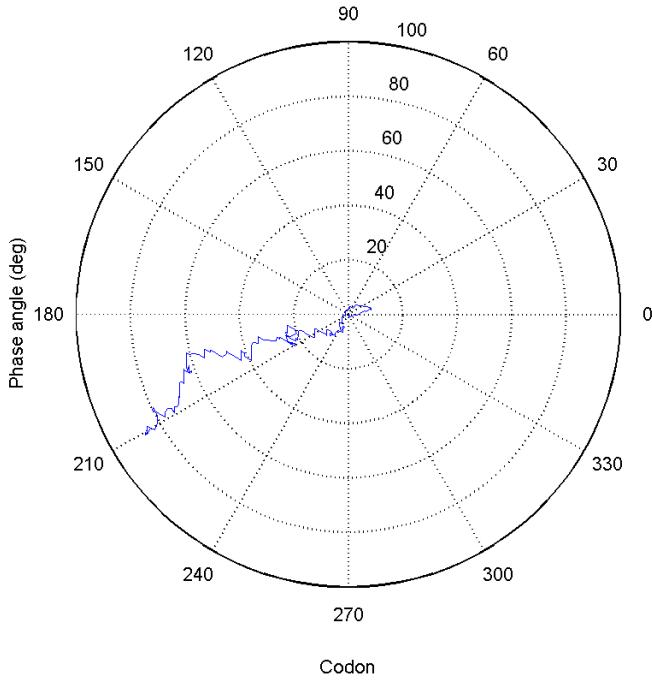
Ponnala et al. represent this with a deterministic model: For each codon, the number of “wait cycles,” a function of the rarity of the tRNA, corresponds to the time the force can increment displacement from the current reading frame. In the deterministic model, the force acts for *exactly* this number of cycles for a given codon. The model then simulates hybridization between the 16S ribosomal subunit and a given mRNA strand: First, the 13-base 16S tail of *E. coli* hybridizes with the first 13 bases of a sequence, which includes a 12-base leader sequence, to determine the free energy value of the first nucleotide of the coding sequence. The algorithm then iteratively calculates the free energy for every nucleotide of the sequence [15].

### 2.3 Frameshifts

Ponnala et al. let a displacement of  $x = 0$  correspond to the zero reading frame and increments of *two* to represent a one-nucleotide change. For example,  $x = 2$  represents the +1 frame. They prove that both  $x = 0$  and  $x = 2$  are stable points, as expected.

A jump from approximately  $x = 0$  to  $x = 2$  in a span of only base pair is then the first indication of a +1 frameshift; it suggests the ribosome skips one entire base pair in the mRNA sequence. Figure 1 shows the displacement plot per this deterministic model for *prfB*, a gene with a unique programmed +1 frameshift, which exists at codon 25. In conjunction with this characteristic plot, a +1 frameshift also displays a characteristic clockwise 120° phase angle rotation

FIGURE 2. Plots of *prfB*: Polar plot



from the species angle, the average phase of the free energy signals of a number of verified *E. coli* genes that stay in frame [8]. We interpret the free energy signal’s alignment with the sudden jump in displacement as a sustained frameshift. Since the free energy signal has a period of one codon [8], a +1 frameshift the free energy signal must undergo a phase shift of a third of an entire period (Figure 2).

### 3 Computational Methods

#### 3.1 Stochastic Displacement Model

As mentioned, the gene *prfB* exhibits a programmed frameshift by jumping in displacement to  $x = 2$ . Certain genes, however, demonstrate ambiguous behavior near  $x = \pm 1$  [8]. In the deterministic model, we lacked the sensitivity needed to clearly discern a programmed frameshift. Worse, the model may not show these unstable behaviors at all. The nondeterministic behavior of translation and the presence of noise limits the model’s ability to achieve better resolution. Therefore, we revamped the model to be stochastic through the incorporation of sinusoidal probability.<sup>1</sup>

At each wait-cycle in elongation, we propose the ribosome makes a decision: stay in the current reading frame, move

<sup>1</sup> The model’s code is available online with detailed documentation.

to the  $\pm 1$  reading frames, or proceed to the next cycle. In addition, because the number of wait cycles is inversely proportional to the tRNA availability of the codon in the current reading frame (the A-site) [3, 8], rarer codons force the ribosome to wait longer for the appropriate aa-tRNA. Ponnala et al., calculated these values from existing research [3], which related them to codon frequency. From this body of work, we created a different algorithm to calculate them (Section 4.5).

Let  $abcd$  be a sequence of four nucleotides, with  $abc$  in the current and  $bcd$  in the +1 reading frame, and let  $x$  be the displacement of the current wait cycle of the ribosome. As the incremental displacement approaches +1, the probability of choosing codon  $bcd$  increases and the probability of choosing codon  $abc$  decreases.

We model this behavior using even powers of cosine and sine functions for  $abc$  and  $bcd$ , defining  $\omega$  as the weight that is directly proportional to the probability. It must meet these criteria: If the ribosome lies completely and thus is stable in the zero frame, then the probability of staying in that frame is one. Consequently, a ribosome fully in the +1 frame has no chance of going to the zero frame, hence the requirement for a period of two base pairs ( $x = 4$ ) in these functions. Thus, we propose

$$\omega_{abc} = \cos^{10} \frac{x\pi}{4} \text{ and } \omega_{bcd} = \sin^{10} \frac{x\pi}{4}. \quad (3.1)$$

Suppose we are on a wait cycle at codon  $abc$  with  $N_{abc}$  total cycles allocated. Let  $P$  be the instantaneous probability of staying in the current reading frame at the next move, which we know (above) is proportional to the weight  $\omega_{abc}$ . Let  $1/n_{abc}$  be the constant of proportionality, implying  $n = \omega_{abc}/P$ .<sup>3</sup> Then the aggregate probability of choosing codon  $abc$  after  $K$  cycles is the probability of *not* failing to change the reading frame  $(1 - P)$  at every cycle for  $K$  cycles. Hence,

$$1 - \prod_{i=1}^K \left(1 - \frac{\omega_i}{n}\right) \text{ where } \omega_i = \cos^{10} \frac{x_i\pi}{4}. \quad (3.2)$$

<sup>2</sup>The cosine and sine functions are taken to the tenth power here. These parameters can change.

<sup>3</sup>We can derive  $n$  as follows: We know  $\omega = \cos^{10}(x\pi/4) \leq 1$ , implying  $P \leq 1/n$ . Assume the probability of moving (1/2) is just as likely as the probability of not moving (1/2). Suppose there are  $N$  wait cycles. Then  $1 - (1 - \omega/n)^N = 1/2$ , implying  $n \leq \sqrt[N]{2}/(\sqrt[N]{2} - 1)$ . We also have  $\lim_{N \rightarrow \infty} \max n = N/\ln 2$ . Thus, the probability of choosing a codon at a cycle is proportional to its TAV because  $N$  is proportional to its TAV, which coincides with intuition.

<sup>4</sup>The weight depends upon the frame choice in question.

## 3.2 Consequences of the Stochastic Model

The stochastic model introduced a concept into the model proposed by Ponnala et al. [8]: The ribosome has a finite probability of “choosing the wrong codon,” in essence going out-of-frame due to a high tRNA availability. This is definitely not a programmed frameshift, which is when the force pushes the ribosome to an unstable point and moves it quickly to the +1 reading frame to regain stability. Programmed frameshifts take place over the span of just one codon; the graph never approaches  $x = 2$  over multiple codons.

An incorrect codon choice can occur when displacement nears  $x = \pm 1$ . Our model assumes that the 0 and the  $\pm 1$  reading frame codons have a finite probability of occupying the A-site in the ribosome. From the probability equations, the ribosome increasingly tends to stabilize around the +1 frame by accident. Unlike a programmed frameshift, incorrect codon choice results from the slow digression from the true alignment and occurs over a number of codons. As the ribosome nears  $x = \pm 1$ , the values for both the sine and cosine functions drop, thus increasing the chance of increasing the time required for translation. This models the physical behavior of ribosome’s tendency to pause as it stabilizes upon an aa-tRNA in the A-site.

Choosing the wrong codon is a purely stochastic phenomenon; only through our new model can we actually track the actions of the ribosome throughout translation. From our physical conceptualization (Section 3.1), we establish the potential to measure translational efficiency below.

## 4 Analysis

To test our computational model, we ran a number of experiments to analyze sequences present or expressed in *E. coli*.

### 4.1 Measures

As this model is stochastic, multiple runs (the sample size) must be analyzed. As such, we propose two sample metrics for analysis.

#### Error-Free Rate

When studying a sequence with a programmed frameshift, *error-free rate* measures the percentage of runs during which

the ribosome chooses the correct codon at every juncture. In the case of a known +1 frameshift, the ribosome must choose the +1 frame at the frameshift codon and stay in the 0 frame before and the +1 frame thereafter for the run to be error-free.

#### Displacement Deviation

We define *displacement deviation* to be

$$d = \sqrt{\frac{\sum_i (x_i - \beta_i)^2}{N}}, \quad (4.1)$$

where  $\beta_i$  is the predicted reading frame at codon  $i$ ,  $x$  is the displacement at codon  $i$ , and  $N$  is the total number of codons as a measure of the deviation of the sequence from the expected reading frame. Usually  $\beta_i = 0$  unless a programmed frameshift exists as in *prfB*. For example, for *prfB*,  $\beta_i = 2$  for all  $i \geq 25$  because *prfB* frameshifts at codon 25 UGA and the model represents a frameshift with +2 displacement per Section 2.3.

### 4.2 *prfB* and Related Sequences

In the first test of our model, we investigated the gene *prfB*<sup>5</sup>, which has a programmed frameshift. Weiss et al. [16] conducted a number of experiments regarding *prfB* to test how mutations in the sequence affect the rate of frameshifting. They present a total of 35 sequences in their paper, along with measures of translational efficiency. We hypothesized that genes found to frameshift at high rates by Weiss et al. should also show high error-free rates under our model.

### 4.3 Ribosomal Proteins

We also hypothesized that displacement deviation provides a suitable metric for translational efficiency: A lower deviation should correspond to a more efficient sequence. This is because if, at each wait cycle, the ribosome fails to stabilize in a period of indecision, the probability that it will change reading frames increases. The translational process resolves this indecision by either stabilizing around the incorrect reading frame or pausing. The former synthesizes an

<sup>5</sup> *prfB* encodes protein release factor 2, which enters the A-site and causes protein synthesis to terminate at the stop codons UGA and UAA. Since *prfB* itself requires translation to continue past a stop codon, this sets up an elegant autoregulatory mechanism. We obtained *prfB*’s nucleotide sequence from NCBI’s Genbank database at <http://www.ncbi.nlm.nih.gov/> with accession number NC\_000913.

incorrect primary structure or causes premature truncation due to a out-of-frame stop codon further downstream; latter preserves fidelity at the expense of speed. That is, while a single run may produce a working primary structure of the protein, the sequence in question ultimately is less efficient than a synonymous sequence with lower aggregate probabilities, whether we measure this as the error-free rate (Section 4.1) or displacement deviation (Section 4.1). In addition, a greater deviation from the zero axis inherently implies a longer time for translation, since proximity to  $\pm 1$  increases the probability of waiting and not choosing either codon. A high value thus reduces translational efficiency, again contributing to the probability of translational failure.

To this end, we performed two tests. In one test, we ran the model on all the 4364 genes provided by the Ecogene database of *E. coli* genes.<sup>6</sup> We predicted the majority to exhibit low displacement deviations based on the assumption that evolution would select for high translational efficiency. In another test, we looked at ribosomal proteins, known to have high levels of expression [10]. For this set we predicted lower mean deviation than those of the *E. coli* gene sample.

#### 4.4 Bovine Growth Hormone

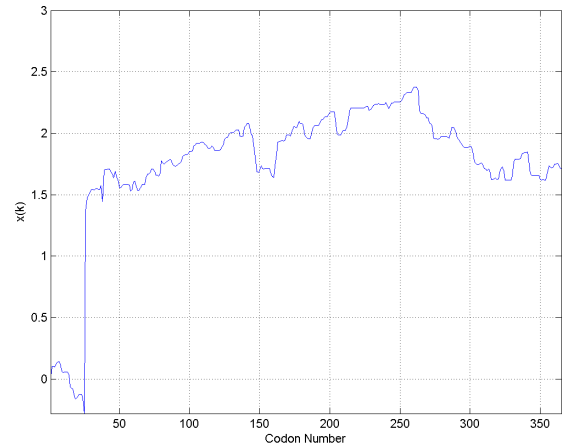
Our final test involved the analysis of a set of sequences known to be translationally regulated. Schoner et al. [12] created a set of eight sequences in an empirical effort to increase the yield of recombinant bovine growth hormone (bGH) synthesized in an *E. coli* expression system. They showed that four sequences were expressed at significantly higher levels than the others. We predicted that these should have lower displacement deviations than others.

#### 4.5 Parameters

For all computational experiments in this report, we use a species angle of  $\theta_{sp} = -30^\circ$  and an initial displacement of 0.1 in accordance with Ponnala et al. [8]. We also explore the effects of changing these parameters on the error-free rate of *prfB*. A parameter that is much more difficult to estimate is the tRNA availability vector (TAV). Ponnala et al. estimated TAV by codon usage, surveying genes from *E. coli*. Although this assumption has experimental basis [3], no concrete evidence supports this assumption. We chose to calibrate the TAV from existing experimental data, designing

<sup>6</sup><http://ecogene.org/>

FIGURE 3. Plots of *prfB* in a stochastic model: Displacement plot



a genetic algorithm to improve these values based on bGH sequences while remaining close to the already determined values.

#### Obtaining the tRNA Availability Vector

We optimize the separation between the displacement deviations of the high-yield and low-yield sequences. First, we generate a list of randomly modified TAVs and calculate the ratio of the displacement deviation (Section 4.1) for the four high-yield sequences to that of the other four. From there, we sort the modified vectors by this ratio and discard the worst half. We repeated choose two of the remaining vectors based on rank, taking a weighted average to spawn a new vector. After creating the next generation of a constant “gene pool” size, we delete the previous generation and repeat. After a fixed number of generations, the algorithm terminates and returns the most optimal vector. This algorithm does not significantly alter the vectors; the average change to each value in the vector was merely 15.99%. We use the new values throughout the paper.

## 5 Results

### 5.1 *prfB*

The gene *prfB*, as mentioned, is known to have a programmed frameshift at the 25<sup>th</sup> codon. Figure 3 shows its displacement plot, again with a distinctive jump at codon 25.

FIGURE 4. Plots of *prfB* in a stochastic model: Sensitivity plot

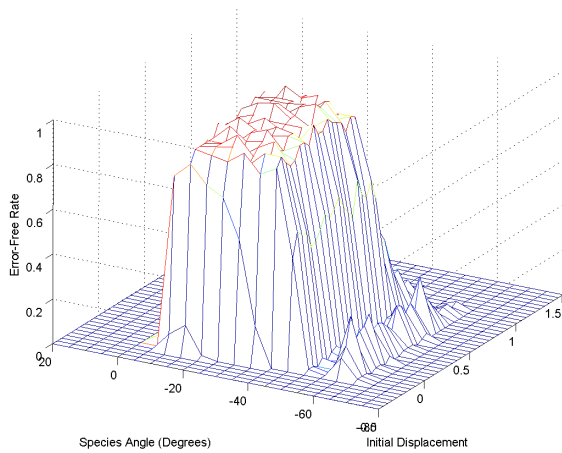
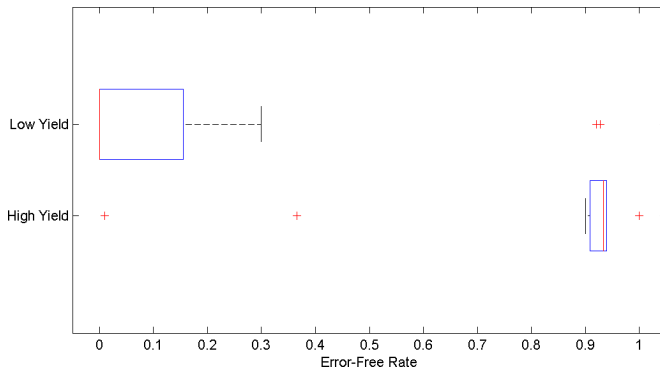


FIGURE 5. Comparison of experimental yield and error-free rate, 500 iterations



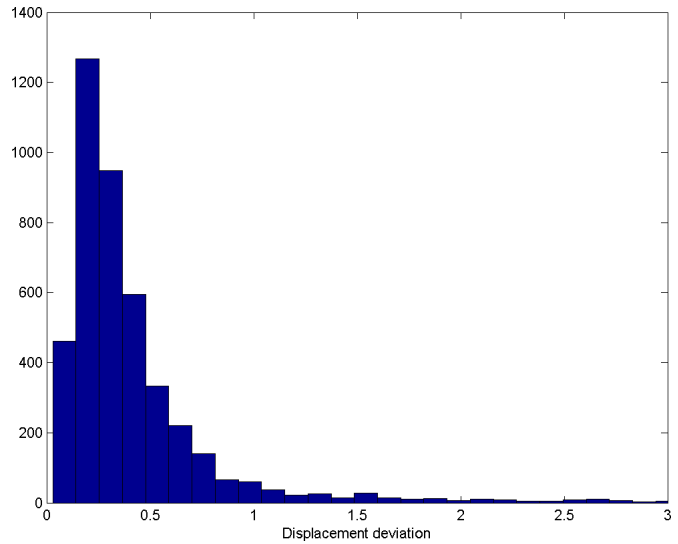
7

Notably, the displacement plot does not reach  $x = 2$  over the span of one codon, as the previous deterministic model predicted. Rather, due to randomness, the ribosome stabilizes the codon of the +1 frame in the A-site before actually reaching a displacement of exactly 2. The propensity to approach and stabilize at  $x = 2$  concurs with experimental evidence indicating the ribosome stays in frame after the *prfB* frameshift to produce full-length RF2. Figure 4 shows the error-free rate of *prfB* as a function of species angle and initial displacement and demonstrates the robustness of our model (Section 6).

We used our computational model to determine whether a correlation exists between error-free rate and the yield of a reporter protein,  $\beta$ -galactosidase ( $\beta$ -gal). Weiss et al.

<sup>7</sup>Note that the polar plot is the same as Figure 2. The new model does not alter the polar plot or the free energy calculations.

FIGURE 6. Investigating a large sample of *E. coli* genes: Displacement deviations



[16] investigated elements of the mRNA sequence that could change the frequency of frameshifting. The frameshift site of *prfB* was fused to the encoding sequence for  $\beta$ -gal so that  $\beta$ -gal activity was dependent on the frameshift occurring, thus serving as an indirect measure of frameshift frequency.

Our metric, error-free rate, showed some ability to divide Weiss et al.'s 35 constructs into two general categories: those with  $\beta$ -gal activity over 1650 whole-cell units and those with a lower activity. This number is in relation to the original, unmodified *prfB* sequence, which exhibited an activity of 6600 units. These computations results suggest that error-free rate could be used to predict protein yield, although it lacks resolution.

Notably, Weiss et al. did not maintain the amino acid structure of the polypeptides. This discrepancy could impact protein folding and half-life. Although they presented no evidence evaluating this idea, such a phenomenon would confound the interpretation of  $\beta$ -gal activity as a measure of frameshift frequency.

## 5.2 *E. coli* Genes

Figure 6<sup>8</sup> is a histogram of the deviation yields of 4364 genes of *E. coli*, encompassing over 80% of the entire genome. As predicted, 93.45% of the genes had deviations in the 0 to 1 interval, agreeing with the assumption of natural efficiency.

<sup>8</sup>We truncate the histogram at three, excluding the outliers and less than 1% of the sample.

FIGURE 7. Investigating a large sample of *E. coli* genes: Comparison to ribosomal proteins

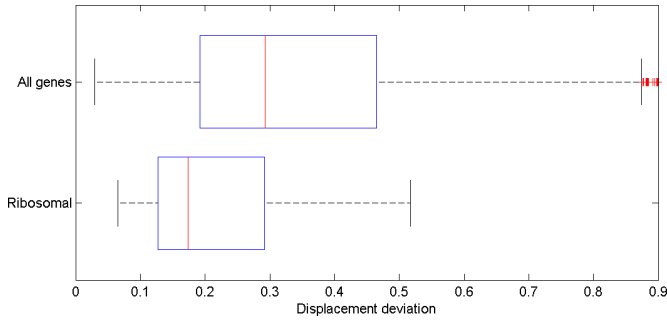
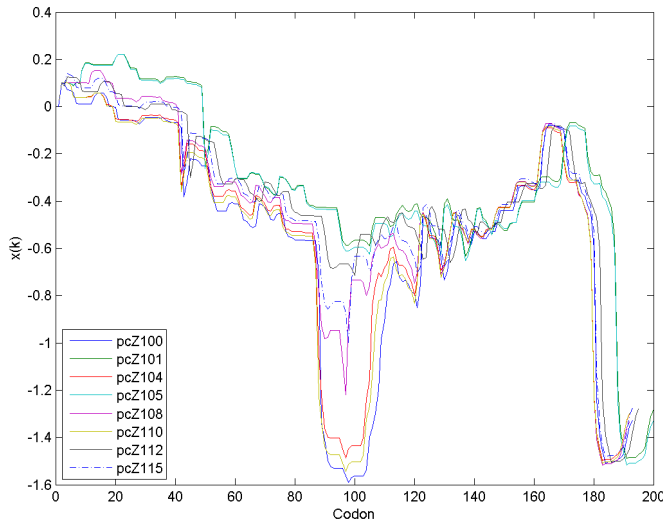


FIGURE 8. bGH: Displacement plot



The average displacement deviation for these *E. coli* genes is 0.4425 ( $\sigma = 0.1537$ ), running 500 iterations per gene.

### 5.3 Ribosomal Proteins

To test further our idea that genes that translate efficiently should exhibit low displacement deviations, we tested our model on ribosomal proteins, which are known to be expressed at especially high levels. The displacement deviation from  $x = 0$  for ribosomal proteins is on average 0.2708 ( $\sigma = 0.0884$ ) in comparison to the average of 0.4425 for our large sample of *E. coli* genes, which is significantly higher with a  $p$ -value of 0.0109 when performing a two-sample  $t$  test on the means.

FIGURE 9. bGH: Deviations with sample size 500

Sequence	$d$	$\sigma(d)$	Yield (% bGH)
pCZ101	0.5146	0.03313	30
pCZ105	0.5139	0.04181	34
pCZ112	0.6612	0.03633	33
pCZ115	0.6721	0.03792	32
pCZ100	0.7107	0.01715	< 0.5
pCZ104	0.7162	0.01433	< 0.5
pCZ108	0.5912	0.05976	1.7
pCZ110	0.7026	0.01966	< 0.5

### 5.4 Bovine Growth Hormone

We investigated the concept of displacement deviation as a predictive parameter for experimental yield using published expression data [12] for bovine growth hormone (bGH). The focus of this research was to modify the bGH mRNA sequence to optimize yield of the protein in an *E. coli* expression system.

Schoner et al. [12] created a number of constructs, primarily modifying the initial codons of a bovine growth hormone sequence. The research found that sequences pcZ101, pcZ105, pcZ112, and pcZ115 have high protein yields in comparison to the four other sequences. We found these aforementioned four sequences in addition to pcZ108 to have the least displacement deviation from  $x = 0$  (Figure 9). Figure 8 shows the displacement plots of all the bGH sequences on the same set of axes.

Therefore, pcZ108 exists as an outlier. Schoner et al. postulates it to have erroneously a low protein yield, attributing it to an experimental error due to its similarity to pcZ114. They believe that the low protein yield is not due to translational effects, a hypothesis that could explain why our model predicts a relatively high translation rate. Excluding this outlier from our data, we find that our data (Figure 9) fully agrees with hers. In addition—and like Weiss et al.—Schoner et al. changed the sequence to encode a different amino acid sequence than the other constructs. We can then attribute the low translational efficiency to the interaction of the primary structure with the ribosome or to protein stability, subjects beyond the scope of our model.

### 5.5 An Artificial Frameshifter

To verify our model's predictive ability, we focused on its ability to predict programmed frameshifts. We designed a

FIGURE 10. Linker sequence: Plasmid construct

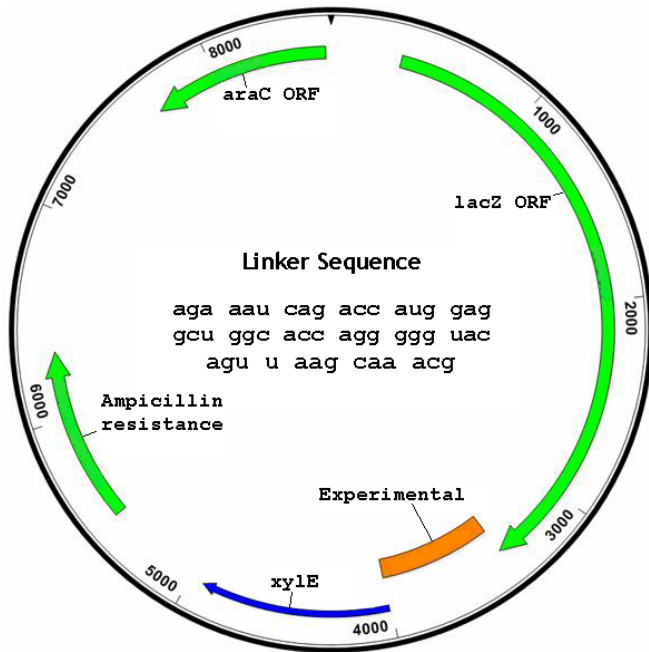
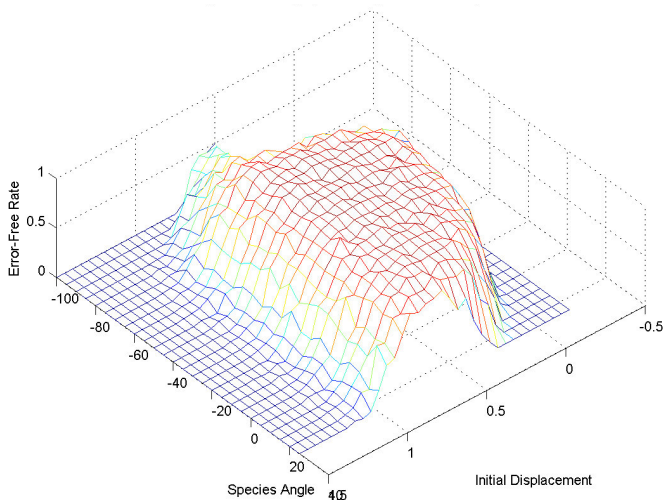


FIGURE 11. Linker sequence: Sensitivity plot



sixteen-codon sequence that should frameshift in *E. coli* into the AAG frame after crossing the sole uracil. Figure 11 shows the error-free rate of the linker sequence as a function of species angle and initial displacement, which is as robust as *prfB*.

Next, we performed a BLAST search, which found no similar sequences in *E. coli*. Then, working with molecular biologists, we helped create a strategy to determine the efficiency of frameshifting using a fusion protein. Biologists will fuse the linker sequence to the 5' to the end of *xylE* [18], which

codes for catechol oxidase, monitored with a colorimetric reaction. A plasmid vector (Figure 10) containing the *xylE* fusion and *lacZ*, a second reporter, genes [11] will be constructed. The *lacZ* and the *xylE* fusion will be co-transcribed on a polycistronic mRNA, but translated separately. The *lacZ* gene product,  $\beta$ -gal, will be expressed constitutively, but the expression of *xylE* will necessitate a frameshift in the linker sequence. As such, standardizing the catechol oxidase to the  $\beta$ -gal activity will serve as a measure of the efficiency. The experimental work is in progress.

## 6 Discussion

The purpose of our studies was to extend and improve the deterministic model of Ponnala et al., which gave a mechanistic perspective of ribosomal movement during translation and a genetic signal approach to translating mRNA sequences. It incorporated a number of parameters, including codon bias [3] and rare codon usage [5], known to affect translation. The model also could predict frameshift locations in the sequence, of value in sequence annotation (frameshift identification). Our studies extended this predictive power to computationally predict translational efficiency.

The deterministic structure limited the model of Ponnala et al. (Section 3.1). We restructured it into a stochastic process to better reflect cellular environment conditions that can affect translation. In essence, the stochastic model paints a more realistic picture of the ribosome, a machine that makes choices nondeterministically due to noise from the cell environment.

In addition to the stochastic version, we also developed two metrics: error-free rate (Section 4.1) and displacement deviation (Section 4.1). Error-free rate provides a measure of an mRNA sequence's propensity for frameshifting by measuring reading frame change frequency. Investigating the constructs used by Weiss et al. [16], our model roughly separated sequences into those with high and low frameshift frequencies. However, the output is not always clear, and we turn to displacement deviation in these cases.

Displacement deviation (Section 4.1), when calculated for a large sample of *E. coli* genes assumed to be translationally efficient, correlated with over 90% of the experimental data, predicting these genes to be efficient. Moreover, the results predicted that ribosomal proteins should be more efficient than average, in accordance with experimental knowledge.

Also, displacement deviations for seven of eight bGH sequences [12] correlated with further experimental data; one sequence was an outlier explained in Schoner's research.

Since the model is preliminary, we must note these outliers. In the case of bGH, Schoner et al. experimentally determined pcZ108 to be of low yield, but the model computationally predicts high yield. Likewise, in the set of constructs used by Weiss et al., one set of sequences had low experimental yields, but the model predicted high translational efficiency. One possible explanation of these discrepancies relates to the chemical structures of the mRNA sequences and the polypeptides they encode. In both cases, Weiss et al. and Schoner et al. did not maintain the amino acid structure while designing the sequences. If a protein-ribosome interaction or post-translational instability resulted from these changes, then such a change could significantly impact protein yield, an indirect measure of translation. However, these effects on translational efficiency are beyond the scope of our model.

We also plan to explore parameter estimation in future studies. Analysis of *prfB* and the proposed linker sequence suggest our model is robust with respect to species angle and initial displacement, but proper estimation of tRNA availability poses a problem. Ponnala et al. [8] based these values solely on codon usage, but research [7] also suggests that tRNA shape also determines overall stability and is known to affect frameshift frequency. The genetic algorithm (Section 4.5) we created thus provides an initial method for a coarse estimation. However, we did not extensively investigate the effect of tRNA shape and thus we limited the difference between our values and those computed by Ponnala et al..

Currently, our model predicts translational efficiency with some accuracy and can distinguish between high and low yield sequences. Using results from our linker sequence (Section 5.5), currently in progress, we will adjust the model to increase its predictive power and consequently its importance in the field of genetics.

## 7 Conclusion

In this paper, we presented a method, based on our newly developed stochastic model and derived metrics that roughly predicts translational efficiency. More research is needed to experimentally validate the model and improve parameter estimates, especially those of TAV values. Despite these

shortcomings, the ability of our model to discern between efficient and non-efficient sequences with respect to translation indicates future potential in the field of recombinant biotechnology in the near future, providing a computational, cost-effective method for gene construct design when using *E. coli* as an expression system. In addition, we may be able to extend our model to other prokaryotic species with similar ribosomes. Having a computational model based on the biological mechanism of translation will further our understanding of the translational process and be practical for the production of recombinant proteins at a commercially useful scale.

## Bibliography

- [1] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Nielson, and D. H. T. II. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Nat. Acad. Sci. USA*, 83:9373–9377, Dec. 1986.
- [2] A. Hui and H. de Boer. Specialized ribosome system: Preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in *Escherichia coli*. *Proc. Nat. Acad. Sci. USA*, 84:4762–4766, 1987.
- [3] T. Ikemura. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. and Evol.*, 2(1):13–34, 1985.
- [4] W. Jacob, M. Santer, and A. Dahlberg. A single base change in the Shine-Dalgarno region of 16S rRNA of *Escherichia coli* affects translation of many proteins. *Proc. Nat. Acad. Sci. USA*, 84:4757–4761, 1987.
- [5] J. F. Kane. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Curr. Opin. in Biotech.*, 6:494–500, 1995.
- [6] M. Kozak. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, 361:13–37, 2005.
- [7] S. Phelps, C. Gaudin, S. Yoshizawa, C. Benitez, D. Foumy, and S. Joseph. Translocation of a tRNA with an extended anticodon through the ribosome. *J. of Mol. Biol.*, 360(3):610–622, 2006.
- [8] L. Ponnala, D. L. Bitzer, A.-M. Stomp, and M. A. Vouk. A model for +1 frameshifts in eubacteria. *Bioinformatics*, 2006.
- [9] L. Ponnala, A.-M. Stomp, D. L. Bitzer, and M. A. Vouk.

Analysis of free energy signals arising from nucleotide hybridization between rRNA and mRNA sequences during translation in Eubacteria. *J. on Bioinformatics and Systems Biol.*, pages 1–9, 2006.

- [10] F. Repoila, N. Majdalani, and S. Gottesman. Small non-coding RNAs, coordinators of adaptation processes in *Escherichia coli*: the *rpoS* paradigm. *Mol. Microbiol.*, 48:855–61, May 2003.
- [11] J. Sambrook, E. Fritsch, and T. Maniatis. *Molecular Cloning: A Laboratory Manual*. CSH Laboratory Press, 1989.
- [12] B. E. Schoner, H. M. Hsuiung, R. M. Belagaje, N. G. Mayne, and R. G. Schoner. Role of mRNA translational efficiency in bovine growth hormone expression in *Escherichia coli*. *Proc. Nat. Acad. Sci. USA*, 81:5403–5407, Sept. 1984.
- [13] J. Shine and L. Dalgarno. The 3′-terminal sequence of *Escherichia coli* 16S ribosomal RNA: Complementarity to nonsense triplets and ribosome binding sites. *Proc. Nat. Acad. Sci. USA*, 71(4):1342–1346, 1974.
- [14] H. P. Sørensen and K. K. Mortensen. Advanced genetic strategies for recombinant protein expression in *Escherichia coli*. *J. of Biotech.*, 115:113–128, 2005.
- [15] J. Starmer. *What can RNA hybrids tell us about translation?* PhD thesis, North Carolina State University, 2006.
- [16] R. B. Weiss, D. M. Dunn, J. F. Atkins, and R. F. Gesteland. Slippery runs, shifty stops, backward steps, and forward hops: -2, -1, +1, +2, +5, and +6 ribosomal frameshifting. *CSH Sym. on Quant. Biol.*, 52:687–693, 1987.
- [17] G. Yusupova, L. Jenner, B. Rees, D. Moras, and M. Yusupova. Structural basis for messenger RNA movement on the ribosome. *Nature*, 444(7117):391–394, 2006.
- [18] M. Zukowski, D. Gaffney, D. Speck, M. Kauffmann, A. Findeli, A. Wisecup, and J.-P. Lecocq. Chromogenic identification of genetic regulatory signals in *Bacillus subtilis* based on expression of a cloned *Pseudomonas* gene. *Proc. Nat. Acad. Sci. USA*, 80:1101–1105, Sept. 1984.