Random Vector Model for Translation Initiation in E. coli K-12

E. E. May*, M. A. Vouk, D. L. Bitzer, and D. I. Rosnick
North Carolina State University, Raleigh, NC USA
***email: eueni1@eos.ncsu.edu**

Abstract

The increase in the availability of genomic data has encouraged the development of computational methods for the analysis of this data. In the work described here, we construct a random vector realization for the protein translation initiation system and evaluate it for Escherichia coli K-12.

If we view protein initiation as a system which receives messenger RNA sequences as input and produces an initiation signal as output, we can classify this system using standard engineering tools and constructs. The input/output relationship of the protein initiation system can not be accurately described using deterministic methods or theories. Complexity and numerous factors that influence the behavior of the protein initiation system, are better served if we can classify the system as a random system. We define a random experiment [1] that captures the binding pattern between the leader regions of the E.coli mRNA sequence and the last thirteen bases of the 3' end of the 16s ribosomal RNA; this region of the rRNA contains the hexamer complementary to the Shine-Dalgarno sequence [2]. Two random variables, $Max1$ and $Max2$ map the outcomes of our random experiment to numeric values. $Max1$ is the maximum number of consecutive hydrogen bonds formed between a thirteen-base mRNA subsequence and the thirteen base subsequence of the 16s rRNA. $Max2$ is the next highest number of consecutive hydrogen bonds formed in this region. These two random variables capture biological characteristics which influence the initiation of protein translation. We design a random vector model to capture the positional interdependence that exists within the mRNA signal[3]. The random vectors V_{Max1} and V_{Max2} are formed using $Max1$ and $Max2$, respectively.

The Escherichia coli K-12 strain MG1655 sequence data (downloaded from the NIH ftp site: ncbi.nlm.nih.gov) is used to test the random vector model. The E.coli genomic sequence is divided into three groups: translated, hypothetically translated, and non-translated. A random vector model is constructed for each sequence grouping and the mean of the $Max1$ and $Max2$ values for each element in the random vector is plotted by position. Position zero signifies the first base of the initiation codon. Negative position values represent upstream regions and positive position values represent regions downstream from the first base of the initiation codon. Our random vector model indicates a significant difference in the behavior of the the three sequence groups in the region between -20 and -10. There are also strong oscillations, with a periodicity of three, present in the -10 to zero region; this is the region which contains the Shine-Dalgarno sequence [2]. These oscillations continue into the downstream region of the mRNA, but the magnitude of the downstream oscillations are less than those in the -10 to zero region. The random vector model indicates a window within the mRNA sequence that can be used to distinguish between translated sequences and non-translated sequences. The model also indicates a synchronization region which may be used by the ribosome to lock on to correct reading frames. The results of the random vector model are encouraging and consistent with the biological model. Work on this model continues.

References

[1] Yannis Viniotis. *Probability and Random Processes for Electrical Engineers.* The McGraw-Hill Companies, Inc., New York, N.Y., 1998.

[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.