Ribosome Binding Model Using a Codebook and Exponential Metric

Mohammad Al Bataineh, Maria Alonso, Siyun Wang, Wei Zhang, Guillermo Atkin, Senior Member IEEE

Abstract - **A model based on a variable length codebook and a metric is used to model the process of translation in gene expression. In this model it is assumed the ribosome decodes the mRNA sequence by using the 3'end of the 16SrRNA molecule as an embedded codebook. The metric uses an exponential algorithm to recognize the Shine Dalgarno (SD) sequence that allows detecting this sequence in each gene without the averaging used in [1]. The E.Coli O157:H7 Sakai sequence data is used in this model and the validity of the results is proved by the ability of detecting the Shine Dalgarno in translation. The initiation codon still can be found by averaging using the method used in [1]. Mutations are also studied for Jacob, Hui and De Boer cases. Results are compared to biological data and prove to be consistent.**

Index Terms – **Bioinformatics; Gene expression; Shine Dalgarno; mRNA; optimization**

I. INTRODUCTION

This work deals with modeling gene expression (information contained in the DNA molecule when transformed into proteins). The process of gene expression involves two main stages. The first one is transcription (related to coding theory) where the information stored in the DNA that has been contaminated with genetic noise is transformed into the messenger RNA (mRNA). The second one is translation (related to detection theory), where the noisy mRNA molecule serves as an instructive for protein synthesis. The accuracy of this process is vital to the survival of the organisms.

Mohammad Al Bataineh, Ph. D. Candidate; Maria Alonso, MS student, and Guillermo Atkin, Ph. D., Senior Member IEEE, Associate Professor; are with the Department of Electrical & Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616; 1-312-567-3417; 1-312-567-8976 Fax (e-mail: albamoh@iit.edu; malonso@iit.edu, atkin@iit.edu). Siyun Wang, Ph. D. candidate, and Wei Zhang, Ph. D., Assistant Professor, are with the Department of Biological, Chemical, Physical Sciences, Illinois Institute of Technology and National Center for Food Safety and Technology, Chicago, IL 60616 (e-mail: swang26@iit.edu, zhangw@iit.edu). This work has been supported by a grant from the Pritzker Institutte, Illinois Institute of Technology.

The analysis of gene expression has many similarities with how engineers analyze a communication process. Analogies between information theory, communications, detection theory, pattern recognition and source and channel coding can be used to analyze problems related to transcription and translation [1][2][3][4][5][6]. These models can be used to predict and/or analyze the process of gene expression and allow developing and introducing new lines of biological research. In practice, these results can lead to better recognition of signals and sequences in gene expression. The use of communications engineering ideas for understanding genetic information has been made possible by the increased availability of genetic data. In this work by using a codebook and a metric specially designed the Shine-Dalgarno sequence is detected. The codebook and metric are common elements used in the detection process of communication systems. Therefore an "electric" model for translation in genome expression is obtained.

II. 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. Fig. 1 shows a general model of gene expression from a communication theory point of view.

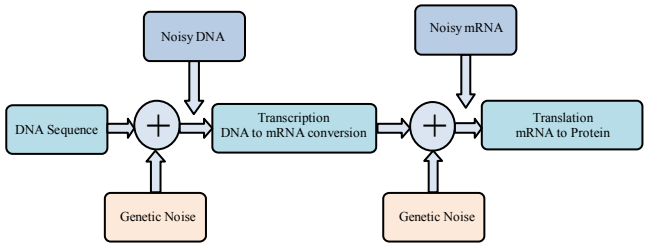


Fig. 1. Transcription and Translation as a Communication Model

Fig. 2 shows a typical mRNA sequence [1]. It is assumed the ribosome binds in the leader region of the mRNA sequence. The leader region is formed by the bases upstream of the initiation codon. These codons, typically AUG, GUG or UUG, are in the start of a coding region that is the part of the mRNA that will translate to a protein.

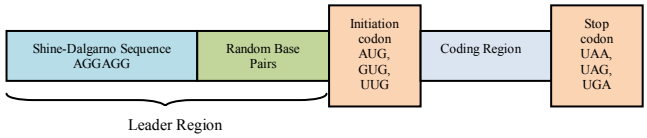


Fig. 2. mRNA Sequence

The input to the model is the noisy mRNA and the last 13 bases of the molecule 16SrRNA (in the ribosome) interact with the leader region of the mRNA to start translation. The mRNA is a noisy version of the mRNA produced in the transcription process due to the addition of genetic noise. The codebook used has variable length N between 2 and 13. Then 13-N+1 codewords are produced by taking a sliding window through the Watson-Crick complement of the sequence of 13 bases (shift one base at a time), [1]. This sequence (UA**AGGAGG**UGAUC) and the resulting codebook for a value N = 5 are shown in Fig. 3 and Table 1 (notice the SD sequence is AGGAGG):

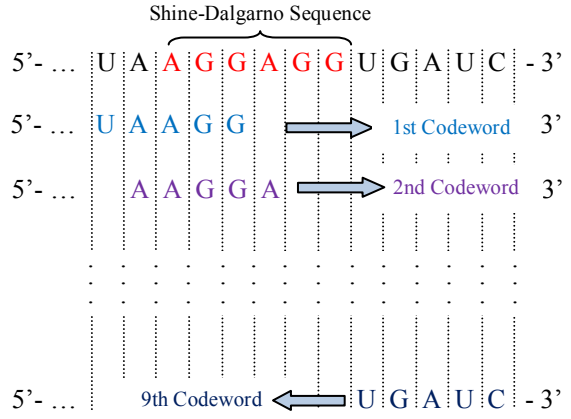


Fig. 3. Codebook Structure Length N = 5

Codebooks of different lengths can be designed following a similar procedure. For example, for N = 7, the first codeword will be UAAGGAG. The second codeword is obtained by selecting the next 7 bases after shifting the selection process by one (i.e., AAGGAGG).

Coderword #	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 length N = 5

A moving window of size N is applied to the received noisy mRNA sequence to select subsequences of length N and match them with the codewords in the codebook. The codeword that results in a minimum weighted free energy exponential metric between doublets (pair of bases) in kcal/mol is selected as the correct codeword (Table 2). The minimum energies are evaluated and plotted to determine the performance of the algorithm.

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

Table 2. Energy Table (Kcal/mol)

To obtain the exponential metric based on the free energy value between pairs of bases the algorithm assigns weights to the doublets such that the total energy of the codeword is increased exponentially with a match and decreased if a mismatch occurs. The algorithm stresses or de-emphasizes the value of the energies when consecutive matches or mismatches occur.

III. Exponentially Weighted Free Energy Decoding Algorithm

The energy function has the following form:

$$E = \sum_{k=1}^{N} w_k \, \delta_k \qquad (1)$$

where δ_k means a match ($\delta_k = 1$) or a mismatch ($\delta_k = 0$) and w_k is the weight applied to the doublet in the k[ith] position. The weights are given by:

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

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

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

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

This algorithm allows deciding the exact position of the Shine-Dalgarno on the genes rather than using an average.

The figures that follow are obtained using the procedure described below:

1. The data used corresponds to the complete genome of the prokaryotic bacteria *E. coli* strain O157:H7
2. The coding and noncoding regions are identified using the start (AUG, GUG, UUG) and the stop (UAA, UAG, UGA) codons in the mRNA sequence
3. Implement the exponentially weighted Free Energy Decoding algorithm that allows identifying the SD sequence without the need of averaging.

IV. Analysis and Simulation Results

In order to test our model, sequences of the complete genome of the prokaryotic bacteria E. coli strain O157:H7 were obtained. These sequences are available in the National Center for Biotechnology Information. Fig. 4 shows the SD detection algorithm for different values of *a* and the results of the algorithm used in [1]. It is noted the weighted algorithm produce a better resolution in detecting the SD sequence.

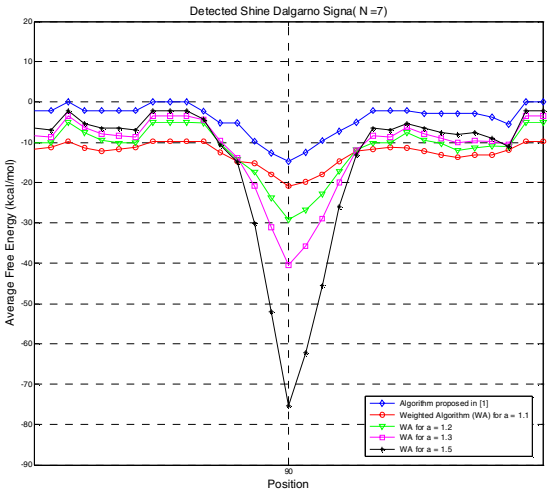
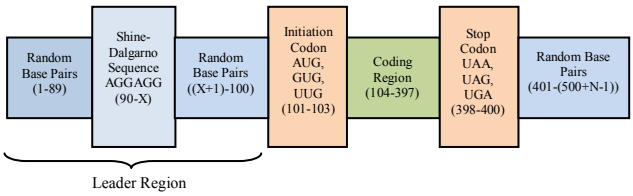


Fig. 4. Detected SD signals as a function of the parameter a and using algorithm in [1]

Also, in Figure 4, there is no averaging of the genes. If the parameter *a* is further increased the resolution of the peak corresponding to the SD sequence will be larger.

The next figures use the following structure for presentation purposes:



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

The results of the proposed weighted algorithm using averaging of the SD, start and stop codon results are shown in Fig. 5. It is noted that the weighted algorithm performs much better than the codebook alone.

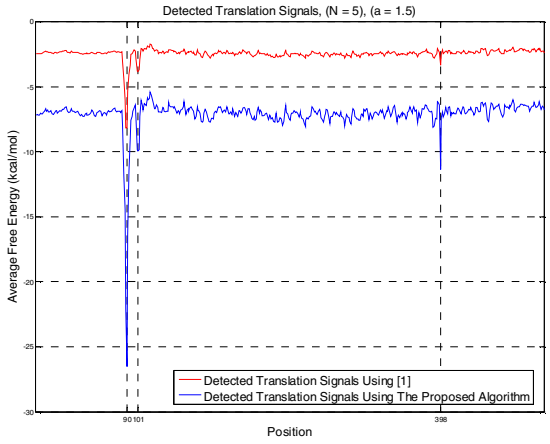


Fig. 5. Comparison of SD signal (position 90), start (position 101) and termination (position 398) codon between the algorithm used in [1] and the weighted algorithm

The published mutations of Hui and De Boer [12] in the 3' end of the 16SrRNA are tested using our model. Biologically speaking, such mutations are lethal for the organism in the sense that the production of proteins stops. Our results prove to be consistent with the previous models and experimental work. They show a complete loss of the SD signal. Therefore, it can be inferred that the translation will never take place. However, detection of the initiation codon is not affected by these mutations and this agrees with biological results. This is illustrated in Fig. 6 which is obtained using the algorithm proposed in [1].

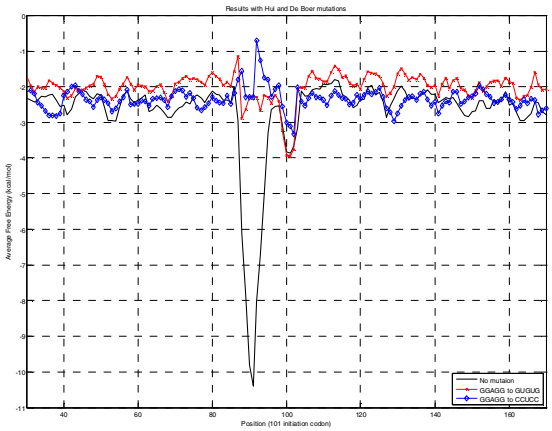


Fig. 6. Results with Hui and De Boer mutations using algorithm in [1]

It is noted in Fig. 6 that the SD peak (position 90) almost complete disappear due to the mutations. The large peak corresponds where no mutations are present.

The same mutations are tested using our model which resulted in a similar result but with a better resolution (note the difference in the y-axis) of the translation signals as illustrated in Fig. 7.

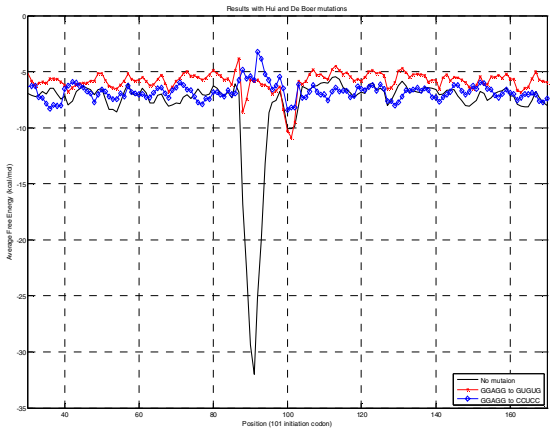


Fig. 7. Results with Hui and De Boer mutations using the our algorithm

Another published record of the behavior of the protein synthesis under mutations in the 3' end of the 16SrRNA, was done by Jacob [13]. This point mutation consisted in a change of the nucleotide C → U in the ribosome small subunit. This is equivalent to make a mutation from G → A in the complement sequence. Specifically, the 5th position in the arrangement illustrated below:

Position	1	2	3	4	5	6	7	8	9	10	11	12	13
Base	U	A	A	G	G	A	G	G	U	G	A	U	C
Mutation	U	A	A	G	A	A	G	G	U	G	A	U	C

This mutation is incorporated in the 16SrRNA based codebook and tested using the proposed weighted algorithm. The result of this mutation in real life is a reduction in the level of protein synthesis. This result can be predicted by noticing the partial loss in the amplitude of the Shine-Dalgarno signal as illustrated in Fig. 8. However, Jacob mutation does not affect the detection of the termination signal. This means that protein synthesis process is normally terminated.

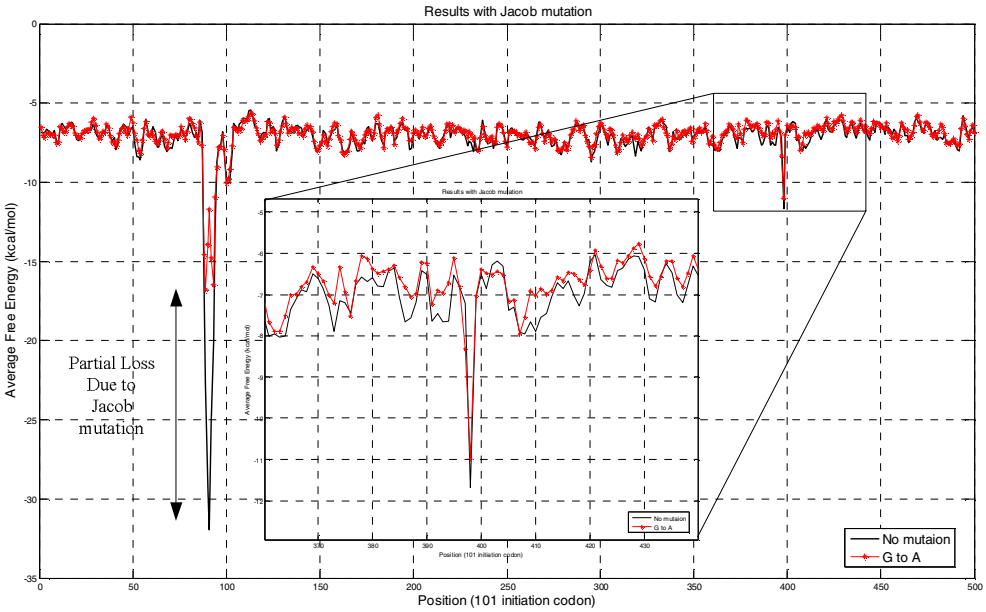


Fig. 8. Results with Jacob mutation

V. 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 them. The algorithm is sensitive to the parameter *a*, and by properly choosing this value the accuracy of the previous work can be improved. By combining this algorithm with the averaging used in [1], we can detect the SD, the initiation and the ending signals and analyze the effect of mutations in the last 13 bases in the 16SrRNA molecule. The results match previous research. The model studied helps testing of mutations in the ribosome molecular structure. Results also 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 reducing the need for time and cost consuming laboratory experimentation.

VI. 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. 125-137, 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. Dalgaruo, "The 3' terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity 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. 4762-4766, 1987.

[13] W. Jacob et al., "A single base change in the Shine Dalgarno region of 16S rRNA of Escherichia coli affects translation of many proteins," Proc. Natl. Acad. Sci., vol. 84, pp. 4757–4761, 1987.