Analysis of Gene Translation Using a Communications Theory Approach

Mohammad Al Bataineh[1], Lun Huang[1], Maria Alonso[1], Nick Menhart[2], and Guillermo E. Atkin[1]

[1]Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL, USA; [2]Department of Biological, Chemical, and Physical Sciences, Illinois Institute of Technology, Chicago, IL, USA

Abstract - Rapid advances in both genomic data acquisition and computational technology have encouraged the development and use of advanced engineering methods in the field of bioinformatics and computational genomics. Processes in molecular biology can be modeled through the use of these methods. Such processes include identification and annotation of all the functional elements in the genome, including genes and regulatory sequences, which is a fundamental challenge in genomics and computational biology. Since regulatory elements are often short and variable, their identification and discovery using computational algorithms is difficult. However, significant advances have been made in the computational methods for modeling and detection of DNA regulatory elements. This paper proposes a novel use of techniques and principles from communications engineering, coding and information theory for modeling, identification and analysis of genomic regulatory elements and biological sequences. The methods proposed are not only able to identify regulatory elements (REs) at their exact locations, but also "interestingly" can distinguish coding from non-coding regions. Therefore, the proposed methods can be utilized to identify genes in the mRNA sequence.

1 Introduction

Communications and information theory has proved to provide powerful tools for the analysis of genomic regulatory elements and biological sequences [1-5]. An up-to-date summary of current research can be found in [6]. The genetic information of an organism is stored in the DNA, which can be seen as a digital signal of the quaternary alphabet of nucleotides $\overline{X} = \{A, C, G, T\}$. An important field of interest is gene expression, the process during which this information stored in the DNA is transformed into proteins. Gene expression codes for the expression of specific proteins that carry out and regulate such processes. Gene expression takes place in two steps: transcription and translation.

This paper is organized as follows. Section 2 describes our previous model for the process of translation in gene expression being compared to the the work done

in [5]. Section 3 presents four new other models for the process of translation with simulation results shown in section 4. Finally, conclusions are drawn in Section 5.

2 Previous Model

The process of translation in prokaryotes is triggered by detecting an RE known as the Shine-Dalgarno (SD) sequence. Physically, this detection works by homology mediated binding of the RE to the last 13 bases of the 16S rRNA in the ribosome [7]. In our work [1] and [2], we have modified this detection/recognition system introduced in [5] by designing a one-dimensional variable-length codebook and a metric. The codebook uses a variable codeword length N between 2 and 13 using the Watson-Crick complement of the last 13 bases of the 16S rRNA molecule. Hence, we obtain (13-N+1) codewords; $\bar{c}_i = [s_1, s_2, \ldots, s_{i+N-1}]$; $i \in [1, 13-N+1]$ where $\bar{s} = [s_1, s_2, \ldots, s_{13}] = [\text{UAAGGAGGUGAUC}]$ stands for the complemented sequence of the last 13 bases. A sliding window of size N applies to the received noisy mRNA sequence to select subsequences of length N and match them with the codewords in the codebook (see Table 1). The codeword that results in a minimum weighted free energy exponential metric between doublets (pair of bases) is selected as the correct codeword and the metric value is saved. Biologically, the ribosome achieves this by means of the complementary principle. The energies involved in the rRNA-mRNA interaction tell the ribosome when a signal is detected and, thus, when the start of the process of translation should take place. In our model, a modified version of the method of free energy doublets presented in [7] is adopted to calculate the energy function (see equation 1). This function represents a free energy distance metric in kcal/mol instead of minimum distance (see Tables 2) [5]. Our algorithm assigns weights to the doublets such that the total energy of the codeword increases with a match and decreases with a mismatch. Therefore, the total energy gets more emphasized or de-emphasized when consecutive matches or mismatches occur. The energy function has the following form:

$$E_k = \sum_{n=1}^{N} w_n E_n \delta_n \tag{1}$$

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^{th} position. The weights are given by:

$$w_n = \begin{cases} \rho_n + a^{\sigma_n} & \text{if match} \\ \max\left\{ w_{n-1} - \left(a^{\tilde{\sigma}_n+1} - a^{\tilde{\sigma}_n}\right), 0 \right\} & \text{if mismatch} \end{cases} \tag{2}$$

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

$$\rho_n = \begin{cases} \rho_{n-1} & \text{if } \delta_n = 1 \\ 0 & \text{if } \delta_n = 0 \;\&\; \rho_{n-1} \le a \\ \max\left\{ w_{n-1} - \left(a^{\tilde{\sigma}_n+1} - a^{\tilde{\sigma}_n}\right), 0 \right\} & \text{otherwise} \end{cases} \tag{3}$$

where a is a constant that will control the exponential growth of the weighting function. The offset variable ρ updated at each step according to equation (3), is introduced for the purpose of keeping track of the growing trend that happens when consecutive number of matches occurs followed by a mismatch. When a mismatch occurs we increment the number of mismatches that is initialized to zero by one, reset the number of matches back to zero, calculate the current weighting factor, and finally reevaluate the offset variable to be used in the next alignment. Without the use of this offset variable, we will have several peaks when we came into a good match of the codeword in that particular alignment.

For larger values of a, the exponential will grow faster as the number of consecutive matches increases (hence increasing the likelihood that the right sequence is enhanced) making the algorithm more sensible to the correlation in the sequence. Not only does this algorithm allow controlling the resolution of detection (by the choice of the parameter a) but also allows identification of the exact position of the best match of the Shine-Dalgarno signal in the genes under study.

Table 1. 16SrRNA Codebook

Cl	Codeword
C1	UAAGG
C2	AAGGA
C3	AGGAG
C4	GGAGG
C6	AGGUG
C7	GGUGA
C8	GUGAU
C9	UGAUC

Table 2. Energy Doublets [7]

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

For the analysis, sequences of the complete genome of the prokaryotic bacteria *E. coli* strain MG1655 and O157:H7 strains were obtained from the National Center for Biotechnology Information. Our proposed exponentially weighting algorithm was not only able to detect the translational signals (Shine-Dalgarno, start codon, and stop codon) but also resulted in a much better resolution than the results obtained when using the codebook alone (without weighting). Fig. 1 shows average results for the detection of the SD, start and stop codons being compared to previous work [5]. It can be seen that the proposed algorithm is able to identify the Shine-Dalgarno (peak at position 90) and the start codon (peak at position 101) and the stop codon (peak at position 398). Moreover, these results support the arguments for the importance of the 16S rRNA structure in the translation process. Different mutations were tested using our algorithm and the results obtained further certified the correctness and the biological relevance of the model.

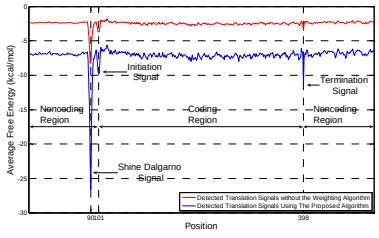

Fig. 1. Detection of translation signals

3 New Models

The previous model discussed in the introduction is based on coding theory (codebook). We have developed different models for the detection process that the ribosome uses to identify and locate translation signals (Shine-Dalgarno, initiation signal, and termination signal) [3]. These models are based on concepts in communications theory as Euclidean distance (model I), matched filter (model II), free energy doublets (model III), and correlation based exponential metric (model IV). The four models are briefly described below.

Model I. Euclidean Distance Based Algorithm

In this model, a Euclidean distance measure can be used to detect a given binding sequence in the mRNA sequence. This measure is calculated at each single base in the mRNA sequence as described in [3]. This method is able to detect the binding sequences in their exact location and accounts for mismatches as well.

Model II. Cross Correlation (Matched Filter)

This model is based on using a matched filter of an impulse response equal to $h(n) = y(-n)$ and an input of $x(n)$ where $y(n)$ is the binding sequence and $x(n)$ is the mRNA sequence [3].

Model III. Free Energy Metric

In this method we use the free energy table (see Table II) to calculate a free energy distance metric in kcal/mol. This metric is calculated at each alignment between the mRNA sequence and the binding sequence under study as described in [3].

Model IV. Exponential Detection Metric

This method detects a binding sequence based on aligning it with the mRNA sequence. An exponential metric related to the total number of matches at each alignment is evaluated as follows:

1. Slide the binding sequence under study along the mRNA sequence one base at a time.
2. At the i^{th} alignment, calculate an exponential weighting function ($W(i)$) using the equation:

$$W(i) = \sum_{n=1}^{N} w(n), \tag{4}$$

where $w(n)$ is the weight applied to the base in the n^{th} position and N is the length of the binding sequence under study. The weights are given by:

$$w(n) = \begin{cases} a^{\sigma} & , \quad if \ match \\ 0 & , if \ mismatch \end{cases}, \tag{5}$$

where a is an input parameter that controls the exponential growth of the weighting function W, and σ is the number of matches at each alignment. .

3. Repeat step 2 for all alignments along the mRNA sequence to get the weighting vector $\overline{W}$:

$$\overline{W} = [w(1), w(2), ..., w(L-N+1)], \quad (6)$$

where L is the length of the mRNA sequence.

4. Plot the weighting vector $\overline{W}$, and detect peaks.

4 Simulation Results

In this section, we show results of applying the four models described briefly in section 3 and we demonstrate their usefulness in pointing out interesting and new biological insights related to the process of translation in gene expression. Without loss of generality and since all of the four models showed similar behavior in detecting translational signals, we chose to show the results of using the Exponential Detection Metric (Model IV) as an example.

In our analysis, sequences of the complete genome of the prokaryotic bacteria E. coli strain MG1655 (also we used O157:H7 strain with similar results) were obtained. These sequences are available in the National Center for Biotechnology Information (NCBI) [8]. For presentation purposes and because of the fact that genes are of different lengths, all the tested sequences were selected to follow a certain structure such that they are all of 500 bases long. The Shine-Dalgarno was set at position 90, the initiation codon at position 101, and the termination codon at position 398. The four new models were used as to detect the last 13 bases of the 16S rRNA molecule in the given mRNA sequence by averaging over all the 500-bases-long test sequences. Simulation results show that the proposed models allow detecting the translational signals at their exact corresponding locations as expected. Furthermore, they allow identifying coding regions (higher ripple region) and the non-coding regions (lower ripple region) as can be observed in figures 2-5. This new result suggests that the last 13 bases sequence of 16S rRNA molecule has a higher correlation with coding regions as compared with non-coding regions. This suggests that the proposed models, which were originally designed for regulatory sequence identification, can help identify genes as well. The four models will be applied to other organisms.

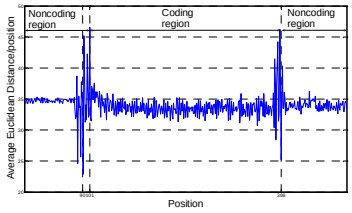

Fig. 2. Euclidean Distance Metric

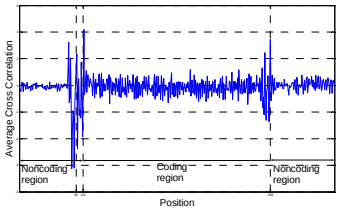

Fig. 3. Cross Correlation (Matched Filter)

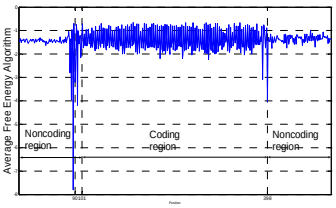

Fig. 4. Free Energy Metric

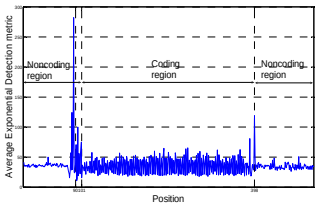

Fig. 5. Exponential Detection

To study the effect of mutations on the detection of translational signals, different types of mutations were incorporated in the last 13-bases sequence and then tested using the developed models. In this work, we have considered Jacob mutation, Hui and De Boer mutations.

Jacob mutation, a mutation in the 5th position of the last 13 bases of 16S rRNA molecule [9], results in a reduction in the level of protein synthesis. This mutation was tested using Model IV. Simulation results in Fig. 6 show a reduction in the amplitude of the Shine-Dalgarno signal compared to the non-mutation case in Fig. 5. This reduction can be interpreted as a reduction in the level of protein synthesis, i.e. the levels of protein production will be reduced but not completely stopped.

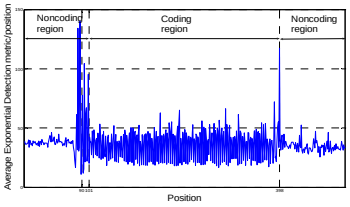

Fig. 6. Jacob Mutation

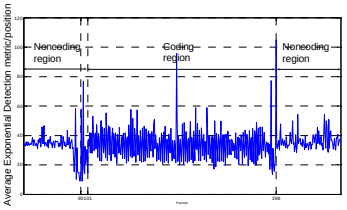

Fig. 7. Hui Mutation

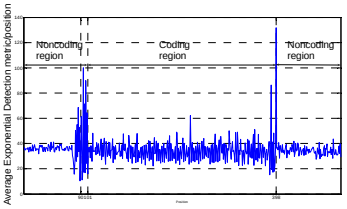

Fig. 8. De Boer Mutation

Hui and De Boer mutations occur in positions 4 to 8 (GGAGG → CCUCC) and positions 5 to 7 (GAG → UGU) of the last 13 bases sequence [10]. The results of both mutations are lethal for the organism in the sense that the production of proteins stop. Fig. 7 shows a complete loss of the Shine-Dalgarno (SD) signal (at position 90). Hence, it can be inferred that the translation will never take place.

5 Conclusion

The increase in genetic data during the last years has prompted the efforts to use advanced techniques for their interpretation. This paper proposes a novel ap-

plication of ideas and techniques from communications and coding theory to model and analyze gene expression and gene and regulatory sequence identification. Different models for regulatory elements identification are developed and investigated. Simulation results verify the correctness, accuracy and biological relevance of these models in detecting regulatory sequences. Moreover, as these models are surprisingly capable of distinguishing coding from noncoding regions, they can help identify genes. Mutations in the 3' end of the 16S rRNA molecule were investigated. The obtained results totally agree with biological experimentations. This further supports the correctness and the biological relevance of the proposed models and hence can serve as a way to introduce new lines of biological research.

References

[1] Mohammad Al Bataineh, Maria Alonso, Siyun Wang, Wei Zhang, and G. E. Atkin., "Ribosome Binding Model Using a Codebook and Exponential Metric," *2007 IEEE International Conference on Electro/Information Technology,* pp. 438-442, 17-20 May 2007.

[2] Mohammad Al Bataineh, Maria Alonso, Siyun Wang, G. Atkin, and W. Zhang, "An Optimized Ribosome Binding Model Using Communication Theory Concepts," *Proceedings of 2007 International Conference for Bioinformatics and Computational Biology,* June 25 – 27, 2007.

[3] Mohammad Al Bataineh, Lun Huang, Ismaeel Muhamed, Nick Menhart, and G. E. Atkin, "Gene Expression Analysis using Communications, Coding and Information Theory Based Models," *BIOCOMP'09 - The 2009 International Conference on Bioinformatics & Computational Biology,* pp. 181-185, July 13-16, 2009.

[4] E. E. May, M. A. Vouk, D. L. Bitzer, and D. I. Rosnick, "Coding theory based models for protein translation initiation in prokaryotic organisms," *Biosystems,* vol. 76, pp. 249-60, Aug-Oct 2004.

[5] Z. Dawy, F. Gonzalez, J. Hagenauer, and J. C. Mueller, "Modeling and analysis of gene expression mechanisms: a communication theory approach," *IEEE International Conference on Communications (ICC),* vol. 2, pp. 815- 819, 2005.

[6] "DNA as Digital Data - Communication Theory and Molecular Biology," *IEEE Engineering in Medicine and Biology,* vol. 25, January/February 2006.

[7] D. Rosnick, "Free Energy Periodicity and Memory Model for Genetic Coding." vol. PhD thesis Raleigh: North Carolina State Univesity, 2001.

[8] "NCBI: National Center for Biotechnology Information. ," from http://www.ncbi.nlm.nih.gov/.

[9] W. F. Jacob, M. Santer, and A. E. Dahlberg, "A single base change in the Shine-Dalgarno region of 16S rRNA of Escherichia coli affects translation of many proteins," *Proc Natl Acad Sci U S A,* vol. 84, pp. 4757-61, Jul 1987.

[10] A. Hui and H. A. de Boer, "Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in Escherichia coli," *Proc Natl Acad Sci U S A,* vol. 84, pp. 4762-6, Jul 1987.