Proceedings of the 22nd Annual EMBS International Conference, July 23-28, 2000, Chicago IL.

Free Energy Periodicity in E.coli Coding

D. I. Rosnick*, D. L. Bitzer, M. A. Vouk, E. E. May

Abstract— **Sequences upstream from coding regions in *E.coli* commonly possess significant complementarity to the exposed part of the 16S rRNA. This region is known as the Shine-Dalgarno sequence. Free energy calculations for binding between homologous sequences suggest that this region is used as a landing site for construction of the ribosome around the mRNA. While strong upstream binding appears to be a condition for translation, it may not be sufficient. Our research suggests that the 16S has a continuing role throughout translation, particularly in ribosomal synchronization with the reading frame. We consider the entire *E.coli* genome of over 2000 forward coding sequences. Presence of strong upstream binding is confirmed, and a definite three-base periodic signal is observed. The distribution of bases parallels that needed to produce a signal of the type observed.**

Keywords— **16S, Shine-Dalgarno, Free Energy**

Introduction

Analyses of features such as coding length and presence of particular sequences upstream from a start codon are often employed to identify true coding sequences from a theoretical standpoint. One example is the Shine-Dalgarno sequence. It is an identifier near the 3' end of the 16S rRNA ribosomal subunit which frequently displays a strong homology between its Watson-Crick complement and regions upstream from the coding sequences [6]. Research suggests that the distribution of bases within a sequence relates to translocation and frameshifting [9]. Though extensively studied and modestly employed in coding detection, the idea of phase bias as a necessary condition of proper translation remains controversial (*e.g.*, Curran [3].

The "Shine-Dalgarno" (SD) sequence we consider here the sequence GAUCACCUCCUUA, read 5'-3' from the 16S, and its DNA complement (cSD) TAAGGAGGTGATC. Homology between cSD and upstream mRNA and free energy release due to Watson-Crick binding between SD and the the upstream sequences have both shown to relate to translation rates of the coding regions to which they correspond. The presence of the cSD in these region is not, however, the sole translation indicator [2].

If upstream presence of the sequence is necessary for initiation of translation, then elongation may mandate continued affinity between the 16S and mRNA well into a coding sequence. Experimental research into downstream homology indicates that the Shine-Dalgarno sequence impacts the efficiency of a shifty stop (*e.g.*, Weiss [9]). Here, those affinities are examined through free energy calculations. Previous examinations have been generally limited to the upstream regions [5], or simpler base preferences [7]. We consider the entire coding region, and eventually, the entire *E.coli* genome. Free energy calculations are performed based on calcuations used in the formation of mRNA secondary structure [4]. For every coding sequence, the minimum free energy at every base position is computed. The average energy at each position on the messenger strand taken over 2095 coding sequences produces an obvious three base periodic signal down the coding region. Preference for one phase of the signal relative to the start codon is also examined. The RF-2 gene is singled out as particularly demonstrative of the phase necessity, as the energy pattern shifts along with the natural +1 frameshift.

Although any unequal distribution of the bases among the three positions should result in a correlation of the free energy values, the observed base preference broadly corresponds to the distribution expected for three-periodic homology with the SD sequence. Because of the high noise level in the energy pattern, it is suggested that a memory mechanism exists within the ribosome to monitor the pattern over the length of the strand.

It is then suggested that the SD portion of the 16S rRNA plays a role as the ribosome maintains frame through the coding region, and therefore on downstream base preference.

Methods

Data we used in these computer experiments is taken from the National Institue of Health genome database, GenBank [1]. We ran original C routines and Matlab analysis tools on a Sun Sparc Ultra-5 workstation running Solaris 2.5.1.

We developed C computer program to read in GenBank datafiles containing a listing of a DNA sequence, annotated with actual and proposed coding sequences, and to then compute the binding energies between the sequence and an input mask from that data. The data is then analyzed and results visualized using simple Matlab routines.

We presuppose that the *E.coli* mRNA transcribes exactly from the DNA sequence listed in the GenBank datafile accession code U00096. We also assume the accuracy and completeness of the GenBank data. This is especially important with regard to the annotation of coding sequences. Particularly, for the purposes of this research, a coding sequence is any region annotated "CDS", although distinctions are made for uncertain sequences listed as "hypothetical," "putative," or simply unknown ORF.

We employ a simple method for computing the free energy binding strengths based on base doublets [4]. Isolated energy calculations are performed using a dynamic programming approach for efficiency in time.

The location of binding sites are identified by the alignment of the 16S and the target strand. Regardless of whether binding takes place or not, the location is determined by the base number where the 3' end of the SD aligns. We ran the program on the *E.coli* K-12 genome

*Department of Computer Science
North Carolina State University
Raleigh, NC 27695-8206 USA
dirosnic@eos.ncsu.edu

0-7803-6465-1/00/$10.00 ©2000 IEEE

with the SD mask for regions 30 bases upstream to 3000 bases downstream from the start codon for every forward coding sequence listed. The program output average binding strengths for each position relative to the respective start.

Results

The average binding strengths around the start codons are seen in Figure 1. Average binding strengths confirm

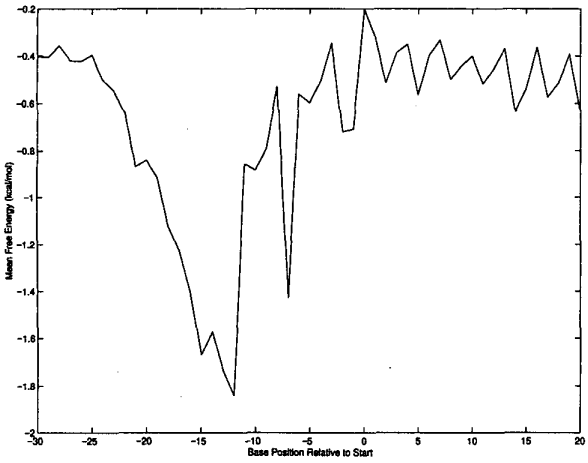

Fig. 1. Average Binding Strength – Near Start

the presence of the SD sequence in the -16 to -12 positions. This region is as close to the beginning as possible without overlapping the start codon. Additionally, there is an energy spike at position -7, which corresponds to start codon affinity to the CA doublet in the SD sequence.

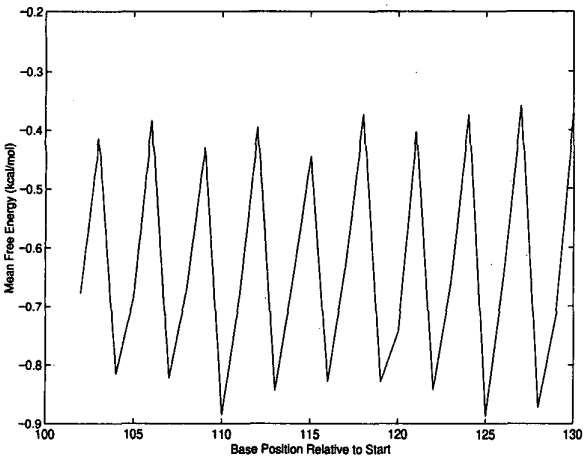

Fig. 2. Average Binding Strength – Downstream

The average binding energies starting 102 bases downstream from the start codon are shown in Figure 2. The binding preference is clearly seen from the sample, but we confirm with Fourier analysis of the signal.

Figure 3 shows the power spectrum density of the energy by binding position. The frequency is given in units of cycles per nucleotide. This figure shows, at frequency

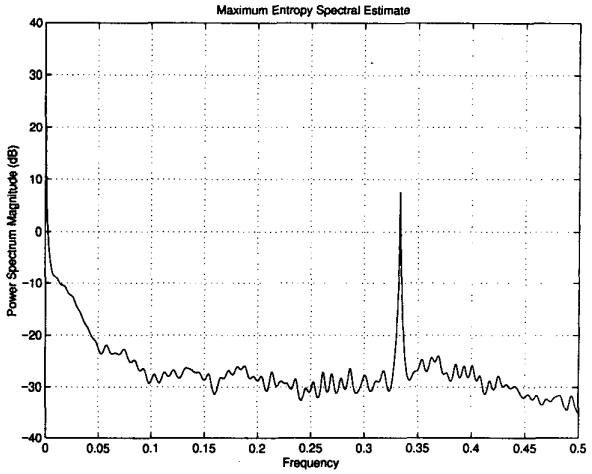

Fig. 3. Power Spectral Density – Coding Regions

one-third, a peak power of 36dB. This is equivalent to a signal to noise ratio on the order of 4,000 to 1. This indicates a pattern of binding between the SD and mRNA every third base position. Furthemore, the pattern, if it exists in individual coding sequences, must be biased toward a particular phase, otherwise the individual signals would cancel each other out. Assuming all the samples are in phase, and a 10db per factor of ten in the number of samples, the power should be reduced in an individual sequence by

$$10 \times \log_{10} 2095 = 33dB$$

Thus, an approximate signal to noise ratio of three decibels may be expected in each coding sequence.

The top spectrum of Figure 4, shows the spectral estimate for the gene aceF, which codes for pyruvate dehydrogenase. It has a 12 decibel signal at period three. We also compute the spectral estimate corresponding to an 1830 base long ORF unlisted in GenBank's annotation. It is located between the putative transport gene emrB, and transport gene srlA. The spectrum of the latter (shown in the bottom of Figure 4) reveals no periodic signal of significance, let alone in the vicinity of period three.

For each coding sequence listed in GenBank, we compute the average free energy for each position modulo 3 relative to the start. The three positions (0, 1, 2) were assigned a binding pattern based on the relative averages (Strong, Medium, Weak). Table I shows the results of these computations. The numbers are shown for both the full CDS listing, and for the certain sequences only. Both the number of sequences, and their average length are shown. Note the significant preference for stronger binding in the third positions relative to the second. 99.1% of the certain coding sequences exhibited this property, as did 97.2% of the uncertain, but suspected, coding regions.

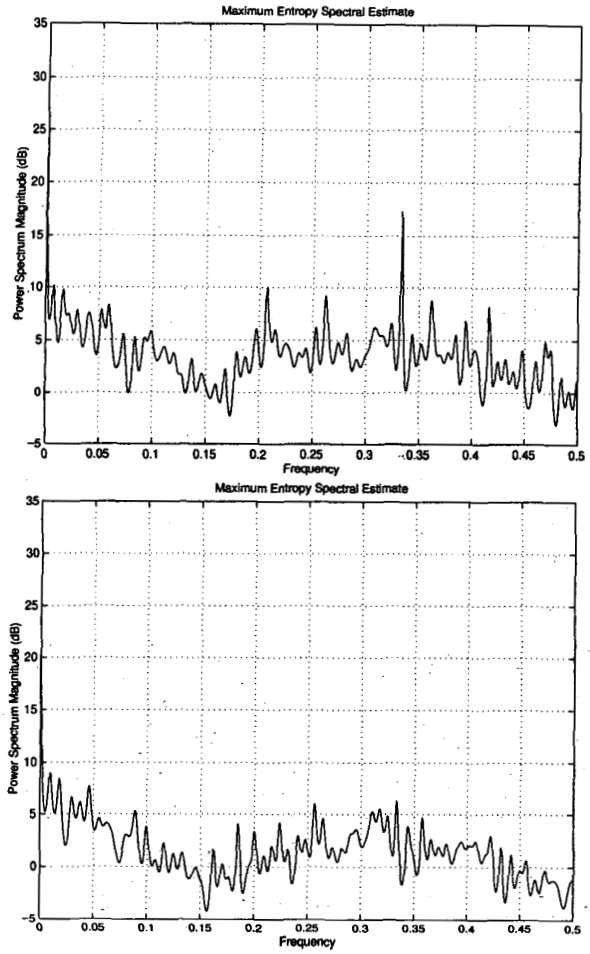

Fig. 4. Example Power Spectrum For Single Sample

TABLE I
BINDING PATTERNS: NUMBER AND AVERAGE LENGTH

Pattern	Certain	Length	All	Length
SMW	3	1042	24	440
*SWM	124	820	426	760
**MWS	598	1178	1570	1033
MSW	1	102	4	320
WSM	3	522	10	365
*WMS	13	426	61	549

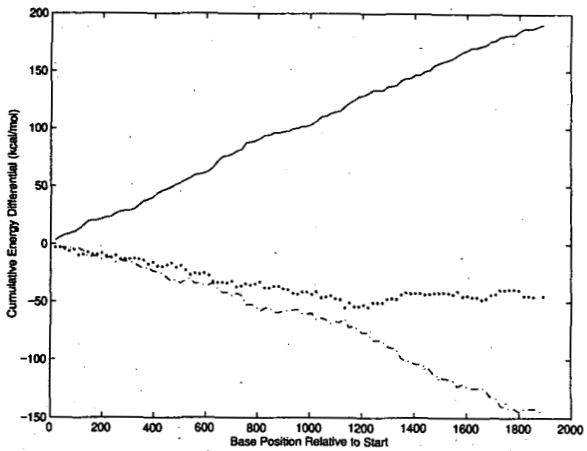

Fig. 5. aceF Energy Differential by Position

As an example of the building energy pattern, the cumulative energies in each position were calculated for aceF. The result is shown in figure 5. For clarity, at every point, the cumulative energy over all upstream positions is subtracted. Thus, the difference in the cumulative sum for each phase, and the overall total, is shown. The dotted line represents the accumulation of energy in the first position. The solid line represents that for the second. Note that the positive energy differential indicates weak binding. Strong binding occurs with negative free energy. The third position is seen with on the broken line. It shows the strongest binding. Thus, aceF falls into the most common category of binding patterns, MWS.

Finally, we examined the RF-2 gene prfB. The binding was computed disregarding the natural frameshift near the beginning of the sequence. We employed the same calculation for prfB as previously with aceF. The result is shown in Figure 6. Note the change in binding preference in the vicinity of the +1 frameshift. prfB exhibits MWS binding for the first 100 or so bases, then shifts to SMW. This is

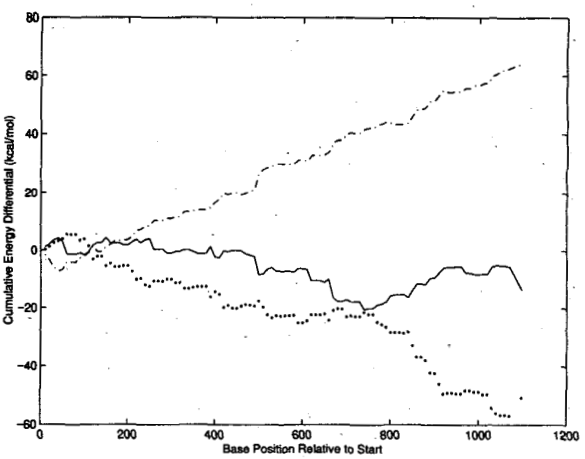

Fig. 6. prfB Energy Differential by Position

TABLE II
BASE DISTRIBUTIONS IN *E.coli* (PERCENTAGE OF REPRESENTATION BY POSITION)

	Theoretical				Observed		
	I	II	III		I	II	III
A	30.1	35.9	12.8	A	24.7	30.4	15.9
T	14.7	33.3	25.6	T	14.0	28.4	24.0
G	35.3	12.8	35.9	G	36.1	18.3	30.6
C	19.9	17.9	25.6	C	25.2	22.9	29.4

identical to MWS, having skipped forward one base.

Next, we investigate the relation of to the energy signal to the actual SD sequence. What must a DNA coding sequence look like in order to produce the binding pattern of the previous section?

To answer this question, we assume that there is an ideal binding pattern, namely, MWS. If there is no bulging of the bases as they bind, each base in the DNA aligns once with each of the 13 bases in the SD. Because the base positions 0, 3, 6, 9, and 12 have neutral binding, no base preference is indicated for these positions. In positions 1, 4, 7, 10, and 13, weak binding is in order, so bases should have a tendency not to pair up in these positions. For positions 2, 5, 8, 11, and 14, the opposite is true.

For example, take the DNA base 12 downstream from the first base of the start codon. Because this base corresponds to the first base of a codon, any indication of preference made with it will correspond to the general preference of the first bases. We now convolve the SD and the DNA to extract the base distribution. For example, we start with the SD so the 5' end (G) aligns with base 12. That leaves the 3' end (A) with the first base of the start codon. This is binding position zero, so there is no binding preference. Advancing the SD by one base, an A now associates with base 12. Because this binding position is weak, this indicates a bias against T in the first base of a codon. Advancing once more, the U indicates a bias in favor of A.

We next apply the thirteen preference conditions for each of the three base positions obtain an estimate of the distributions in each base position. Where there no preference is indicated, we assume the bases are equally distributed. When a base is biased against, it is assumed that the distribution is equal among the remaining bases. And where there is a bias in favor, it is assumed that the base is always the desired one. For each position, the thirteen distributions are given equal weight, resulting in the final base distribution of Table II. The observed distribution was obtained from all the certain forward coding sequences of length 1500nt or greater. Although the individual probabilities differ by as much as 5.5%, the relative order of the bases is conserved by position. The G/A preference over T in position 1, the A/T preference over G in position 2, and the weak representation of A in position 3 are all seen clearly in the data, despite the extreme simplicity of the model. With respect to guanine preference, this agrees with Trifonov's GHN phase bias [7].

DISCUSSION

The results show that there is a strong three base period present in the coding region. They also suggest that the signal disappears in noncoding sequences. Examinations bear this out in the samples we examine. Because the DNA coding is presumed to be efficient in prokaryotic organisms, it seems unlikely that such a signal fails to serve a real purpose with regard to the actual genetics. The preference for a particular phase of the signal relative to the start codon indicates that the signal may be used to synchronize the ribosome as it travels along the coding sequence in a three base fashion as expected by the genetic code.

Furthermore, even a simple model indicates that the expected base preference parallels the observed distribution of bases, although statistics do not indicate that signal maintanance is the sole influence on the base preference.

We conjecture that three major conditions must be met for proper translation of a coding sequence. First, a sufficiently long coding region is necessary for the production to be obeservable. Second, there must be a sufficiently strong cSD sequence upstream from the region to be translated. This may aid in initializing the energy within the ribosome as well as attract the 16S to begin construction thereof. Third, there must be a periodic synchronization signal to maintain the reading frame. The addition of the third condition has two major consequences. First, it could be used as a major identification mechanism in DNA analysis of coding sequences. Second, it constrains the sequences that are actually codable, which then gives additional reason for the multiple codings to each amino acid.

REFERENCES

[1] D. A. Benson et al. GenBank. *Nucleic Acids Res.*, 26(1):1–7, January 1988.

[2] P. Bermel, E. Eni, M. Vouk, and D. Bitzer. On the import of the Shine-Dalgarno series to the expression of mRNA sequences. Department of Computer Science, North Carolina State University.

[3] J. F. Curran and B. L. Gross. Evidence that GHN Phase Bias does not Constitute a Framing Code. *J. Mol. Biol.*, 235:389–395, 1994.

[4] B. Lewin. *Genes VI.* Oxford University Press, New York, NY, 1997.

[5] T Schurr, E. Nadir, and H. Margalit. Identification and characterization of *E.coli* ribosomal binding sites by free energy calculation. *Nucleic Acids Res.*, 21:4019–4023, 1993.

[6] J. Shine and L. Dalgarno. The 3'-terminal sequence of E.coli 16S ribosomal RNA: Complementarity to nonsense triplets and ribosome binding sites. In *Proc. Natl. Acad. Sci.*, volume 71, pages 1342–1346, 1974.

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

[8] E. N. Trifonov. Recognition of the correct reading frame by the ribosome. *Biochimie*, 74:357–362, 1992.

[9] R. B. Weiss et al. Reading frame switch caused by base-pair formation between the 3' end of the 16S rRNA and the mRNA during elongation of protein synthesis. *EMBO J.*, 7:1503–1507, 1988.