Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

A computational model for reading frame maintenance

L. Ponnala[1], D. L. Bitzer[2], A. Stomp[3], M. A. Vouk[2]
[1]Department of Electrical and Computer Engineering, [2]Department of Computer Science,
[3]Department of Forestry
North Carolina State University, Raleigh, NC 27695 USA

Abstract— **The free energy released during the interaction of the 16S rRNA tail with the mRNA sequence during translation contains a weak sinusoidal pattern of frequency 1/3 cycles/nucleotide. We hypothesize that this signal encodes information related to the maintenance of reading frame during elongation. In the case of the well-studied +1 frameshifter, *prfB* in *E. coli*, we have observed a direct relationship between cumulative signal phase and reading frame. Based on this observation, we have developed a model that indicates how likely it is for the ribosome to stay in frame throughout the process of elongation. We validate this model by analyzing verified coding sequences in *E. coli*.**

I. INTRODUCTION

Our group has been utilizing signal processing analysis to better understand the decoding mechanisms that are used in the prokaryotic translational process of protein synthesis [1][2][3][4]. This approach has revealed a periodic signal encoded in the mRNA nucleotide sequence. This signal is revealed by calculating the variable free energy of hybridization of the 3'-terminal nucleotides of the 16S rRNA with the mRNA as it moves through progressive alignments during elongation. The phase and magnitude of this free energy signal can be estimated. In doing so, we discovered that the signals corresponding to coding regions of genes in each bacterial species have a roughly constant phase and that the mean signal phase is a function of the genome (G+C) content [5].

Prior work of Weiss and co-workers [6][7] has established the role of hybridization between the rRNA tail sequence with the mRNA in the regulation of the programmed frameshift in the *E. coli* gene, *prfB*. This work combined with our signal discoveries led us to hypothesize that the free energy signal could encode information that is utilized during translation elongation to maintain the correct reading frame. Maintenance of the correct reading frame is absolutely critical to the fidelity of protein synthesis. Therefore, identifying the factors that contribute to reading frame maintenance is fundamental to understanding the rules that define gene sequence structure, that regulate recombinant protein yields in host cells, and that govern the correct structuring of genes for gene therapy. We developed a computational model based on observations of the translational process, utilizing our signal analysis approach, to test our hypothesis. The results of this work are presented here.

II. COMPUTATIONAL METHODS

A. The Conceptual Model

For the purposes of developing our model, we utilized a simplified view of translation elongation. The cycle starts with the tRNA carrying the nascent polypeptide occupying the P-site. The A-site is open and accessible for binding of the correct amino acyl-tRNA. The cycle waits to advance until the proofreading mechanism determines that the correctly charged tRNA occupies the A-site. Proofreading allows a conformation change which aligns the nascent peptide and the new amino acid in the peptidyl transferase site facilitating the formation of the peptide bond, and the transfer of the nascent polypeptide chain to the tRNA occupying the A-site. Subsequent to this step, the hydrolysis of GTP coupled with the release of elongation factor Tu energetically favor conformational changes that result in the translocation of the ribosome the distance of one codon (three nucleotides) bringing a new codon into the A-site, moving the tRNA carrying the nascent peptide to the P-site and moving the now uncharged tRNA previously found in the P-site into the E-site for release from the ribosomal complex. The cycle is complete and repeats itself until the entire polypeptide is synthesized.

Our model attempts to capture the periodic motion of the ribosome and provide a mechanism by which this motion can maintain the precision required for correct translation of the mRNA, i.e. correct maintenance of reading frame. Our assumption is that the actual translocation step of three nucleotides is imprecise in that it can sometimes overshoot or undershoot. In our model, this positional error, also referred to as displacement, is caused by the interaction of the 16S rRNA tail with the mRNA.

The amount of energy that causes displacement arises from the variable free energy signal, whose variation is a direct result of the mRNA nucleotide sequence. This positional error would be captured by the phase of the directional vector representing signal change between adjacent codons. In the majority of translocation steps, the error is random and small so that the correct codon is located in the A-site with sufficient precision to maintain reading frame for correct translation. However, significant deviation from this assumption could occur. If the mRNA sequence was such that these errors did not cancel but were additive, each translocation cycle would find the codon position in the A-

1-4244-0033-3/06/$20.00 ©2006 IEEE.

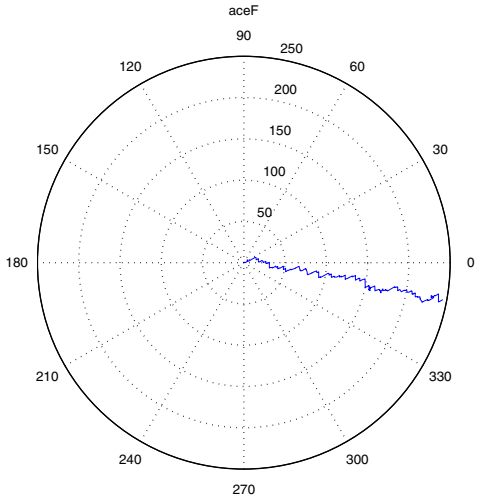

Fig. 1. Polar plot for *aceF*

site being driven closer and closer to the tolerance limits for the correct reading frame. If this were to occur, at each cycle, there would be an increased probability that a codon in the wrong reading frame would be translated. To capture this idea, the model needs to have memory of the directional vectors.

B. Cumulative Phase and Magnitude

Since we know that the free energy signal is periodic with frequency (1/3), we need three registers in our memory model [8]. Our hypothesis is that the ribosome needs to "see" a specific phase to stay in frame. This is in agreement with the "shifty sites" model of frameshifting [7]. The memory registers are filled with the position-specific free energy estimates, in a cumulative fashion. We fit a sine-wave to the memory contents at every step, and estimate its magnitude M_k and phase θ_k by solving a set of simultaneous equations [3]. These cumulative vectors can be visualized using a polar plot, with the radial coordinate indicating magnitude and the angular coordinate representing phase. Fig. 1 shows the polar plot for *aceF*, a normal, i.e. non-frameshifting, gene in *E. coli* K-12. Note that the cumulative phase stays roughly constant at about -20°. The polar plot for *prfB* shows a phase change around the location of frameshift (Fig. 2 and Fig. 3). The cumulative phase is initially at the species-specific angle, but changes through -140° after the frameshift at codon 26. This is clearly seen using our cumulative model since the frameshift occurs relatively close to the start of the sequence. In cases where the frameshift occurs towards the end of the sequence, the change in cumulative phase may not be significant due to the already "energy-heavy" memory registers. Since the phase change may not be a good way to locate frameshifts in longer genes, especially if they occur in the latter half of the sequence, we propose a new model that is closely tied to the physical mechanism of elongation.

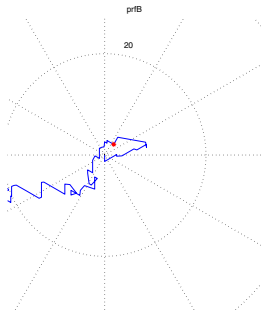

Fig. 2. Polar plot for *prfB*: zoomed-in, * indicates location of frameshift

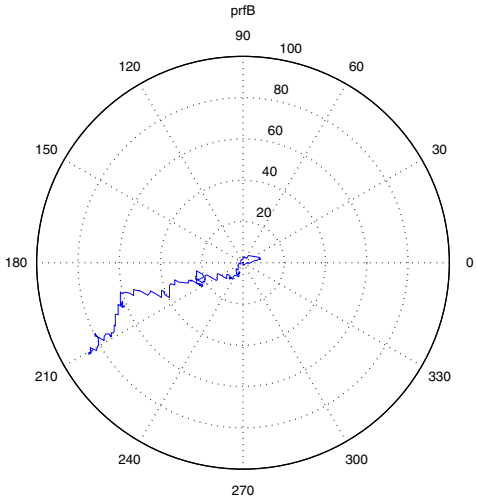

Fig. 3. Polar plot for *prfB*

C. Quantifying Displacement

Our hypothesis is that the elongation step consists of two parts: the approximate translocation of the ribosome and an incremental displacement in the positioning of the mRNA codon in the A-site. The periodic signal contained in the memory registers could be shifted away in either direction from the "perfect lock" position. We could quantify this displacement by using the relationship between the phase of a sine wave and its linear position (see Fig. 4). Since the periodicity of our free energy signal corresponds to the length of each codon, we may assume that this displacement x indicates the "exposure" of the codon in the A-site. If $x = 1$, the first base of the codon in the +1 frame is also exposed in the A-site (see Fig. 5).

The rate at which the polar plot changes direction can be calculated by taking the derivative of the cumulative vector $M_k e^{j\theta_k}$ with respect to codon number k. This derivative turns out to be a vector in itself and we call it the differential vector $\mathbf{D}_k$, when evaluated at codon k. The magnitude and phase of the differential vector, referred to as "differential magnitude" and "differential phase", are given by (1) and (2) respectively. Since the cumulative magnitude

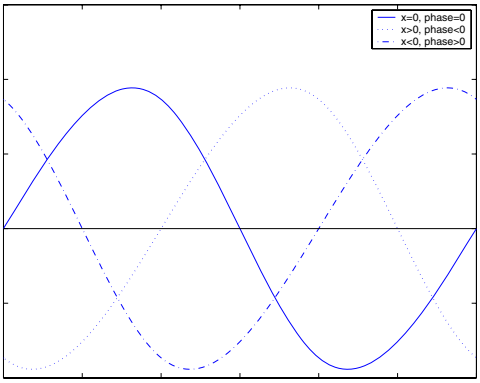

Fig. 4. Illustration of phase and displacement for a sine wave of period 3

Reading frame 0, Perfect exposure, x = 0

```
-------------
|   |   |   |
| U | G | A |
|   |   |   |
-------------
```

Imperfect exposure, x = 1

```
  -------------
  |   |   |   |
| U | G | A | C
  |   |   |   |
  -------------
```

Reading frame +1, Perfect exposure, x = 2

```
-------------
|   |   |   |
| G | A | C |
|   |   |   |
-------------
```

Fig. 5. Exposure of the codon in the A-site and its relationship to displacement x. Shown here is the 26th codon of the *prfB* gene where a shift into the +1 frame occurs.

and phase are both functions of codon position, we compute their derivatives ($\frac{dM}{dc}$ and $\frac{d\theta}{dc}$) using function approximation techniques.

$$|\mathbf{D}_k| = \sqrt{\left(\frac{dM}{dc}\right)^2 + \left(M_k \frac{d\theta}{dc}\right)^2} \tag{1}$$

$$\angle\mathbf{D}_k = \theta_k + \arctan\left(\frac{M_k \frac{d\theta}{dc}}{\frac{dM}{dc}}\right) \tag{2}$$

The incremental displacement dx can be calculated from the change in the cumulative phase, using (4). The accumulation of these incremental displacements gives the total displacement x, as shown in (5). θ_{sp}, also known as species-specific phase angle, is the mean phase angle of the signals corresponding to a set of verified genes in the species being studied. Using 1673 verified genes in *E. coli*, we have found θ_{sp} to be -20.73°. The value of C_1 is fixed at 0.005 to make the equations work for *prfB*.

$$\theta_{dx} = \frac{\pi x}{3} + \theta_{sp} \tag{3}$$

$$dx_k = -C_1 \, |\mathbf{D}_k| \sin(\angle\mathbf{D}_k + \theta_{dx}) \tag{4}$$

$$x_k = \sum_{j=1}^{k} dx_j \tag{5}$$

Finally, there is the issue of the waiting time required to allow the proofreading mechanism to confirm that the correct tRNA is occupying the A-site. We reasoned that this wait-time would be proportional to the availability of tRNAs. We estimate the tRNA availability, γ, using the codon distribution of the verified genes in *E. coli* K-12. This approach is based on the experimentally verified assumption that the availability of tRNAs is proportional to the codon frequency in the mRNA [9]. We will need to take into consideration the fact that multiple codons are decoded by the same tRNA [10]. We use a set of 2438 verified coding sequences in *E. coli*, which have 1053360 codons in all, to estimate the frequency of each codon, f_i, as shown in (6).

$$f_i = \frac{N_i}{N}, \quad i = 1 \ldots 64 \tag{6}$$

where N_i is the number of codons of type i and N is the total number of codons in the dataset. The availability of tRNA for each amino acid is then calculated using (7).

$$\gamma_p = \sum_{i=1}^{n_p} f_i, \quad p = 1 \ldots 21 \tag{7}$$

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 (8).

$$\tau_p = \frac{\max(\gamma) - \gamma_p}{\min(\gamma)} \tag{8}$$

Since stop codons have no tRNAs, they are assumed to have an arbitrarily large wait-time. In the event of an imperfect setting of the codon in the A-site, we choose the wait-time based on the codon having greater exposure. During the wait-time, the incremental displacement, which is calculated relative to the current position, gets added onto the total displacement x_k.

III. RESULTS

For a gene to have no errors during elongation, the displacement x must stay well within the range $-1 < x < 1$. Proximity to either $+1$ or -1 indicates a likelihood of erroneous tRNA recruitment or random frameshift. The displacement plot for *aceF* (Fig. 6) shows that the ribosome always reads the in-frame codon (displacement stays within the range $-1 < x < 1$), indicating no programmed frameshifts within the gene. We also observe that there are no regions where the displacement gets close to ± 1,

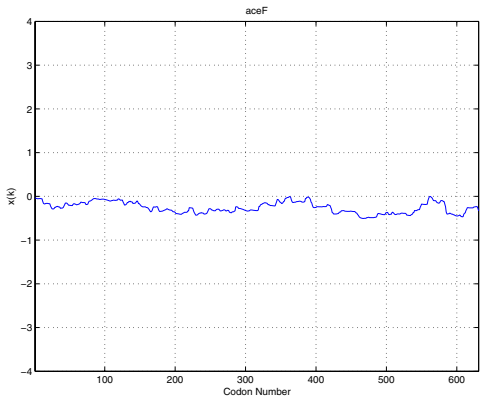

Fig. 6. Displacement plot for *aceF*

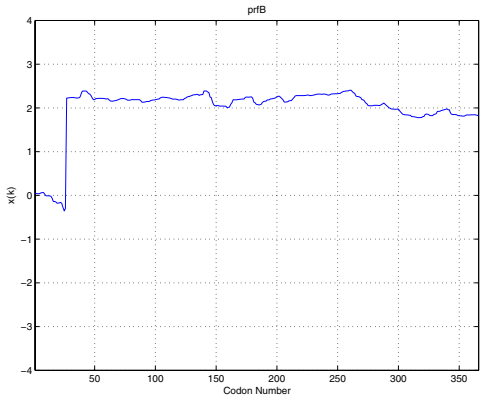

Fig. 7. Displacement plot for *prfB*

indicating a relatively "correct" translation of this gene.

In the case of *prfB* (Fig. 7), the incremental displacement dx is positive at codon 26. Since the stop codon in view has a large wait-time, the accumulated total displacement goes from $x \approx 0$ to $x \approx 2$, and stays roughly constant till the end of the gene.

In order to validate our model, we used 1673 verified genes from *E. coli*, each of which is longer than 200 codons. We found that 1548 of them have displacement tracks confined to the range $-1 < x < 1$. We suspect that the 125 genes that failed the test have random frameshifts in them, which could affect their yield.

IV. DISCUSSION

Using the sequence of a known programmed frameshift gene, *prfB*, we have constructed a model that captures the mechanism of reading frame maintenance in eubacteria. We have validated the model using verified genes in *E. coli*, and found that it explains reading frame maintenance in most of them. With the model in hand, a number of studies are possible. We have found a few genes that do not conform to our expectations, and therefore, warrant further investigation. More detail as to how the subtle patterns of mRNA coding sequences affect signal phase would increase our understanding of the role synonymous codon bias plays in translational efficiency. These studies are currently underway. Although the 3'-terminal nucleotide sequences of bacterial 16S rRNAs are highly conserved, there are differences in sequence. The role these differences play in maintenance of reading frame has not been explored.

We could refine our model in the following ways:

1) The error associated with species-specific phase angle θ_{sp} could be improved by considering known genes of high-yield
2) Under "borderline" conditions, i.e. when $x = 1$, it is ambiguous as to which codon will be chosen. A thorough investigation of this issue is pending.
3) To evaluate an overall probability of error-free elongation, we could evaluate the likelihood of choosing the right codon at each step, based on the exposure x.

Finally, the model needs experimental validation through gene expression studies using synthetic genes for which our model makes specific predictions of their behavior in translation. When validated, our model will be a valuable guide for editing gene sequences used for recombinant protein production and gene therapy.

V. ACKNOWLEDGEMENT

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

REFERENCES

[1] D. Rosnick, D. Bitzer, M. Vouk, and E. May, "Free energy periodicity in E.coli coding," in *Proc. 22nd Annual EMBS International Conference*, 23-28 Jul 2000, pp. 2470–2473.
[2] M. Mishra, S. Vu, D. Bitzer, M. Vouk, and A. Stomp, "Coding sequence detection and free energy periodicity in prokaryotes," in *Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, May 26-27 2004.
[3] L. Ponnala, T. M. Barnes, D. L. Bitzer, and M. A. Vouk, "The search for the optimal ribosome 3 tail end in E. coli," in *Proceedings of the 26th Annual International Conference of the IEEE EMBS*, vol. 4, Sep 2004, pp. 2824 – 2827.
[4] L. Ponnala, T. Barnes, D. Bitzer, M. Vouk, and A. Stomp, "A signal processing-based model for analyzing programmed frameshifts," in *Proc. IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS 2005)*, May 22-24 2005.
[5] L. Ponnala, A. Stomp, D. L. Bitzer, and M. A. Vouk, "Analysis of free energy signals in eubacteria," *EURASIP Journal on Bioinformatics and Systems Biology*, 2006 (submitted).
[6] 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," in *Cold Spring Harb Symp Quant Biol*, vol. 52, 1987, pp. 687–693.
[7] R. B. Weiss, D. M. Dunn, A. E. Dahlberg, J. F. Atkins, and R. F. Gesteland, "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*, vol. 7, no. 5, pp. 1503–1507, 1988.
[8] D. Rosnick, "Free energy periodicity and memory model for genetic coding," Ph.D. dissertation, North Carolina State University, 2001.
[9] T. Ikemura, "Codon usage and tRNA content in unicellular and multicellular organisms," *Mol Biol Evol*, vol. 2, no. 1, pp. 13–34, Jan 1985.
[10] R. F. Weaver, *Molecular Biology*, 2nd ed. McGraw Hill, 2003.