Coding Model for Translation in E. coli K-12

Elebeoba E. May*, Mladen A. Vouk, Donald L. Bitzer, and David I. Rosnick
*Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC
*email: euenil@eos.ncsu.edu

Abstract— **Various computational methods for determining protein producing regions in a given genome are based on engineering constructs such as Hidden Markov Models and neural networks. This work presents a coding theory approach to modeling the process of protein translation initiation. The messenger RNA (mRNA) is viewed as a noisy encoded message and the ribosome as a block code minimum distance decoder. The results of applying this model to the Escherichia coli K-12 are presented.**

Keywords— **Coding Theory, Translation Initiation**

I. Introduction

If genetic information in the DNA sequence is encoded in a manner equivalent to block encoding, the received message, the mRNA, should conform to the block coding model [1]. Using basic block coding principles we developed a method and a decoding model based on chemical and biological characteristics of the ribosome and the ribosome binding site, located in the leader region of the mRNA [2].

II. Methods

We used principles of base pairing, wobble pairing, and translation initiation [2] to define a coding alphabet on the field of 5. The alphabet consists of inosine (I=0), adenine (A=1), guanine (G=2), cytosine (C=3), and uracil (U=4). We model the genetic encoder as a block code whose output is a systematic zero parity check code [1] [3]. An (n,k) code was developed based on the last thirteen bases of the 3' end of 16s ribosomal RNA, which contains the hexamer complementary to the Shine-Dalgarno sequence [2]. We designed a minimum distance decoder to verify the block coding model of translation initiation.

III. Results

The Escherichia coli K-12 strain MG1655 sequence data (downloaded from the NIH ftp site: ncbi.nlm.nih.gov) was used to test the model. Figure 1 shows the resulting mean minimum distance by position for the (5,2) block code model. The smaller the value on the vertical axis, the stronger the bond formed between the ribosome and the mRNA. Zero on the horizontal axis corresponds to the alignment of the first base of a codeword with the first base of the initiation codon.

IV. Discussion

As Figure 1 illustrates there is a significant difference between the translated, hypothetical and the non-translated sequence groups. For the translated and hypothetically translated sequence groups, a minimum distance trough occurs between the -15 and -10 regions. All the sequence groups in the (5,2) model achieve a global minimum mean

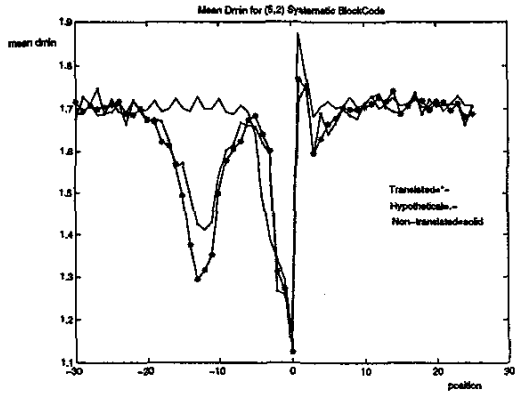

Fig. 1. Results of Minimum Distance Block Decoding Model for (5,2) Code

distance value in the -5 to 0 region. The -15 to 0 region contains large synchronization signals which can be used to determine valid protein coding sequences or frames. There are also smaller synchronization signals outside the -15 to 0 region which seem to oscillate with a frequency of three. The results of the longer (8,2) block code model (presented in [3]) illustrate the effect of two or more codons while the (5,2) block code model is affected by at most two codons.

V. Conclusion

The block code method shows distinction between translated sequence groups and non-translated sequence groups. The oscillations present in the mean minimum distance plot suggest that the leader sequence of the mRNA contains synchronization information used by the ribosome to lock on to the correct reading frame. The results of our work suggest that it may be possible to design a coding theory based model that can distinguish between protein coding and non-protein coding genomic sequences by "decoding" the leader region of the mRNA, and probably more.

Acknowledgments

This research is supported in part by an NSF Graduate Fellowship Grant.

References

[1] Peter Sweeney, *Error Control Coding an Introduction*, Prentice Hall, New York, NY, 1991.
[2] Benjamin Lewin, *Genes V*, Oxford University Press, New York, NY, 1995.
[3] Elebeoba E. May, "Comparative Analysis of Information Based Models for Initiating Protein Translation in Escherichia coli K-12", Master's thesis, NCSU, December 1998.