

A model for +1 frameshifts in eubacteria

Lalit Ponnala^{1*}, Donald L. Bitzer², Anne-Marie Stomp³, Mladen A. Vouk²

¹Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC

27695 ²Department of Computer Science, North Carolina State University, Raleigh, NC 27695

³Department of Forestry, North Carolina State University, Raleigh, NC 27695

ABSTRACT

Motivation: This work applies the methods of signal processing and the concepts of control system design to model the maintenance and modulation of reading frame in the process of protein synthesis. The model shows how translational speed can modulate translational accuracy to accomplish programmed +1 frameshifts and could have implications for the regulation of translational efficiency.

Results: A series of free energy estimates were calculated from the ribosome's interaction with mRNA sequences during the process of translation elongation in eubacteria. A sinusoidal pattern of roughly constant phase was detected in these free energy signals. Signal phase was identified as a useful parameter for locating programmed +1 frameshifts encoded in bacterial genes for release factor 2. A displacement model was developed that captures the mechanism of frameshift based on the information content of the signal parameters and the relative abundance of tRNA in the bacterial cell. Results are presented using experimentally verified frameshift genes across eubacteria.

Availability: A set of MATLAB® programs that implement our methods are available upon request from the corresponding author.

Contact: lalit.p@gmail.com

Supplementary information: Supplementary results are available at *Bioinformatics* online.

1 INTRODUCTION

Maintenance and modulation of ribosome reading frame during the translation process is a long-standing problem in RNA biology. In prokaryotes, the idea that hybridization between the 3'-terminal nucleotides of the 16S rRNA (the *tail* of the 16S rRNA) and the mRNA is somehow involved with ribosomal reading frame during translation has a long history. As early as 1979, Sedlacek *et al.* analyzed complementarity between the twelve, 3'-terminal nucleotides of the 16S rRNA tail sequence of *E. coli* and a limited number of mRNAs. Their observation of non-random complementarity led to speculation that nucleotide sequence bias of mRNA coding regions existed as part of a mechanism to modulate ribosomal motion, and hence reading frame, during elongation. Experimental evidence of 16S rRNA tail involvement was obtained by Weiss *et al.* (1988) who proposed a mechanism to control programmed frameshifting of the *prfB* gene that utilizes scanning of the mRNA by the 16S rRNA tail. Trifonov (1992) hypothesized that three regions of 16S rRNA sequence, one of which is the 16S rRNA tail, modulate ribosomal reading frame through their variable hybridization to mRNA. Both the sequence analysis and

experimental studies offer strong evidence for involvement of the 16S rRNA tail sequence in reading frame maintenance and modulation, but do not offer a model of a proposed mechanism.

To further progress towards a model for control of reading frame, we applied electrical engineering concepts used for control system design. In electrical devices, input signals control device states. If the translating ribosome followed this design, its reading frame states, Frame 0, Frame +1 and Frame +2 (or -1), would be controlled by an input signal. In electrical devices, control system design takes the form of a mathematical model of a control system algorithm which decodes input signals to determine device state. The analytical tools of signal processing provide methods for detecting signals, extracting them from noise, characterizing signal parameters, and identifying the parameters and parameter behaviors that are predictive of device states. To use these tools requires a mathematical model of the machine and an algorithm that simulates the machine process.

Our previous work (Ponnala *et al.*, 2006a) has shown that a free energy signal containing a periodic component of frequency $f = 1/3$ can be extracted for each mRNA of a specific eubacterium. Signal extraction is done using an algorithm that creates successive alignments of the bacterium's 16S rRNA 3'-terminal nucleotide tail with the mRNA sequence. For each sequence alignment, a free energy of hybridization is calculated, the value of which is a function of the degree of complementarity. This algorithm simulates scanning of the mRNA by the 16S rRNA tail, as suggested by Weiss *et al.* (1988).

Our hypothesis is that the free energy signal arising from hybridization of the 16S rRNA tail with the mRNA is the input signal that controls reading frame. Modulation of reading frame could be accomplished through this signal if it supplied a force that adjusted the position of the mRNA relative to the ribosome. The first step towards validation of this hypothesis is the development of a mathematical model that defines ribosome position as a function of free energy signal parameters. The second step involves experimental testing of model predictions. This paper presents the development of the mathematical model describing control system design.

2 SIGNAL CHARACTERIZATION AND EXTRACTION

Our previous work (Ponnala *et al.*, 2006a) has shown that the free energy signal contains a periodic $f = 1/3$ component embedded in noise. A suitable model for the free energy signal is

$$y_n = \mu + A \sin \left(2\pi \frac{1}{3} n + \phi \right) + z_n, \quad n = 0 \dots (L-1) \quad (1)$$

*to whom correspondence should be addressed

where L is the number of nucleotides in the mRNA sequence, and z_n is additive IID noise with mean 0 and variance σ^2 . Estimates of signal amplitude A and phase ϕ were obtained using a regression procedure. We found that genes belonging to a specific organism had a roughly constant phase ϕ in their free energy signals and that the mean phase angle of all genes in the species (θ_{sp}) varied linearly with species (G+C) content (Ponnala et al., 2006a). However, the statistical error associated with these estimates was large.

The free energy signal is noisy, resulting in a low signal-to-noise ratio (SNR). The signal periodicity of three nucleotides can be used to improve the signal to noise ratio. The noise component of the signal can be reduced by calculating nucleotide-based averages of free energy triplets. This approach will result in the SNR growing linearly with the number of codons.

2.1 Method of accumulation

A hypothetical memory for the ribosome system can be created consisting of a stack of 3 registers. The memory system maintains updates of the free energy released due to the interaction between the 16S rRNA tail and the mRNA sequence. As the energy values accumulate in the memory registers, information pertaining to the reading frame gets updated.

We denote the register contents by the vector $\mathbf{R}^{(k)}$, $k = 1 \dots \frac{L}{3}$, where $\frac{L}{3}$ is the number of codons in an mRNA sequence. We store the first three energy values (computed from alignments of the 16S rRNA tail with the first 3 bases of the mRNA sequence, i.e. the first codon) in consecutive registers i.e.

$$\mathbf{R}^{(1)} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \end{bmatrix}$$

We then accumulate, or update, the free energies from the first codon by adding to them the free energy values corresponding to the second codon position, resulting in

$$\mathbf{R}^{(2)} = \begin{bmatrix} y_0 + y_3 \\ y_1 + y_4 \\ y_2 + y_5 \end{bmatrix}$$

After accumulating the signal for a length of k codons, the register contents will be

$$\mathbf{R}^{(k)} = \begin{bmatrix} R_1^{(k)} \\ R_2^{(k)} \\ R_3^{(k)} \end{bmatrix} = \begin{bmatrix} \sum_{n=0}^{k-1} y_{3n} \\ \sum_{n=0}^{k-1} y_{3n+1} \\ \sum_{n=0}^{k-1} y_{3n+2} \end{bmatrix}$$

This procedure is repeated until the last mRNA codon is reached, i.e., until $k = \frac{L}{3}$.

2.2 Cumulative magnitude and phase

The register contents $\mathbf{R}^{(k)}$ represent a snapshot of the free energy signal pattern. The three points have a sinusoidal nature due to the dominant periodicity of the energy pattern. This allows us to calculate the cumulative magnitude M_k and phase θ_k by interpolation. As a result, $\mathbf{R}^{(k)}$ can be represented as a phasor

$M_k e^{j\theta_k}$ (Giancoli, 1989). We equate the contents of the registers, after subtracting their mean, to points on a sine-wave and solve Equations (2), (3) and (4) for M_k and θ_k .

$$r_1^{(k)} = R_1^{(k)} - \left(\frac{\sum_{n=1}^3 R_n^{(k)}}{3} \right) = M_k \sin(\theta_k) \quad (2)$$

$$r_2^{(k)} = R_2^{(k)} - \left(\frac{\sum_{n=1}^3 R_n^{(k)}}{3} \right) = M_k \sin\left(\theta_k + \frac{2\pi}{3}\right) \quad (3)$$

$$r_3^{(k)} = R_3^{(k)} - \left(\frac{\sum_{n=1}^3 R_n^{(k)}}{3} \right) = M_k \sin\left(\theta_k + \frac{4\pi}{3}\right) \quad (4)$$

2.3 Signal-to-Noise Ratio

Based on our free energy signal model (Equation (1)), we have

$$r_1^{(k)} = (kA) \sin(\phi) + \left(\sum_{j=0}^{k-1} z_{3j} \right) - \frac{1}{3} \sum_{j=0}^{3k-1} z_j \quad (5)$$

$$r_2^{(k)} = (kA) \sin\left(\frac{2\pi}{3} + \phi\right) + \left(\sum_{j=0}^{k-1} z_{3j+1} \right) - \frac{1}{3} \sum_{j=0}^{3k-1} z_j \quad (6)$$

$$r_3^{(k)} = (kA) \sin\left(\frac{4\pi}{3} + \phi\right) + \left(\sum_{j=0}^{k-1} z_{3j+2} \right) - \frac{1}{3} \sum_{j=0}^{3k-1} z_j \quad (7)$$

Therefore,

$$M_k = kA$$

and

$$\sigma_k^2 = \left(\frac{2k}{3}\right) \sigma^2$$

where σ_k^2 is the noise variance of the contents of the memory register $\mathbf{R}^{(k)}$. The SNR of the register contents is given by

$$\Gamma_k = \frac{M_k^2}{2\sigma_k^2} = \frac{3k}{2} \left(\frac{A^2}{2\sigma^2} \right)$$

Thus, the accumulation of points corresponding to the same sinusoidal pattern causes the SNR to grow linearly with the number of codons.

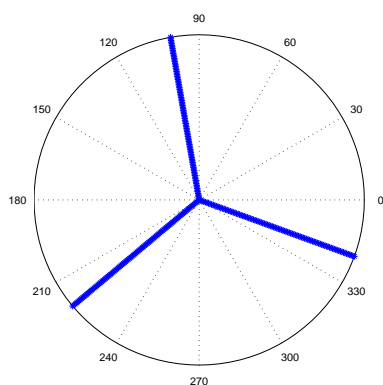


Fig. 1. Thick lines indicate phase boundaries for each reading frame, relative to an initial signal phase of -20°

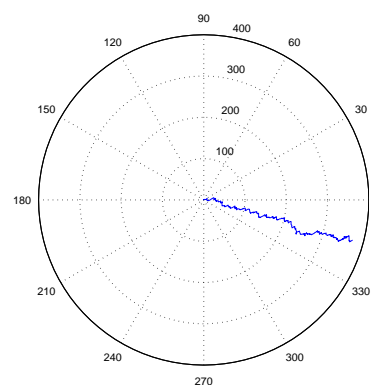


Fig. 2. Polar plot for gene *aceF* in *E. coli*

2.4 Visualization using polar plots

The magnitude M_k and phase θ_k of the register contents can be visualized on a polar plot, with the radial coordinate representing magnitude and the angular coordinate representing phase. Because the free energy signal frequency equals $1/3$ cycles/nucleotide, each 120° sector of the polar plot represents one nucleotide (see Figure 1). For the free energy signal to play a role in reading frame determination, it would be expected that variation in M_k and/or θ_k would correlate with shifts in reading frame. To determine if such a correlation might exist, two genes were selected: *aceF*, a gene which does not encode a frameshift, and *prfB*, a well-studied gene whose mRNA sequence is known to encode a programmed frameshift at codon 26 (Farabaugh, 1996).

Although the polar plot for *aceF* (Figure 2) shows some variation, the cumulative phase stays roughly constant at about -15° , within the sector of one nucleotide. Similar phase constancy was observed in all the 1673 verified genes in *E. coli* of length 200 codons or greater (Ponnala *et al.*, 2006b). However, considerable variation in track within the nucleotide sector can occur (see Figure 3). By comparison, the polar plots of *prfB* (Figures 4 and 5) are quite different. The plot starts in the same nucleotide sector as that for *aceF*, but around codon 26 it swings through approximately 240° . When the phase change is complete, the plot re-establishes itself within a different nucleotide sector and remains there, with small variation, to the end of the gene. Although provocative and consistent with our hypothesis, analysis of other genes known to encode frameshifts would strengthen the correlation.

RECODE¹ is a database of non-canonical translational events such as frameshifts, ribosomal hops and codon redefinition (Baranov *et al.*, 2001)(Baranov *et al.*, 2003). Experimentally verified *prfB* gene sequences for twelve prokaryotes other than *E. coli* were obtained and their free energy signals were calculated using the corresponding species' 16S tail, and signal parameters were generated using the cumulative method. The *prfB* polar plots for all the examined species are shown in the Supplementary Data. A significant phase change is observed around the frameshift location in all these genes, consistent with the results obtained using the *prfB* gene in *E. coli*.

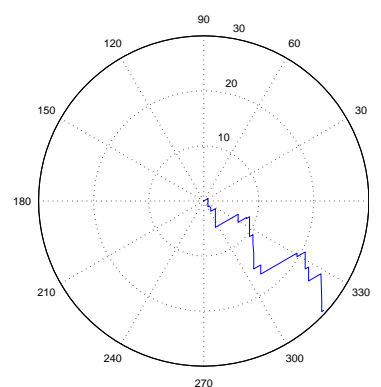


Fig. 3. Polar plot for gene *tsf* in *E. coli*

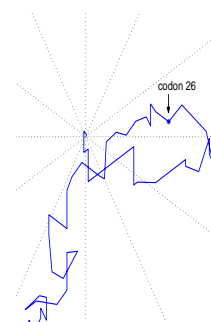


Fig. 4. Partial polar plot for gene *prfB* in *E. coli*; arrow points to the location of frameshift, marked by a *

2.5 Drawbacks

Our cumulative model of signal phase, although useful for revealing frameshift sites encoded in gene sequences, has one significant drawback. For every additional codon, a greater perturbation of the free energy signal will be needed to shift the cumulative phase. This means that the model will have difficulty identifying frameshifts if they occur towards the end of a long gene sequence. Also,

¹ <http://recode.genetics.utah.edu/>

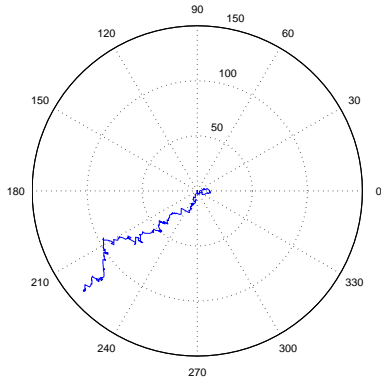


Fig. 5. Polar plot for gene *prfB* in *E. coli*

there is no experimental evidence that indicates that the entire gene sequence upstream of a frameshift site has a controlling influence on the frameshift. The sequence elements that result in a shift in reading frame during translation are small and can be localized in a short sequence within the coding region (Farabaugh, 1996). To accommodate these concerns we developed a new model that estimates instantaneous signal phase at each codon.

3 DISPLACEMENT MODEL

3.1 Calculation of displacement

For a gene without a frameshift, the polar plot would lengthen itself radially (due to growth in magnitude) but stay at a roughly constant phase angle ($\theta_k \approx \theta_{sp}$). When a +1 frameshift happens, the phase moves to a new nucleotide sector, +240° or -120° away. From the *prfB* polar plot, we see that the phase shifts about 60° before it gets to the frameshift location (from approximately -20° to approximately +40°), the equivalent of one-half of a nucleotide. Then it begins its track at the angle that reestablishes it in the new nucleotide sector, +240° from where it originated. We designate $x = 0$ as the initial state, i.e., reading frame 0, as one of the two stable states of the ribosome-mRNA system. We assign unit increments in x for every 60° increment in phase, i.e. for every $\frac{1}{2}$ nucleotide-shift in the mRNA sequence. If the ribosome shifts a whole nucleotide, as it does in the +1 frameshift, we have $x = 2$. So a +1 frameshift can be modeled as a state transition from $x = 0$ to $x = 2$. The intermediate value $x = 1$ can be thought of as a *boundary* point, where there is equal likelihood of picking either the codon in Frame 0 or the codon in Frame +1.

As stated earlier, the cumulative energy signal, owing to its sinusoidal nature, can be represented as $\mathbf{V}_k = M_k e^{j\theta_k}$. We will refer to \mathbf{V}_k as the *cumulative vector*. The contents of \mathbf{V}_k contain a summation of the entire free energy signal up to codon k . The derivative of \mathbf{V}_k with respect to codon position k gives the instantaneous energy available at codon k .

$$\mathbf{D}_k = \frac{d}{dk} \left(M_k e^{j\theta_k} \right) = M_k \frac{d}{dk} \left(e^{j\theta_k} \right) + e^{j\theta_k} \frac{dM_k}{dk} \quad (8)$$

The magnitude and phase of the differential vector \mathbf{D}_k , referred to as differential magnitude and differential phase, are given by Equation (9) and Equation (10) respectively.

$$|\mathbf{D}_k| = \sqrt{\left(\frac{dM_k}{dk} \right)^2 + \left(M_k \frac{d\theta_k}{dk} \right)^2} \quad (9)$$

$$\angle \mathbf{D}_k = \theta_k + \arctan \left(\frac{M_k \frac{d\theta_k}{dk}}{\frac{dM_k}{dk}} \right) \quad (10)$$

To calculate $|\mathbf{D}_k|$ and $\angle \mathbf{D}_k$, we will need the derivatives ($\frac{dM_k}{dk}$ and $\frac{d\theta_k}{dk}$), which can be evaluated using function approximation techniques (Cheney and Kincaid, 1999). A second order polynomial can be fitted to a window of points centered around M_k , to evaluate its derivative, $\frac{dM_k}{dk}$. An identical procedure is followed for computing $\frac{d\theta_k}{dk}$.

We observe that for a signal that stays roughly in phase, $\frac{d\theta_k}{dk} \approx 0$, and so, $|\mathbf{D}_k| \approx \frac{dM_k}{dk}$ and $\angle \mathbf{D}_k \approx \theta_k$. We know, from previous work that the free energy signals in a given eubacterium have a roughly constant phase (Ponnala et al., 2006a). For *E. coli*, that angle is $\theta_{sp} \approx -20^\circ$. For a normal, non-frameshifting gene of length L nucleotides in *E. coli*, we see that $\theta_k \rightarrow \theta_{sp}$ as $k \rightarrow \frac{L}{3}$. Within the context of our hypothesis, the differential vector \mathbf{D}_k represents a force acting on the ribosome at codon k that adjusts the position of the ribosome relative to the mRNA, i.e., that modulates reading frame.

Another element believed to play an integral part in programmed frameshifts is ribosomal pausing (Farabaugh, 1996). Siple and Goldman (1993) provide experimental evidence that supports a frameshift model in which ribosomal pause time is a major determinant of frameshift probability, with pause time a function of tRNA availability. Therefore, we introduce the concept of *wait-time*, a measure of how long the ribosome waits for the tRNA to associate with the ribosome A-site, into our displacement model.

3.2 Estimating wait-time

The actual availability of tRNA, estimated using two-dimensional polyacrylamide gel electrophoresis, was found to be proportional to codon frequency for moderately expressed genes (Ikemura, 1985). Using a set of mRNA sequences in *E. coli* that have N codons in all, the frequency of each codon (except the stop codons) can be calculated as

$$f_i = \frac{N_i}{N}, \quad i = 1 \dots 61 \quad (11)$$

where N_i is the number of codons of type i . If a particular tRNA recognizes only one codon, then the codon frequency would be indicative of its availability. If there is more than one codon recognized by a tRNA isoacceptor, then the availability of that isoacceptor will be the sum of the individual codon frequencies. We estimate the availability of each tRNA isoacceptor using

$$\gamma_p = \sum_{i=1}^{n_p} f_i, \quad p = 1 \dots 20 \quad (12)$$

where n_p is the number of codons that code for amino acid p .

Codons having abundant tRNAs would have short wait-times, and vice-versa. We assume a decreasing linear relationship between the wait-time τ and the tRNA availability γ , as shown in Equation (13). The wait-time gives an approximate number of cycles for which the ribosome can adjust itself while waiting for the appropriate tRNA.

Codon	Amino-acid	Number of wait-cycles
aac	Asn	7
ccu	Pro	16
acg	Thr	13
cuu	Leu	13
uuc	Phe	7
gca	Ala	2

Table 1. Wait-times for a few sample codons in *E. coli*

The number of wait cycles for a few sample codons are shown in Table 1.

$$\tau_p = \frac{\max(\gamma) - \gamma_p}{\min(\gamma)} \quad (13)$$

3.3 The complete model

The vector \mathbf{D}_k represents a force that could produce a linear movement of the ribosome one way or the other until the corresponding tRNA is found for the codon in the A-site. The displacement at each codon position is calculated incrementally (Δx), with the sign of Δx indicating the direction of movement (+ = downstream, - = upstream). The total displacement x_k is obtained by accumulating Δx for the corresponding number of wait cycles. When the ribosome is in reading frame 0, we define $x = 0$ and when it moves into the +1 frame, we define $x = 2$. We claim that the following equation captures the behavior in both reading frame states:

$$\Delta x_k = -C |\mathbf{D}_k| \sin \left(\angle \mathbf{D}_k + \frac{\pi x_k}{3} - \theta_{sp} \right) \quad (14)$$

The argument of the sine function contains the instantaneous measurement of phase:

$$\theta_{\Delta x} = \frac{\pi x_k}{3} - \theta_{sp} \quad (15)$$

Observe that when $x = 0$, the cumulative phase is at the species angle i.e., $\angle \mathbf{D}_k = \theta_{sp}$, leading to $\Delta x = 0$. When $x = 2$, we have $\angle \mathbf{D}_k = \theta_{sp} + \frac{4\pi}{3}$, again leading to $\Delta x = 0$. To calculate Δx , we introduce a constant of proportionality C , and calibrate it using the *prfB* signal. Mathematically, C measures the rate at which the ribosome adjusts itself to perturbations in x . For each unit of wait-time (also referred to as a *wait-cycle*), the incremental displacement Δx_k^j gets added onto the current position x_k^j . The total displacement is then assigned to the next codon $k + 1$. Note that we are using the superscript j to index increments made during the wait-time of the ribosome. If the ribosome waits for τ cycles at codon k , the total initial displacement at codon $k + 1$ would be assigned as

$$x_{k+1}^0 = \sum_{j=1}^{\tau} \Delta x_k^j \quad (16)$$

3.4 Stability

In practice, all the above equations hold approximately, so it is important to establish stability of the ribosome-mRNA system in a rigorous manner (Strogatz, 1994). Equation (14) can be written as a recursive relation

$$x_k^{j+1} = x_k^j - C |\mathbf{D}_k| \sin \left(\angle \mathbf{D}_k + \frac{\pi x_k^j}{3} - \theta_{sp} \right) \quad (17)$$

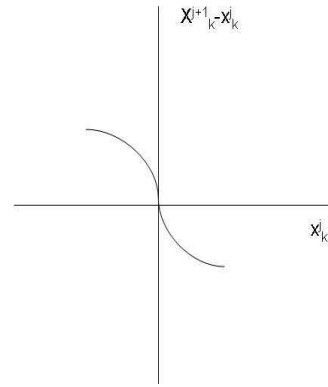


Fig. 6. Vector field generated by Equation (14)

3.4.1 Stability of $x^* = 0$ When the ribosome is in reading frame 0, $x_k^j = 0$ and $\angle \mathbf{D}_k = \theta_{sp}$. Substituting $x_k^j = 0$ into Equation (17) leads to $x_k^{j+1} = x_k^j$, and hence, $x^* = 0$ is a fixed point. Let $\eta_j = x_k^j - x^*$ be a small perturbation away from x^* . To see whether the perturbation grows or decays, we substitute $x_k^j = \eta_j + x^*$ into Equation (17). The recursive relation can now be written as

$$x^* + \eta_{j+1} = x^* + \eta_j - C |\mathbf{D}_k| \sin \left(\angle \mathbf{D}_k + \frac{\pi(x^* + \eta_j)}{3} - \theta_{sp} \right)$$

Substituting $x^* = 0$, we get

$$\eta_{j+1} = \eta_j - C |\mathbf{D}_k| \sin \left(\frac{\pi \eta_j}{3} \right) \quad (18)$$

Since η_j is small, we have

$$\eta_{j+1} \approx \eta_j - C |\mathbf{D}_k| \frac{\pi \eta_j}{3} = \left(1 - C \frac{\pi |\mathbf{D}_k|}{3} \right) \eta_j$$

By making C fairly small, it can be ensured that $\left(C \frac{\pi |\mathbf{D}_k|}{3} \right) < 1 \forall k$. This implies that η_j decays to zero as j gets large, since $\left(1 - \frac{\pi |\mathbf{D}_k|}{3} \right) < 1$. Thus, small perturbations cause the displacement to converge to the fixed point $x^* = 0$. The idea is illustrated in Figure 6.

3.4.2 Stability of $x^* = 2$ When the ribosome is in reading frame +1, $x_k^j = 2$ and $\angle \mathbf{D}_k = \theta_{sp} + \frac{4\pi}{3}$. Substituting these into Equation (17) yields $x_k^{j+1} = x_k^j$, so $x^* = 2$ is a fixed point. For a nearby point $x_k^j = x^* + \eta_j$, the recursive relation takes the form

$$x^* + \eta_{j+1} = x^* + \eta_j - C |\mathbf{D}_k| \sin \left(\angle \mathbf{D}_k + \frac{\pi(x^* + \eta_j)}{3} - \theta_{sp} \right)$$

Substituting $x^* = 2$, we get an equation identical to Equation (18). Following identical steps, we may establish the stability of the fixed point $x^* = 2$.

The above arguments have established that the Equations (14) and (15) are structured so that the states $x = 0$ and $x = 2$ represent stable fixed points of the ribosome-mRNA system. Transition between the states is governed by the differential vector \mathbf{D}_k and the time τ for which the ribosome waits at codon k .

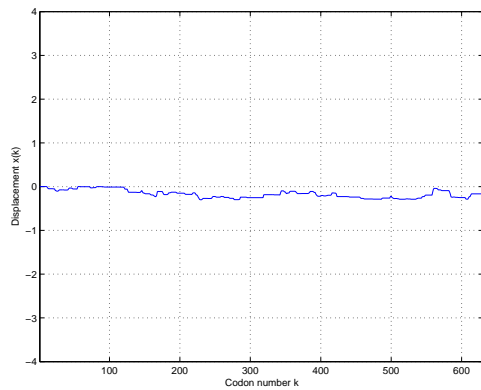


Fig. 7. Displacement plot for gene *aceF* in *E. coli*

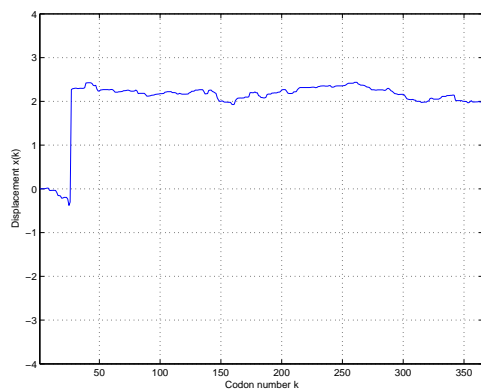


Fig. 8. Displacement plot for gene *prfB* in *E. coli*

4 RESULTS

Two model parameters, the species phase angle, θ_{sp} , and the constant, C , must be specified to generate displacement values. The species phase angle θ_{sp} is the mean phase angle estimated from the set of verified genes as annotated in GENBANK², using the method described in Ponnala et al. (2006a). For *E. coli*, the estimated value is $\theta_{sp} = -13^\circ$. For gene *prfB* in *E. coli*, the value of $C = 0.005$ gave the highest resolution of a jump in displacement at codon 26. These values of θ_{sp} and C were used for subsequent analyses of other genes in *E. coli*. The values of these parameters for other bacteria are listed in the Supplementary Data. At the first codon of a gene sequence, the ribosome is locked into Frame 0, so we use $x_1 = 0$. The stop codons are assigned a large number of wait-cycles, typically 1000.

The displacement plots for the *aceF* and *prfB* genes of *E. coli* are given in Figures 7 and 8, respectively. Several features of these plots are of note. The displacement plot for *aceF* (Figure 7), a gene lacking a frameshift, shows that $x \approx 0$ for the entire length of the coding region. This behavior of x indicates that our method does not detect a frameshift in this gene, the expected result. In contrast, the displacement plot for the *prfB* gene (Figure 8) shows a sudden shift in x at codon 26, the absolute value of which is slightly greater than

2 and it is in the positive direction. Our algorithm is scaled such that a displacement value of $x = 2$ indicates a shift of one nucleotide, so in this case, the displacement indicates a +1 nucleotide shift in reading frame. This is also an expected result given that codon 26 is the location of a +1 frameshift in the *prfB* gene. For the remainder of the sequence, i.e., from codon 27 to the end of the gene, the value of x remains roughly at $x = 2$. This indicates that the gene stays in the new reading frame. The *prfB* displacement plots for the remaining bacteria that we analyzed are given in the Supplementary Data.

Link et al. (1997) assessed the *in vivo* abundances of proteins in *E. coli* using electrophoresis, and ranked the genes in decreasing order of yield. We calculated the free energy signals for 87 such genes in *E. coli*, and analyzed them using our model. We found that for 86 of these genes, $-1 < x_k < 1$ for all values of k , indicating that the ribosome stays in frame for the entire length of each sequence. For the one remaining gene, we found slight deviation from the boundary value of $x_k = 1$ at $k = 70$, indicating a low probability of picking the in-frame codon at that location. The polar plots and displacement plots for 10 of these genes are included in the Supplementary Data.

5 DISCUSSION

Our previous work defined an algorithm that simulates possible hybridization between the 3'-terminal nucleotides of the 16S rRNA and the mRNA. The algorithm revealed a periodic, free energy signal in the coding regions of the genes in a number of bacterial species (Ponnala et al., 2006a). Based on the ideas of Weiss et al. (1988), Trifonov (1992) and others, we hypothesized that this free energy signal could be supplying the information to modulate reading frame.

Using the free energy signal we developed a mathematical model optimized to precisely predict the codon location of the frameshift site within the *prfB* coding sequence. The model is an adaptive algorithm that estimates the displacement of the ribosome from its original reading frame (Frame 0). This algorithm enables us to track the state of the ribosome-mRNA system. The physical interpretation of the differential vector, \mathbf{D}_k , in the model is that it represents the amount of force available at codon k to adjust the position of the mRNA. The amount of this adjustment potential that is actually realized is proportional to the time the ribosome waits for a tRNA to occupy the A-site. If the tRNA is relatively abundant, little of the adjustment is realized; if the tRNA is rare implying a long pause before the A-site is occupied, more adjustment of the mRNA relative to the ribosome occurs. The displacement x , captures the position adjustment. In a recursive form, the model starts with the previous position, derived from the energy signal for all the codons up to but not including the current codon, and uses the new displacement value to update the position, or state, of the mRNA relative to the ribosome.

In the course of developing our model, we have made several approximations and assumptions. One model assumption is that the presence of rare codons is the only factor modulating elongation rate. This assumption is consistent with Spirin (1999) who asserts that the wait time due to the relative abundance of the tRNA can be assumed to be a dominating factor in inducing frameshifts. Although mRNA secondary structure is believed to result in ribosomal pausing, its absence from our model is based on the observation that a strong correlation has not been observed in all

² <http://www.ncbi.nlm.nih.gov/Genbank/>

cases between mRNA secondary structure and frameshifting (Kontos *et al.*, 2001).

A second assumption concerns the proportionality between frequency of tRNA isoacceptor (calculated using Equation (12)) and actual tRNA availability. This proportionality is found to break down at low frequencies for genes encoding highly abundant proteins (Ikemura, 1985). The codon bias in such genes is extreme, and this implies that the actual tRNA availability may be more than that estimated using our simple frequency calculation. This introduces a small error into the wait-time estimated using Equation (13). However, this small error would not significantly impact our overall results obtained by assuming that the wait-time is inversely proportional to our estimated tRNA availability. Another approximation involves the calculation of species mean phase angle θ_{sp} . We have used *all* the coding sequences annotated as “verified” in the GENBANK database, leading to a large variance in the estimate of θ_{sp} . A more confident estimate may be obtained by using genes whose authenticity has a greater degree of certainty, such as the genes studied by Link *et al.* (1997).

Our model has utility as both a tool that could be used for sequence annotation and for its implications as to the mechanism of reading frame maintenance and frameshifting. Sequence annotation is an early objective for genome sequencing projects. Frameshift sites are difficult to recognize (Moon *et al.*, 2004) for current gene annotation programs such as GENMARK (Borodovsky and McIninch, 1993) and GLIMMER (Delcher *et al.*, 1999). Our model implies that a free energy signal that is used to maintain reading frame is encoded in the coding regions of authentic genes. The existence of this signal can be visualized using either polar plots of signal phase and magnitude or in displacement plots. We are currently exploring this approach with the objective of developing an annotation program that can identify authentic coding regions and frameshift locations.

The utility of this model from the mechanistic perspective is that it suggests how both reading frame maintenance and reading frame shifts could be encoded in mRNA sequences using translational speed to modulate positional accuracy. The model captures the idea that the instantaneous component of hybridization energy, D_k (whose amount is a function of the mRNA sequence), is available to the ribosomal complex to adjust the position of the mRNA relative to the ribosomal decoding center by an amount that is proportional to the time required for a tRNA or release factor to fully occupy the A-site. The model implies that the codon bias of mRNAs could reflect the existence of a position-adjusting mechanism to maintain reading frame. Through codon selection, each mRNA sequence carries the information to fine-tune the position of each codon in the decoding center taking into consideration variable translational speed.

One consequence of our interpretation of the functional significance of codon bias is that it could give insight into the empirically demonstrated connection between native and recombinant protein yields and codon bias. Using the free energy signal parameters as indicators of elongation accuracy, one way to think about our model is that it yields a qualitative estimate of the frameshift tendency within a coding sequence. To the degree that protein yield losses are determined by elongation errors, such as incorrect recruitment of tRNA, our model can show where such errors are most likely to occur in the coding sequence. Our model can also determine which possible sequence modifications would reduce the likelihood of such errors. By fitting a likelihood function

to the displacement data x_k , we could quantify the “correctness” of a coding sequence for translation. These predictions would then need to be experimentally tested.

Our model also illustrates the value of applying engineering concepts to biological systems. The translation process operates with high reliability in potentially variable environments. As such, it can be considered a dynamic process in which the existence of a control system for reading frame maintenance is a reasonable engineering assumption. Mathematical modeling of control systems for dynamic processes has been the subject of considerable research (Maybeck, 1979). Signal processing techniques have been used with considerable success to estimate the various states of a dynamic process using noisy measurements. The Kalman filter (Kalman, 1960)(Brown and Hwang, 1992) is one of the most useful control system models. This filter uses recursive updating of the process state based on discrete sampling of input signal information. One example application is maintaining a ship’s geographical position despite drift, a problem that bears some similarity to the problem faced by the ribosomal complex in maintaining reading frame.

Each cycle of translation elongation requires the ribosomal complex to return to the same “position”, i.e., the positioning of the tRNA carrying the nascent polypeptide chain in the P-site. The precision of this position is critical as the P-site tRNA spatially defines the A-site boundary in the ribosomal complex (Baranov *et al.*, 2004). The translational process must accomplish precise positioning of the P-site tRNA in the face of considerable process variation, including potentially changing environmental conditions of salt concentration, temperature, pH, and variable process components such as tRNAs and mRNA sequences. The requirement for the ribosomal complex to return to position in the face of environmental perturbations is analogous to the drift problem encountered in the ship example. In our model the equation for calculating instantaneous phase (Equation (15)) is analogous to the *measurement equation* of a Kalman filter, and the recursive relation (Equation (17)) is analogous to its *state update* equation. We have identified two states $x = 0$ and $x = 2$ corresponding to reading frames 0 and +1, respectively. The ribosome-mRNA system is shown to be stable in each of these two states, i.e., small perturbations to the state x_k arising from minor signal deviations will die out eventually. Our algorithm lays the ground work for using adaptive filtering techniques to detect frameshifts in coding sequences. The logical next step is to design an algorithm that describes the transition into the -1 frame, and thereby develop a generalized model of reading frame maintenance in bacteria.

ACKNOWLEDGEMENT

This work was supported in part by NC State DURP funds.

REFERENCES

- Baranov, P. V., Gurchich, O. L., Fayet, O., Prere, M. F., Miller, W. A., Gesteland, R. F., Atkins, J. F., and Giddings, M. C. (2001). RECODE: a database of frameshifting, bypassing and codon redefinition utilized for gene expression. *Nucleic Acids Res*, **29**(1), 264–267.
- Baranov, P. V., Gurchich, O. L., Hammer, A. W., Gesteland, R. F., and Atkins, J. F. (2003). RECODE 2003. *Nucleic Acids Res*, **31**(1), 87–89.

- Baranov, P. V., Gesteland, R. F., and Atkins, J. F. (2004). P-site tRNA is a crucial initiator of ribosomal frameshifting. *RNA*, **10**, 221–230.
- Borodovsky, M. and McIninch, J. (1993). GENMARK: Parallel gene recognition for both dna strands. *Computers Chem.*, **17**(19), 123–133.
- Brown, R. G. and Hwang, P. Y. C. (1992). *Introduction to Random Signals and Applied Kalman Filtering*. John Wiley and Sons, Inc., 2 edition.
- Cheney, W. and Kincaid, D. (1999). *Numerical Mathematics and Computing*. Brooks/Cole Publishing Company, 4 edition.
- Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Research*, **27**(23), 4636–4641.
- Farabaugh, P. J. (1996). Programmed translational frameshifting. *Microbiol Rev*, **60**(1), 103–134.
- Giancoli, D. C. (1989). *Physics for Scientists and Engineers*. Prentice Hall.
- Ikemura, T. (1985). Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol*, **2**(1), 13–34.
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME - Journal of Basic Engineering*, **82 (Series D)**, 35–45.
- Kontos, H., Naphine, S., and Brierley, I. (2001). Ribosomal pausing at a frameshifter RNA pseudoknot is sensitive to reading phase but shows little correlation with frameshift efficiency. *Mol Cell Biol*, **21**(24), 8657–8670.
- Link, A. J., Robison, K., and Church, G. (1997). Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli*. *Electrophoresis*, **18**, 1259–1313.
- Maybeck, P. S. (1979). *Stochastic models, estimation, and control*, volume 141 of *Mathematics in Science and Engineering*.
- Moon, S., Byun, Y., Kim, H.-J., Jeong, S., and Han, K. (2004). Predicting genes expressed via -1 and +1 frameshifts. *Nucleic Acids Res*, **32**(16), 4884–4892.
- Ponnala, L., Stomp, A.-M., Bitzer, D. L., and Vouk, M. A. (2006a). Analysis of free energy signals arising from nucleotide hybridization between rna and mrna sequences during translation in eubacteria. *EURASIP Journal on Bioinformatics and Systems Biology*, **2006**, Article ID 23613, 9 pages. doi:10.1155/BSB/2006/23613.
- Ponnala, L., Bitzer, D. L., Stomp, A., and Vouk, M. A. (2006b). A computational model for reading frame maintenance. In *Proceedings of the 28th IEEE EMBS Annual International Conference*, pages 4540–4543. IEEE. ISBN: 14244-0033-3.
- Sedlacek, J., Fabry, M., Rychlik, I., Volny, D., and Vitek, A. (1979). The arrangement of nucleotides in the coding regions of natural templates. *Mol Gen Genet*, **172**(1), 31–6.
- Shah, A. A., Giddings, M. C., Parvaz, J. B., Gesteland, R. F., Atkins, J. F., and Ivanov, I. P. (2002). Computational identification of putative programmed translational frameshift sites. *Bioinformatics*, **18**(8), 1046–1053.
- Sipley, J. and Goldman, E. (1993). Increased ribosomal accuracy increases a programmed translational frameshift in *Escherichia coli*. *Proc Natl Acad Sci USA*, **90**(6), 23152319.
- Spirin, A. S. (1999). *Ribosomes*. Springer.
- Strogatz, S. H. (1994). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*. Perseus Books, Cambridge MA.
- Trifonov, E. N. (1992). Recognition of correct reading frame by the ribosome. *Biochimie*, **74**(4), 357–362.
- Weiss, R. B., Dunn, D. M., Dahlberg, A. E., Atkins, J. F., and Gesteland, R. F. (1988). Reading frame switch caused by base-pair formation between the 3' end of 16S rRNA and the mRNA during elongation of protein synthesis in *Escherichia coli*. *EMBO J*, **7**(5), 1503–1507.