An Optimized Ribosome Binding Model Using Communication Theory Concepts

Mohammad Al Bataineh, **Maria Alonso**, and **Guillermo Atkin**, Senior Member IEEE, Department of Electrical & Computer Engineering, Illinois Institute of Technology. **Siyun Wang**, and **Wei Zhang**, Department of Biological, Chemical, Physical Sciences, Illinois Institute of Technology and National Center for Food Safety and Technology, Chicago, IL 60616

Abstract - *Regulatory elements in genetic data are often short and variable, their identification and discovery using computational algorithms is difficult. A model based on communication theory concepts and used by Dawy, et al., [1], is used to model the process of translation in gene expression. The model assumes the ribosome decodes the mRNA sequence by using the 3' end of the 16SrRNA molecule as an embedded codebook. In this work we used an optimized algorithm to obtain the Shine-Dalgarno (SD) sequence that allows detecting this sequence in the genes without the averaging used in [1]. The validity of the model is established by the ability of detecting the Shine-Dalgarno in translation, the initiation codon still can be obtained by averaging as in [1]. Mutations are also studied for De Boer case. Results are compared to biological data and proved to be consistent.*

Keywords: Gene expression; Shine-Dalgarno; translation; mRNA; optimization.

1 Introduction

This work deals with modeling gene expression (information contained in the DNA molecule when transformed into proteins). The accuracy of this process is vital to the survival of the organisms. Gene expression involves two main stages. The first one is transcription (related to coding theory) where the information stored in the DNA is transformed into the messenger RNA (mRNA) by the RNA polymerase molecule. The resulting mRNA corresponds to a complementary copy of the template strand except that the base T (Thymine) is substituted by U (Uracil). The second one is translation (related to detection theory), where the mRNA molecule serves as an instructive for protein synthesis. The ribosome molecule translates the mRNA into a sequence of amino acids - a protein. Hereby, triplets of bases are converted to amino acids according to the mapping rule described by the genetic code (see Figures 1 and 2). Analyzing gene expression, many similarities with the way engineers send digital information come into view. Concepts of information theory, communications, detection theory, pattern recognition and source and channel coding can be used to find out analogies between these fields [1][2][3][4][5][6]. At the same time the analysis of the results made possible by developing these models can serve as a way to introduce new lines of biological research. In practice, these results can lead to better recognition of signals in gene expression. The use of communications engineering ideas for understanding genetic information has been promoted by the increased availability of genetic data. In this work we study a communication theory based model for translation in genome expression.

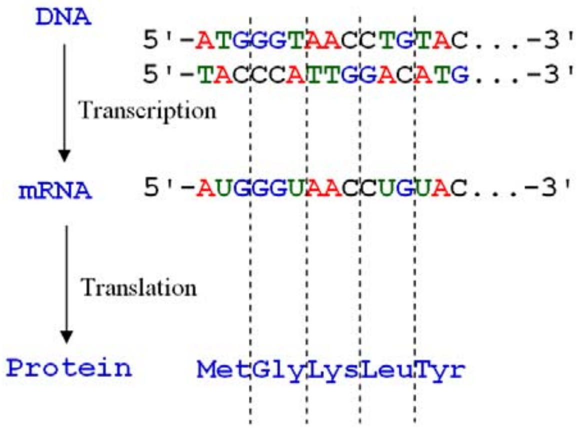


Figure 1. Protein Synthesis

This work has been supported by a grant from the Pritzker Institute, Illinois Institute of Technology

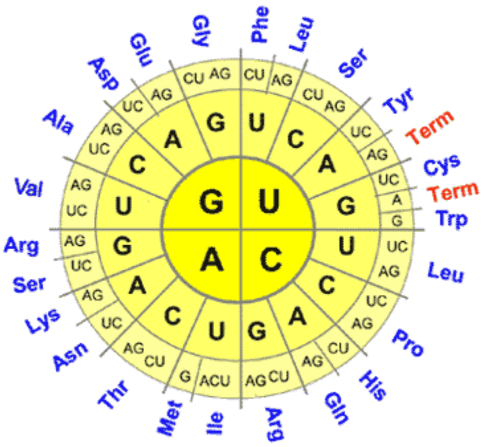
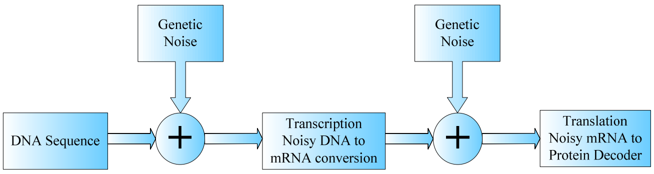


Figure 2. Genetic Code (www.medigenomix.de)

2 Communication Theory – Based Modeling

Our work focuses on modeling translation, specifically in the E. coli bacteria [7][8][9]. During translation the ribosome scans and searches the mRNA [10][11] for a translation initiation signal. Figure 3 shows a general model of gene expression from a communication theory view.



Genetic Noise
Genetic Noise
DNA Sequence
+
Transcription Noisy DNA to mRNA conversion
+
Translation Noisy mRNA to Protein Decoder

Figure 3. Gene Expression as a Communication Model

The ribosome binds in the leader region of the mRNA sequence. The leader region is composed by the bases upstream of the initiation codon. This codon, typically AUG, marks the start of a coding region that is the part of the mRNA that will code for a protein. Figure 4 shows a typical mRNA sequence [1].

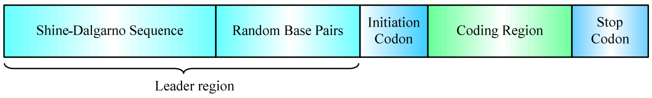


Figure 4. mRNA Sequence

The analysis is based on using as input to the model the noisy mRNA. The last 13 bases of the molecule 16SrRNA (inside the ribosome) interact with the leader region of the mRNA to start translation. A codebook is built of variable length N between 2 and 13. 13-N+1 codewords are generated by taking a sliding window through the Watson-Crick complement of the sequence of 13 the base (shift one base at a time), [1]. This sequence and the resulting codebook are shown in Figure 3 and Table 1, below for N=5:

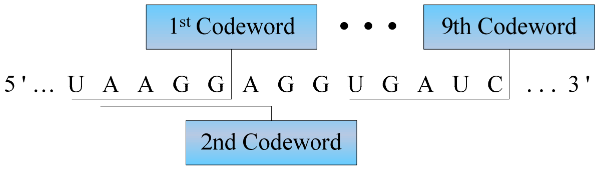


Figure 5. Structure of the Codebook

Codeword #	Codeword
C_1	UAAGG
C_2	AAGGA
C_3	AGGAG
C_4	GGAGG
C_5	GAGGU
C_6	AGGUG
C_7	GGUGA
C_8	GUGAU
C_9	UGAUC

Table 1. Codebook

A sliding window on the received noisy mRNA sequence is used to select subsequences of length N and compare them with all codewords in the codebook. The codeword that results in a minimum weighted free energy distance metric between doublets in kcal/mol is selected as the right codeword (see Table 2). The minimum energies are stored and plotted to show the performance of the algorithm.

Free Energy Doublets			
AA -0.9	AG -2.3	GA -2.3	GG -2.9
AU -0.9	AC -1.8	GU -2.1	GC -3.4
UA -1.1	UG -2.1	CA -1.8	CG -3.4
UU -0.9	UC -1.7	CU -1.7	CC -2.9

Table 2. Energy Table (Kcal/mol)

To obtain the weighted minimum free energy value the following algorithm is used assigning weights to the doublets such that the total energy of the codeword is increased exponentially with a match. At the same time the energy is decreased if a mismatch occurs. The algorithm emphasizes or de-emphasizes the value of the weights when consecutive matches or mismatches occur.

3 Exponentially Weighted Free Energy Decoding Algorithm

The energy function has the following form:

$$E = \sum_{k=1}^{N} w_k \, \delta_k \tag{1}$$

where δ_k denotes a match ($\delta_k = 1$) or a mismatch ($\delta_k = 0$) and wk is the weight applied to the doublet in the kth position. The weights are given by:

$$w_k = \begin{cases} \rho + a^{\sigma} & \text{if } \delta_k = 1 \\ \max\left\{ w_{k-1} - \left(a^{\tilde{\sigma}+1} - a^{\tilde{\sigma}} \right), 0 \right\} & \text{if } \delta_k = 0 \end{cases} \tag{2}$$

where σ and $\tilde{\sigma}$ are the numbers of consecutive matches or mismatches and ρ is an offset variable updated as it follows:

$$\rho = \begin{cases} \rho & \text{if } \delta_k = 1 \\ 0 & \text{if } \delta_k = 0 \;\&\; \rho \le a \\ \max\left\{ w_{k-1} - \left(a^{\tilde{\sigma}+1} - a^{\tilde{\sigma}} \right), 0 \right\} & \text{otherwise} \end{cases} \tag{3}$$

where *a* is a number that will determine the exponential growth of the weighting function. For larger values of *a* exponential will grow faster making the algorithm more sensible to the correlation in the sequence.

This algorithm allows determining the exact position of the Shine-Dalgarno signal on each gene rather than using an average.

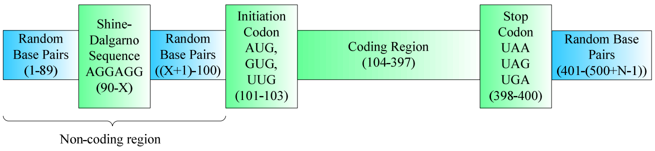
For the next figures the procedure used is as follows:

1. Obtain the complete genome of the prokaryotic bacteria E. coli strain MG1655
2. Identify coding and noncoding regions using the start and stop positions in the mRNA sequence
3. Implement the weighted Free Energy Ribosome Decoding algorithm that allows identifying the SD sequence in the genes without the need of averaging.

4 Analysis and Simulation Results

In order to test our model, sequences of the complete genome of the prokaryotic bacteria E. coli strain MG1655 were obtained. These sequences are available in the National Center for Biotechnology Information. Figure 6 shows the Shine-Dalgarno detection algorithm for different values of *a* and the results of the algorithm used in [1]. It is seen that using the algorithm and by a proper choice of the parameter *a*, which determines the exponential growth of the weighting function, we can not only detect the Shine Dalgarno signal, but also with a much better resolution compared with the results obtained in [1].

The next figures use the following structure for presentation purposes:


Random Base Pairs (1-89)
Shine-Dalgarno Sequence AGGAGG (90-X)
Random Base Pairs ((X+1)-100)
Initiation Codon AUG, GUG, UUG (101-103)
Coding Region (104-397)
Stop Codon UAA UAG UGA (398-400)
Random Base Pairs (401-(500+N-1))
Non-coding region

Where X represents the position of the last G of the Shine-Dalgarno signal in the above sequence structure (i.e. 90 + SD length). N is the codeword length used to design the codebook.

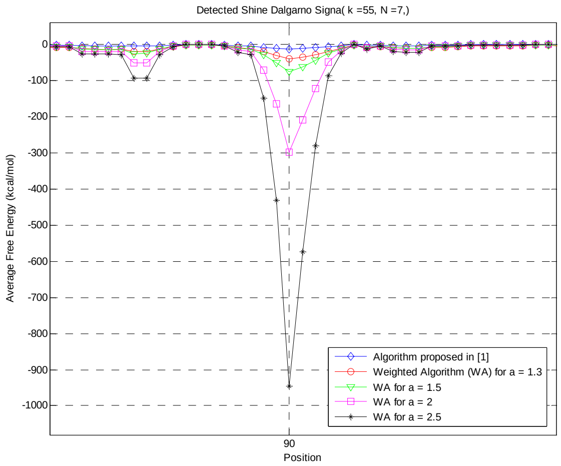


Figure 6. Detected SD signals as a function of the parameter a and using algorithm in [1] (*k* is the gene number)

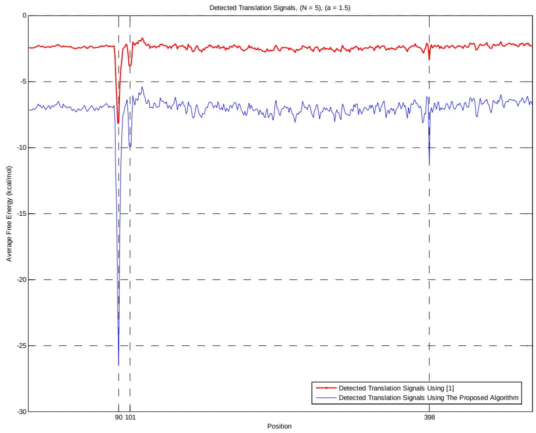


Figure 7. Comparison of SD signal (position 90), start (position 101) and termination (position 398) codon between the algorithm used in [1] and the weighted algorithm

Results from De Boer [12] mutation (3 point mutations in the positions 5 to 7 in the last 13 bases in the 16SrRNA molecule (GAG → UGU)) are shown in Figure 8. Such mutation is considered to be lethal for the organism in the sense that the whole translation process will be stopped, and hence no protein will be synthesized. This mutation was tested using our proposed model. The results show a complete loss of the SD signal. Hence, it can be inferred that the translation will never take place as happens in real life laboratory experimentation. However, the detection of the initiation codon is not affected by the mutation and this agrees with biological results. These results are consistent with the previous models and experimental work.

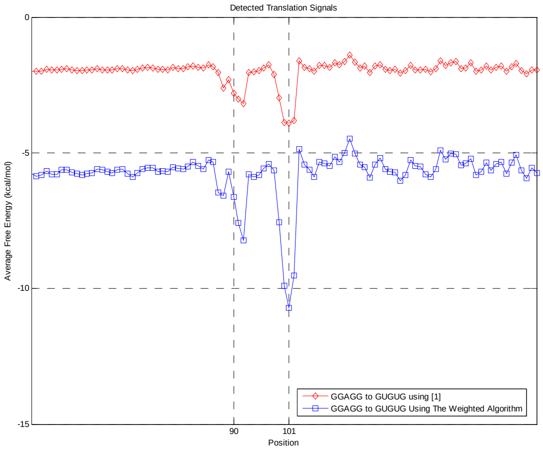


Figure 8. SD detection with De Boer Mutation

5 Conclusions

The new proposed weighted algorithm allows a better resolution of the SD sequence. It lets detecting this sequence in the genes without the need of averaging over a large set of genes. The algorithm is sensitive to the parameter a, and by properly choosing this value we can increase the accuracy of the previous work. By combining this algorithm with the averaging used in [1], we can detect the SD, the initiation and the termination signals, and analyze the effect of mutations in the last 13 bases in the 16SrRNA molecule. The model investigated facilitates testing of mutations in the ribosome molecular structure. Results match previous published experimental work. This shows the relevance of the model, its biological accuracy, and its flexibility to incorporate and study structural changes. Also, the proposed algorithm allows testing various combinations of mutations without the need for time and cost consuming laboratory experimentation.

6 References

[1] Dawy, Z.; Gonzalez, F.; Hagenauer, J.; Mueller, J.C.: Modeling and Analysis of Gene Expression Mechanisms: A Communication Theory Approach. - In: IEEE International Conference on Communications (ICC 2005), Seoul, South Korea, May 2005, vol. 2, S. 815-819

[2] H. Yockey, Information theory and molecular biology. Cambridge: Cambridge University Press, 1992.

[3] T. Schneider, "Theory of molecular machines I. Channel capacity of molecular machines," Journal Theoretical Biology, vol. 148, pp. 83123, 1991.

[4] T. Schneider, "Theory of molecular machines II. Energy dissipation from molecular machines," Journal Theoretical Biology, vol. 148, pp. 125137, 1991.

[5] G. Battail, "An engineer's view on genetic information and biological evolution," Submitted to Elsevier Science, July 2003.

[6] M. Eigen, "The origin of genetic information: viruses as models," Gene, vol. 135, pp. 37-47, 1993.

[7] E. May, Analysis of Coding Theory Bases Models for Initiating Protein Translation in Prokaryotic Organisms. PhD thesis, North Carolina State University, Raleigh, January 2002.

[8] E. May, M. Vouk, D. Bitzer, and D. Rosnick, "Coding model for translation in E.coli K-12," Proceedings of The First Joint BMESIEMBS Conference, October 1999.

[9] M. Kozak, "Initiation of translation in prokaryotes and eukaryotes," Gene, vol. 234, pp. 187-208, 1999.

[10] J. Steitz and K. Jakes, "How ribosomes select initiator regions in mRNA: base pair formation between the 3' terminus of 16S rRNA and the mRNA during initiation of protein synthesis in E. coli," Proc. Nat!. Acad. Sci., vol. 72, pp. 4734-4738, 1975.

[11] J. Shine and L. Dalgarno, "The 3' terminal sequence of Escherichia coli 16S ribosomal RNA: complementary to nonsense triplets and ribosome binding sites," Proc. Nat!. Acad. Sci., vol. 71, pp. 1342-1346, 1974. [11] T. Schneider, "Consensus sequence Zen," Applied Bioinformatics, vol. 3, pp. 111-119, 2002.

[12] A. Hui and H. D. Boer, "Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in Escherichia coli," Proc. Nat!. Acad. Sci., vol. 84, pp. 47624766, 1987.