STATISTICAL SIGNIFICANCE AND BIOLOGICAL RELEVANCE OF A SINUSOIDAL PATTERN DETECTED IN TRANSLATIONAL FREE ENERGY SIGNALS

L. Ponnala, A. Stomp, D.L. Bitzer, M.A. Vouk

North Carolina State University, Raleigh, NC, USA

ABSTRACT

Interactions of the ribosome with the mRNA constitute the mechanism of translation in bacteria. We hypothesize that a variable free energy pattern arising from repeated hybridizations of the rRNA tail with a moving window of mRNA sequence could encode information required to maintain translational reading frame. If present, this signal should exist in all coding regions and signal processing methods should prove valuable in characterizing and decoding this signal. Our analysis shows that a variable energy pattern does exist and it can be modeled as a discrete sinusoidal signal of frequency 1/3 cycles/base in approximately 65% of verified coding regions. For the remaining 35% of coding regions, a low signal-to-noise ratio is the most likely explanation of the apparent absence of a detectable periodic signal.

1. INTRODUCTION

Molecular biology studies of bacteria have established that the single-stranded, 3'-terminal nucleotides of the 16S rRNA (rRNA tail) interact via hydrogen bonding with the mRNA during the entire translational process [1][2][3]. Thermodynamics shows that the formation of these hydrogen bonds will result in a change of free energy that favors their formation. Using this approach, software programs, such as MFOLD, have been created that accurately predict RNA secondary structure based on minimization of the free energy of hybridization [4]. Utilizing specific sequence mutations, Weiss et al [1][2] have shown that the degree of sequence complementarity of the rRNA tail with the mRNA can control the shift in translational reading frame in the programmed frameshift of the *prfB* gene in E. coli. Based on these observations, we hypothesize that an information signal could be encoded as variations in free energy that arise as a function of the changing alignments of the rRNA tail with the mRNA as the ribosome moves along the mRNA during translation. If such a signal is present, the methods of signal processing could be used to model information decoding of the signal. Therefore, our first task was to rigorously determine if an energetic signal is present and to characterize this signal. If successful, this work would form the basis for signal modeling to reveal the information regulating the precise shift of translational reading frame encoded in the *prfB* gene sequence.

2. COMPUTATIONAL METHODS

An algorithm developed by Starmer and co-workers [5] was used to calculate the free energy change associated with alignments of an individual gene coding sequence. The algorithm utilizes dynamic programming to identify the minimal free energy conformation of the alignment and the Individual Nearest Neighbor Hydrogen Bond model [6] to estimate the associated free energy value for that conformation. This approach generates a set of free energy values for an entire mRNA sequence indexed by nucleotide position. Our subsequent analysis assumes that this linear array of free energy values constitutes a discrete signal, which can be examined using methods of time series analysis. The only difference is that our signal points are indexed by nucleotide position, instead of time. We use GenBank (http://www.ncbi.nlm.nih.gov/) as our source of data. Based on the annotation provided for E.coli K-12 (accession number: NC_000913), we divide the coding sequences into two categories: (a) verified sequences (i.e. genes with a clearly annotated function) and (b) hypothetical sequences (i.e. genes listed as "hypothetical" or "putative"). We also extract intergenic regions which do not code for any protein and categorize them as non-coding sequences. The 13 bases of the 3'end of the 16S rRNA of E. coli constitute the decoding "rRNA tail". We treat the set of free energy estimates as a discrete signal and verify that it is stationary using the Augmented Dickey-Fuller unit-root test [7]. This allows us to use the periodogram to estimate the power spectral density of the signal [8][9]. The periodogram of the free energy signal for a sample gene *aceF* reveals a dominant frequency of 1/3 cycles/base (Figure 1). The absence of other periodic components suggests that this signal can be modeled as a sum of a sine wave of frequency $f = 1/3$ and noise. We would like to check if this model is appropriate for other coding sequences in E.coli. If a periodic component of frequency $f = 1/3$ does not exist, the signal would be just white noise. So we perform a statistical test of our hypothesis (that a free energy signal contains only white noise) against the alternate hypothesis that it contains a dominant frequency component of $f = 1/3$.

We calculate the discrete Fourier transform (DFT) of our signal $\{x_0,\, x_1,\, \ldots,\, x_{N-1}\}$ as

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N},\ \ k = 0 \ldots (N-1) \tag{1}$$

The periodogram is defined as

$$I_k = \frac{1}{N} \left| X_k \right|^2,\ \ k = 0 \ldots (N-1) \tag{2}$$

The model for our signal is

$$x_n = \mu + A\cos(2\pi f n) + B\sin(2\pi f n) + Z_n \tag{3}$$

where Z_n is Gaussian white noise with variance σ^2, A and B are non-random constants and the specified frequency $f = 1/3$ in this case. The sum-of-squares of the signal can be partitioned by periodic components, based on which we can construct a test of hypothesis [8]. Our null hypothesis is

$$H_0 : A = B = 0$$

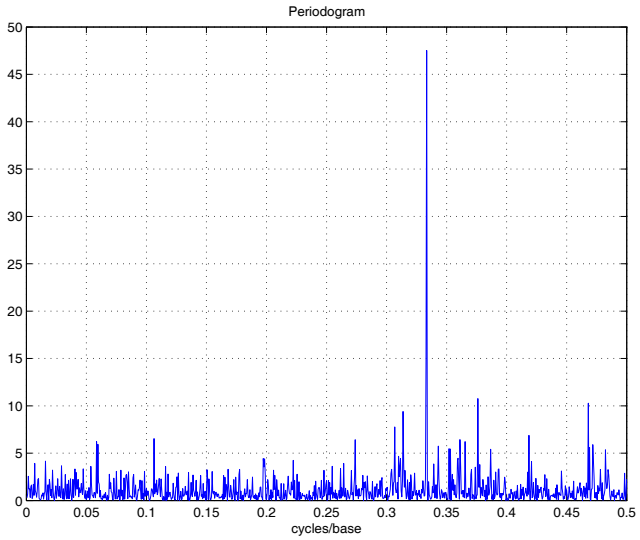

Fig. 1. Periodogram for *aceF*

Type	*Sample Size*	*Passed*
Verified	2438	1579
Hypothetical	1799	956
Non-coding	532	50

Table 1. Detection results

and our alternate hypothesis is

$$H_1 : A \; and \; B \; are \; both \; not \; zero$$

Under H_0,

$$(2I_{N/3}) \;\sim\; \sigma^2\chi^2(2)$$

and $I_{N/3}$ is independent of

$$(\sum_{i=0}^{N-1} x_i^2 - I_0 - 2I_{N/3}) \;\sim\; \sigma^2\chi^2(N-3)$$

We reject H_0 in favor of H_1 at level α if

$$\left[(N-3)I_{N/3}\right] / \left[\sum_{i=0}^{N-1} x_i^2 - I_0 - 2I_{N/3}\right] > F_{1-\alpha}(2, N-3) \quad (4)$$

3. RESULTS

The results of this test for the three categories of E. coli sequences are given in Table 1. The test is performed at level $\alpha = 0.01$. "Sample Size" indicates the number of sequences in each category. "Passed" indicates the number of sequences whose free energy signal shows only one periodic component of the assumed frequency for the hidden periodicity statistical test, i.e., $f = 1/3$. We observe that 64.8% of verified sequences and 53.1% of hypothetical sequences demonstrate strong periodicity at $f = 1/3$ in their free energy signals. On the contrary, only 9.4% of non-coding sequences demonstrate such periodicity.

4. DISCUSSION

It is clear that a periodic signal is encoded in the free energy variation resulting from changes in the degree of hybridization between the rRNA tail and the mRNA as they move by each other during translation. To maintain the correct reading frame, the ribosome must translocate three nucleotides after each amino acid is incorporated into the polypeptide product of the translation process. Therefore, a signal with a dominant frequency of 1/3 cycles/base could encode information to maintain reading frame. If this signal did encode regulatory information for the maintenance of the correct reading frame, we would expect that it should be present in all coding regions. The results of our test (Table 1) indicate that approximately 65% of coding regions have such a signal at the robust, $\alpha = 0.01$ level, a result consistent with our hypothesis. However, 35% of verified coding regions apparently lack this signal. The most likely explanation of this unexpected result is that a low signal-to-noise ratio (SNR) in these sequences could make the signal undetectable at the $\alpha = 0.01$ level chosen for our statistical test. The typical SNR for free energy signals in E.coli is about -18 dB. It would only take an SNR of -20 dB to make the signal statistically undetectable. It is possible that the verified coding sequences that failed to show a strong frequency component of $f = 1/3$ could have higher noise levels. Our test is also not robust if there is more than one periodic component in the free energy signal. These possible alternative explanations are currently under investigation.

5. REFERENCES

[1] 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*, 1987, vol. 52, pp. 687–693.

[2] 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.

[3] E. N. Trifonov, "Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences," *J Mol Biol*, vol. 194, pp. 643–652, 1987.

[4] M. Zuker and P. Steigler, "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information," *Nucleic Acids Res*, vol. 9, no. 1, pp. 133–148, 1981.

[5] J. D. Starmer, "Free2bind: Tools for computing minimum free energy binding between two separate ribonucleic acid molecules," http://sourceforge.net/projects/free2bind/.

[6] T. Xia, J. SantaLucia Jr., M.E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner, "Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs," *Biochemistry*, vol. 37, no. 42, pp. 14719–14735, Oct 20 1998.

[7] J.C. Brocklebank and D.A. Dickey, *SAS for Forecasting Time Series*, Wiley-SAS, 2 edition, 2003.

[8] P.J. Brockwell and R.A. Davis, *Time Series: Theory and Methods*, Springer-Verlag, New York, 2 edition, 1991.

[9] A.V. Oppenheim and R.W. Schafer, *Digital Signal Processing*, Prentice-Hall, 1 edition, 1975.