Proceedings of the 26th Annual International Conference of the IEEE EMBS
San Francisco, CA, USA • September 1-5, 2004

The search for the optimal ribosome 3' tail end in E. coli

L. Ponnala[1], T.M. Barnes[2], D.L. Bitzer[2], MA. Vouk[2]
[1]Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, USA
[2]Department of Computer Science, North Carolina State University, Raleigh, NC, USA

Abstract—**The 16s ribosomal tail end has been conjectured to play an important role in the regulation of protein production and of translation efficiency. Using E. coli K-12 as our model organism, we generate sequences of 13 base pairs as hypothetical ribosome tail ends. We analyzed the distributions of these random hypothetical ribosome tail ends and found the actual E. coli ribosome tail end to be significantly different from a randomly generated ribosome tail in the magnitude of the lock and synchronization signals, and the signal to noise ratio. We then designed and ran a Genetic Algorithm to optimize hypothetical ribosome tail ends simultaneously for these three signal criteria. We found that the actual E. coli ribosome tail end was among the best by these measures.***

Keywords—**translation, coding, ribosome, bacteria**

I. Introduction

Transgenic protein production is an important biotechnological advance, offering a method for producing large quantities of necessary proteins at low cost. Its effectiveness and efficiency, which strongly affect its cost, are determined by adjusting foreign messenger RNA (mRNA) to be acceptable to both the host environment and the host ribosome. Without these adjustments, proteins may not be produced in sufficient quantities. However, the process of determining necessary adjustments is complex and often involves much trial and error [1].

The tail end of the 16s ribosomal subunit appears to play an important role in the translation process in prokaryotic organisms. An improved understanding of this role and the interactions of the 16s tail with mRNA may therefore lead to significant advances in genetic engineering.

An important feature of the ribosome is the strong affinity of its exposed 3' tail end to an identifier called the Shine-Dalgarno sequence that is located roughly 13 bases upstream of the start codon. This interaction can be modeled by calculating the free energy released due to the binding of the ribosome's tail to the messenger RNA. This free-energy release is interpreted as a signal that appears to be a good indicator of the regulation that takes place during protein production, and of the translation efficiency. Previous work suggests that this regulation has two parts: a "lock" and a synchronization signal. The lock is located at or just before the start codon. It appears to reflect the need to bind or pause the ribosome long enough close to start codon for it to lock into the reading frame and to start protein production. Once the lock is achieved and protein production starts, the synchronization signal must be strong enough, and in the right phase to maintain the reading frame. In this model, the tail end of the ribosome should be able to detect these features in the genome for optimal translation.

In this research, we explore this model by creating random sequences of 13 base pairs as hypothetical ribosome tail ends for *E. coli*, and assessing each based on signal criteria. We found the actual *E. coli* ribosome tail end to be significantly different from a randomly generated ribosome tail in the magnitude of the lock and synchronization signals, and the signal to noise ratio. We then designed and ran a genetic algorithm (GA) to optimize hypothetical ribosome tail ends simultaneously for these three signal criteria. We found that the actual *E. coli* ribosome tail end was among the best by these measures.

In the following section we introduce the concepts of free energy calculations and genetic algorithms. In Section II, we discuss our statistical analysis of a random ribosome tail end and the genetic algorithm we used to search for optimal ribosome tail ends. We present and discuss our findings in Section III. Section IV concludes our paper and suggests paths for future work.

A. Ensemble Average Signal

In this section, we discuss a method to analyze the free energy released during genetic translation using signal processing techniques. A particular alignment of the 3' exposed tail end against the messenger RNA is referred to as a conformation. The binding energy released in this conformation, also referred to as free energy, is estimated using the method of base-doublets [2]. This calculation penalizes mismatches and rewards consecutive base pairing in the conformation. A shift along the mRNA by one base position results in a new conformation, and the calculation of the free energy estimate is repeated. The binding energies for matching doublets are determined by experiment and are listed in [3].

The set of free energy estimates for all possible conformations along the mRNA sequence constitutes a discrete signal that can be analyzed using methods of discrete-time signal processing [4]. The signal is calculated for each individual coding sequence along the forward strand in *E. coli*, and the ensemble average of 531 such signals is plotted (See Appendix for equations). For the remainder of this paper, we refer to this as the ensemble

[*] This work was supported in part by the DOE SciDAC grant DE-FC02-01ER25484, and the IBM Corp. Shared University Research Program.

average signal, which, for each conformation, is expressed in units of kcal/mol, referred to as E.

Fig.1 demonstrates the ensemble average signal calculated by averaging the signals obtained by matching the tail end of the *E. coli* 16s ribosomal subunit to the 531 certain coding sequences for *E. coli* K-12 available in Genbank [5] (See Appendix). The dip in this signal, interpreted as a "lock", occurs roughly 13 bases upstream from the start codon, indicating strong affinity of the tail end to the Shine-Dalgarno consensus sequence that resides here [6]. About 90 bases downstream, we observe that the signal becomes strongly 3-base periodic. We will refer to this downstream signal as the "synchronization" signal, since it appears to reflect how the ribosomal subunit moves along the mRNA sequence till the formation of the polypeptide chain is complete [2].

B. Genetic Algorithms

Genetic Algorithms (GAs) are numerical optimization techniques based on a generalized theory of evolution and natural selection, and have been used to solve a variety of problems such as the selection of optimal convolutional codes [7] and table-based codes for genetic translation initiation [8]. There are 4^{13} different sequences that may be considered as hypothetical tail ends for the 3' end of the 16s ribosomal subunit, a number which would be prohibitive for performing an exhaustive search. As discussed above, consecutive base-pairings between the ribosome tail end and the mRNA result in higher free energy release, suggesting that the complements of frequent mRNA base sequences will be important patterns in candidate tail ends. Since GAs emphasize patterns such as these in searching for optima, we chose to use a genetic algorithm to search for optimal ribosome tail ends.

TABLE I
Signal characteristics for E. coli ribosome tail end

Sync. signal magnitude	0.1274 E
Lock magnitude	–0.7605 E
Signal to noise ratio	–10.46 dB

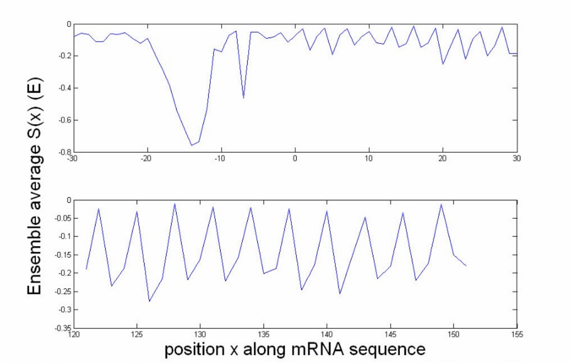

Fig. 1. Ensemble average S(x) for actual E. coli ribosome tail

II. Methodology

Our experiment consists of three main parts: 1) calculation of the ensemble average signal and evaluation of its characteristics, 2) generation and analysis of 100 random hypothetical tails, and 3) design of a Genetic Algorithm to search for optimal solutions.

A. Ensemble Average Signal and Characteristics

For each candidate ribosome tail end, we calculate the ensemble average signal (see Appendix) over a set of 531 certain coding sequences obtained from Genbank [5]. From this signal, we determine three parameters: 1) the magnitude of the synchronization signal, 2) the magnitude of the lock signal (which will be negative, since it represents a free energy "release"), and 3) the signal-to-noise ratio. Table 1 lists these parameters for the actual tail end in *E. coli.*

The magnitude of the synchronization signal (sync) is estimated using a method that takes advantage of our prior knowledge of its periodicity. We calculate running averages of every third position along the signal, and interpolate a sine wave through the three resulting points (see Appendix). This method of calculating the magnitude works well in the presence of immense noise, which is characteristic of the *E. coli* genome [2].

An estimate of the "pure" signal is obtained using the calculated magnitude and phase. This estimated signal is subtracted from the noisy ensemble-average signal to get the noise signal. The ratio of the variance of the estimated "pure" signal to the variance of the noise yields an estimate of the signal-to-noise ratio (SNR).

The affinity of the 16s tail end to the Shine-Dalgarno sequence is measured by the minimum magnitude of the signal, between positions 16 and 12 bases upstream from the start codon. This is referred to as the "lock" magnitude.

B. Randomly Generated Hypothetical Ribosome Tails

To confirm our hypothesis that the actual *E. coli* ribosome tail end is significantly different from a random sequence of 13 bases, we plotted the distribution of the lock, synchronization, and SNR criteria for 100 randomly generated hypothetical ribosome tail ends.

Figs. 2, 3, and 4 show the probability distributions for the synchronization, lock, and SNR characteristics of the ensemble average signals for 100 random 13-base sequences treated as ribosome tail ends. In each figure a point shows the value for the actual ribosome tail. We can visually deduce that the actual ribosome tail end is significantly different from a random sequence of 13 bases in these characteristics. Since the actual ribosome has extreme values in these characteristics, we can also deduce that few random sequences of 13 bases will have extreme values similar to the actual tail end. These observations were important in constructing our fitness function to evaluate hypothetical ribosome tail ends.

C. The Genetic Algorithm Search for Optimal Tail Ends

Since we hypothesized that the most important features of a ribosome tail end for optimal translation are the lock and synchronization signal magnitudes, and the SNR of the ensemble average signal, we designed an objective function to simultaneously optimize for these features.

The objective function, or fitness, for a given ribosome tail end is computed in three steps. First, we compute the lock, synchronization, and SNRs as discussed above. We then assume that each of these features is normally distributed and calculate one-sided p-values for each feature. Third, the total fitness of a candidate 13-base tail end is the sum of these three p-values, and the GA optimizes for the minimum of this sum. Our GA, with population size 100, mutation rate 0.1, and no crossover, was then run using this function to search for optimal ribosome tail ends.

III. Results & Discussion

The first five rows of Table 2 list five optimal hypothetical ribosome tail ends whose fitness functions were better than that of the actual *E. coli* ribosome tail end. For each of these, the lock and synchronization signal magnitudes were higher than those of the actual tail. For all but tail 5, their SNRs were also better than that of the actual. On inspection we see that these five, as well as tails 7 and 8, have a strong lock signal that corresponds to having strong complementarity to the Shine-Dalgarno sequence. In other words, these sequences contain subsequences similar to UCCUCC, differing in one base or less.

We also list some hypothetical tails whose fitness was better than that of the actual, demonstrating the ability of the GA to optimize on all three signal criteria. Tail 6 has a better signal to noise ratio than the actual tail, but its lock magnitude is much worse than that of the actual, and may not be sufficient for a lock. Visually, we can confirm that this tail does not have a subsequence with a good match to the Shine-Dalgarno sequence.

Tails 7-8 in Table 2 have a decent lock magnitude and match to the Shine-Dalgarno sequence. However, their signal to noise ratios are lower than that of the actual. Tail 7 has a higher synchronization signal magnitude than the actual, while tail 8 does not. Since we believe that good performance in all three parameters is important, these results show that the fitness function was appropriate in optimizing for the lock and synchronization signal magnitudes and the signal to noise ratio.

TABLE 2
Optimal 13-base ribosome 3' tail ends

#	Ribosome	Lock (E)	Sync. Mag. (E)	Sync. Phase (rad)	SNR (dB)
1	GACCUCCUCCUGC	-0.919	0.224	1.507	-9.42
2	GACCUCCUCCUGA	-0.912	0.192	1.440	-10.13
3	GACCUCCUCCUAA	-0.863	0.179	1.494	-10.36
4	GACCUCCUUCAGU	-0.778	0.173	1.444	-9.84
5	GUCCUCCACCUGA	-0.843	0.168	1.247	-10.80
Actual	**AUUCCUCCACUAG**	**-0.760**	**0.127**	**-0.161**	-10.46
6	GCCUUGACCUGCU	-0.453	0.222	-0.485	-8.66
7	CUCCGCCUCUUGA	-0.633	0.154	0.853	-12.64
8	UCCUCAAGCUCGU	-0.712	0.109	-0.681	-12.36

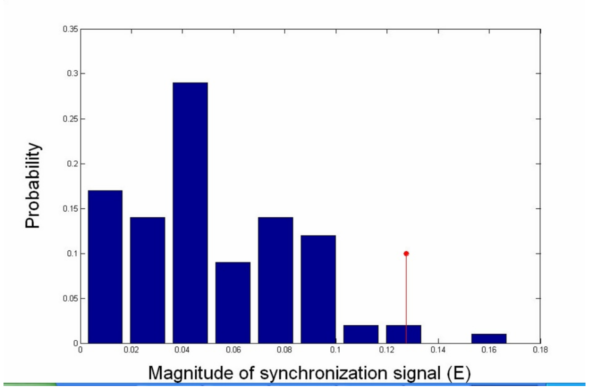

Fig. 2. Distribution of Sync. Magnitudes for Random Sample

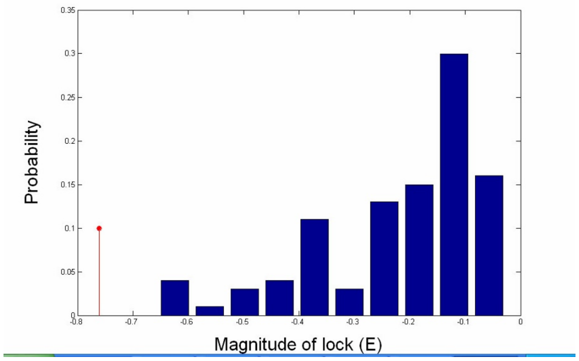

Fig. 3. Distribution of Lock Magnitudes for Random Sample

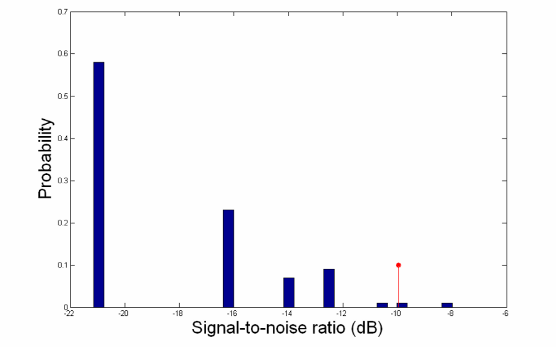

Fig. 4. Distribution of SNRs for Random Sample

IV. Conclusion

In this work, we investigate signal-processing characteristics of the tail end of the 16s ribosomal subunit, to obtain a better understanding of the interactions of the 16s tail with mRNA. This understanding can lead to significant advances in genetic engineering.

This investigation had two main parts: 1) to show that the actual *E. coli* ribosome tail differed significantly from a random hypothetical tail in signal characteristics, and 2) to find other candidate ribosome tails with similar signal characteristics. The first result supports the conjecture that the actual ribosome tail may have been selected by nature, using the lock, synchronization, and SNR characteristics, to be effective in translating the genes of the species. The second provides a list of candidate tails that can be analyzed to discover other properties that are important for the ribosome tail.

This research showed that the actual *E. coli* ribosome tail did differ from a random one in the lock and synchronization signal magnitudes, and also in the signal to noise ratio. This finding suggests that the natural selection of the *E. coli* ribosome tail was not random relative to these characteristics. Because of their similarity to the actual ribosome tail, the hypothetical tails found by the GA offer additional evidence that the selection of the ribosome tail is not random.

The fitness function was able to distinguish candidate tail ends, and optimize for only those with favorable values for all three signal criteria. Future work can extend this fitness function to encompass other criteria that may be important in genetic translation.

In conclusion, our findings suggest that the "ideal" ribosome tail needs all of these characteristics: a strong lock to initiate protein production, and a strong synchronization signal, that is well-differentiated from noise, to drive it along. In future work we plan to extend the method to other prokaryotes (and possibly eukaryotes) to find "optimal" ribosome tail ends.

References

[1] B. E. Schoner et al. "Role of mRNA translational efficiency in bovine growth hormone expression in Escherichia coli." Proc. Natl. Acad. Sci. USA, 81:5403-5407, 1984.

[2] D. I. Rosnick. "Free Energy Periodicity and Memory Model for Genetic Coding," PhD Dissertation, North Carolina State University.

[3] S. M. Freier et al. "Improved free-energy parameters for predictions of RNA duplex stability," *Proc. Natl. Acad. Sci. USA*, 83: 9373-9377, 1986.

[4] A. V. Oppenheim and R. W. Schafer. *Discrete Time Signal Processing*, Prentice Hall, March 1989.

[5] The genome of Escherichia Coli K-12, available online at ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/ Escherichia_coli_K12/U00096.gbk

[6] J. Shine and L. Dalgarno. "The 3'-terminal sequence of E.coli 16S ribosomal RNA:Complementarity to nonsense triplets and ribosome binding sites" Proc. Natl. Acad. Sci. USA, 71(4):1342{1346, 1974.

[7] Tiffany M. Barnes, "Using Genetic Algorithms to Find the Best Generators for Half-Rate Convolutional Coding," North Carolina State University, Raleigh, NC, 1994.

[8] E. E. May, "Analysis of Coding Theory Based Models for Initiating Protein Translation in Prokaryotic Organisms," PhD Dissertation, North Carolina State University.

VI. Appendix

Let $e_i(x)$ denote the free energy score estimate for gene *i* at position *x* along the mRNA as computed in [2]. Let *N* denote the number of genes in the sample. The ensemble average signal $S(x)$ at position *x* is:

$$S(x) = \frac{1}{N}\sum_{i=1}^{N} e_i(x) \qquad (1)$$

In the following, we calculate the magnitude M and phase ϕ of the synchronization signal. For a coding sequence of length *k* codons, we calculate three quantities, *A*, *B* and *C*. We begin these calculations 90 bases, i.e. 30 codons, downstream from the start codon, after which strong periodicity of the signal is observed [3].

$$A = \frac{1}{k-30}\sum_{x=90,93,96,...}^{3k-3} S(x) \qquad (2)$$

$$B = \frac{1}{k-30}\sum_{x=91,94,97,...}^{3k-2} S(x) \qquad (3)$$

$$C = \frac{1}{k-30}\sum_{x=92,95,98,...}^{3k-1} S(x) \qquad (4)$$

These quantities represent the average signal over the entire coding sequence. We subtract the constant DC term from these quantities to remove any bias, resulting in the points *a*, *b*, and *c*, given in (6).

$$DC = (A + B + C)/3 \qquad (5)$$

$$a = A - DC, b = B - DC, c = C - DC \qquad (6)$$

To estimate the synchronization signal, we interpolate a sine wave (of magnitude *M* and phase ϕ) through these three points using the formulae given below:

$$a = M\sin(\phi) \qquad (7)$$

$$b = M\sin(\phi + 2\pi/3) \qquad (8)$$

$$c = M\sin(\phi + 4\pi/3) \qquad (9)$$

$$\phi = \arctan(\sqrt{3}a/(a+2b)) \qquad M = a/\sin(\phi) \qquad (10)$$