# Gene Expression Analysis using Communications, Coding and Information Theory Based Models

**M. Al Bataineh[1], L. Huang[1], I. Muhamed[2], N. Menhart[2], and G. 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 -** *Identification and annotation of all the functional elements in the genome, including genes and regulatory sequences, 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 no coding regions. Therefore, the proposed methods can be utilized to identify genes in the mRNA sequence.*

**Keywords:** Gene expression analysis, translation, ribosome, coding theory, communications theory.

## 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 (Figure 1).

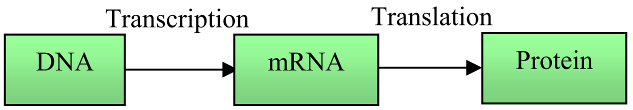

**Figure 1**: The process of protein synthesis (gene expression)

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

## 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 [8]. In our work [1] and [2], we have modified this detection/recognition system done in [4] 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; $\bar{c}_i$ = [$s_1$, $s_2$, …, $s_{i+N-1}$]; i ∈ [1, 13-N+1] where $\bar{s}$ = [$s_1$, $s_2$, …, $s_{13}$] = [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) [4]. 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:

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

where $\delta_k$ 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:

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

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

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

where *a* is a constant that will control the exponential growth of the weighting function. The offset variable $\rho$ 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.

**Table 1.** 16SrRNA Codebook

| Cl | Codeword | C5 | GAGGU |
|---|---|---|---|
| C1 | UAAGG | C6 | AGGUG |
| C2 | AAGGA | C7 | GGUGA |
| C3 | AGGAG | C8 | GUGAU |
| C4 | GGAGG | 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 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.

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). Figure 2 shows average results for the detection of the SD, start and stop codons being compared to previous work [4]. 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.

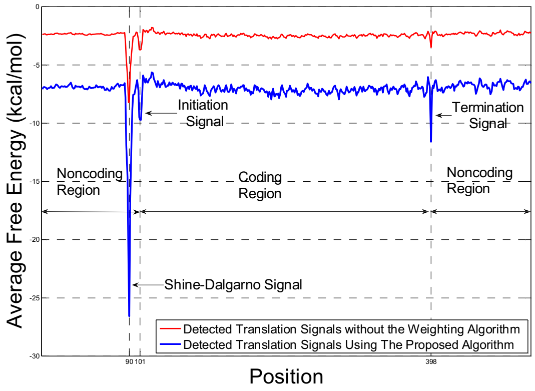

**Figure 2:** Detection of translation signals

## 3 New Models

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

### Model I. Euclidean Distance Based Algorithm

In this method, 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 follows:

1) Map both mRNA sequence and the binding sequence under study to their equivalent numerical quaternary representations using (A = 0, C = 1, G = 2, and U = 3).

2) Slide the binding sequence along the mRNA sequence and find the Euclidean distance at each alignment position.
3) Sum the resulting Euclidean distance vector and save the result as a function of base position.
4) Plot the resulting vector in step 3 and detect minimal points.

A minimal point (dip) of amplitude of zero in the resulting plot corresponds to a total match of the binding sequence. The next minimal point is a partial match of the binding sequence. Therefore, 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)

In telecommunication, a matched filter is obtained by correlating a known signal, or template, with an unknown signal to detect the presence of the template in the unknown signal. This is equivalent to convolving the unknown signal with a time-reversed version of the template. The matched filter is the optimal linear filter for maximizing the signal to noise ratio (SNR) in the presence of additive stochastic noise. Model II is based on using a matched filter of an impulse response equal to h(n)=y(-n) and an input of *x(n)* (see Figure 3) where *y(n)* is the binding sequence and *x(n)* is the mRNA sequence.

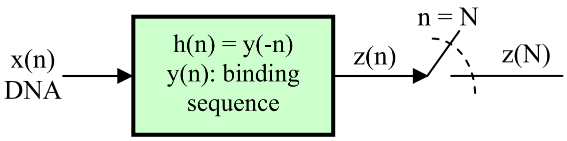

**Figure 3:** Matched Filter

1) Map both the mRNA sequence *x(n)*, and the binding sequence *y(n)*, under study to their equivalent binary representation using (A = 00, C = 01, G = 10, and T = 11).
2) Convert each zero in the resulting binary sequences to (-1) to get a better correlation form.
3) Correlate both sequences using

$$\begin{aligned} \mathrm{z(n)} &= \mathrm{x(n)} \otimes \mathrm{y(n)} = \mathrm{x(n)} * \mathrm{y}^{*}(\mathrm{-n}) \\ &= \sum_{\mathrm{n}=-\infty}^{\infty} \mathrm{x(k)y(n+k)} \end{aligned} \qquad (4)$$

where ( $\otimes$ ) corresponds to cross correlation and ( $*$ ) corresponds to convolution. Correlation is equivalent to convolution of the sequence *x(n)* with an inverted version of the sequence *y(n)*. This can be done by first flipping the sequence *y(n)* and then convolving it with the sequence *x(n)*.
4) Plot the cross correlation function and detect the maximal points.
5) Convert the binding sequence detected positions ( a maximal point in the plot) to their corresponding locations in the original mRNA sequence using:

$$\mathrm{DP}_{\mathrm{mRNA}} = \left\lceil \left(\mathrm{DP}_{\mathrm{plot}} \ \text{-2 BSL} + 1\right) \big/ 2 \right\rceil \qquad (5)$$

where $\mathrm{DP}_{\mathrm{mRNA}}$ is the **d**etected **p**osition in the mRNA sequence, $\mathrm{DP}_{\mathrm{plot}}$ is the **d**etected **p**osition in the plot; BSL is **b**inding **s**equence **l**ength, and $\lceil X \rceil$ rounds the value X to the nearest integer larger than X.

## Model III. Exponential Detection Metric

In this model, a binding sequence is detected by 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 $\mathrm{i^{th}}$ alignment, calculate an exponential weighting function ( $W(i)$ ) using the equation:

$$W(i) = \sum_{n=1}^{N} w(n), \qquad (6)$$

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

$$w(n) = \begin{cases} a^{\sigma} & , \quad if \ \ match \\ 0 & , if \ mismatch \end{cases}, \qquad (7)$$

where $a$ is an input parameter that controls the exponential growth of the weighting function *W*, and $\sigma$ is the number of matches at each alignment. .
3) Repeat step 2 for all alignments along the mRNA sequence to get the weighting vector $\overline{W}$ :

$$\overline{W} = [w(1), w(2), ..., w(L \quad N+1)], \qquad (8)$$

where L is the length of the mRNA sequence.
4) Plot the weighting vector $\overline{W}$ , and detect peaks.

This model considers the total number of matches at each alignment rather than the consecutive number of matches and mismatches as in the codebook model discussed in 3.1.1.

## Model IV. Free Energy Metric

In this model, we use the free energy table (see Table III) 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 follows:

1) Align the binding sequence with the mRNA sequence and shift it to the right one base at a time.
2) At the i$^{th}$ alignment, calculate the free energy metric using the equation:

$$\mathrm{E}(i) = \sum_{n=1}^{N-1} E\left(y_n y_{n+1}\right) . \, \delta(n) \qquad (9)$$

where $N$ is the length of the binding sequence. $\overline{y}$ denotes the binding sequence vector and is given by $\overline{y} = [y_1, y_2, \ldots, y_N]$. Let $\overline{x}$ denote the mRNA sequence vector where $\overline{x} = [x_1, x_2, \ldots, x_L]$.

$E(y_n y_{n+1})$ is the energy dissipated on binding with the nucleotide doublets $y_n y_{n+1}$ and is calculated from Table III. $\delta(n)$ is given by:

$$\delta(n) = \begin{cases} 1 \ , if \ y_n y_{n+1} = x_n x_{n+1} \ (match) \\ 0 \ , if \ y_n y_{n+1} \neq x_n x_{n+1} \ (mismatch) \end{cases} \qquad (10)$$

3) Repeat step 2 for *i=1,2,…,L-N+1*, where *L* is the length of the mRNA sequence vector,
4) Plot the free energy vector E and detect minimal points.

To show how the four previous models behave, we arbitrarily selected a 71-bases-long mRNA sequence as a test sequence. Then, we chose an 11-bases-long sequence starting at position 13 to be a hypothetical binding sequence. This binding sequence was also inserted at position 53 with two bases being changed to get a partial match of the original sequence. The four previous models were applied to detect these binding sequences. Figures 4, 5, 6, and 7 show that these methods are accurately detecting the binding sequence as expected. A total match occurs at position 13 (longer peak/dip), and a partial match occurs at position 53 (shorter peak/ dip).

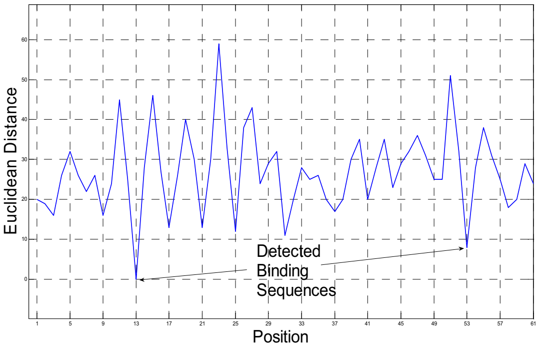

**Figure 4**: Model I: Euclidean Distance Metric

Figure 5 shows that the binding sequence has been detected at positions 46 and 126. According to equation 5, these positions correspond to positions 13 ($\lceil (46 - 22 + 1)/2 \rceil$=13) and 53 ($\lceil (126 - 22 + 1)/2 \rceil$=53) in the original mRNA sequence, respectively.

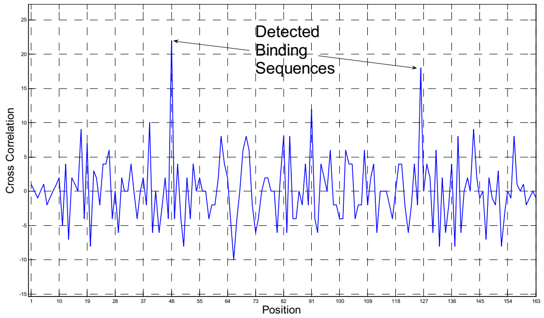

**Figure 5**: Model II: Cross Correlation (Matched Filter)

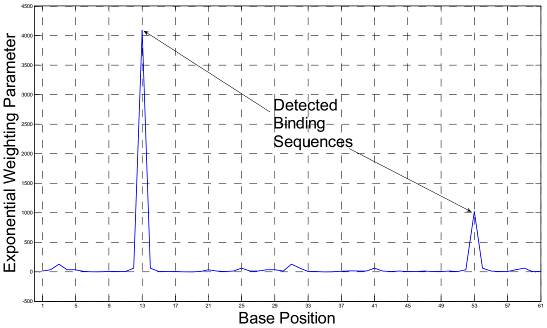

**Figure 6**: Model III: Exponential Detection

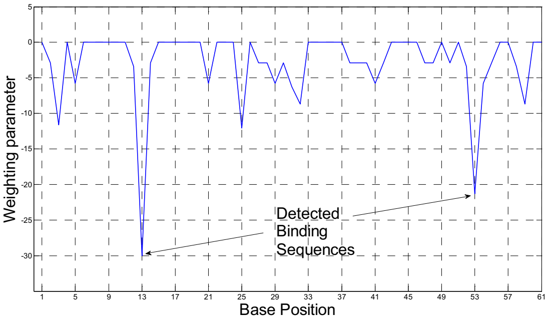

**Figure 7**: Model IV: Free Energy Metric

## 4 Simulation Results

For the analysis, sequences of the complete genome of the prokaryotic bacteria *E. coli* strain MG1655 and o157:H7 were used. Applying these methods to detect the last 13 bases of the 16S rRNA molecule in the given mRNA sequence allows detecting the translational signals at their exact corresponding locations (as previous methods mentioned before, Figure 2). Moreover, these methods "interestingly" allow for distinguishing coding from noncoding regions. This new finding suggests that the last 13 bases of 16S rRNA molecule has a higher correlation with coding regions. Simulation results for the four methods are shown in Figure 8, 9, 10, and 11. From these figures, coding and noncoding regions can be apparently identified. This interesting result will be further analyzed and explored using other binding sequences in future work. Figures 8-11 show that all the proposed methods are obviously able to identify coding from noncoding regions. These interesting findings suggest that the proposed methods, which were originally designed for regulatory sequence identification, can be utilized to identify genes (coding regions) in the mRNA sequence.

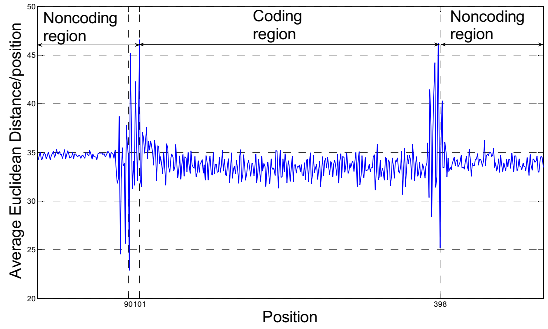

**Figure 8**: Model I: Euclidean Distance Metric

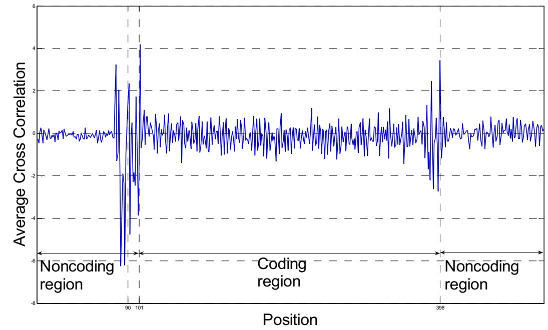

**Figure 9**: Model II: Cross Correlation (Matched Filter)

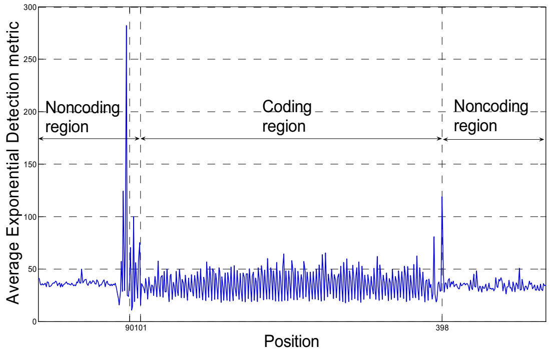

**Figure 10**: Model III: Exponential Detection

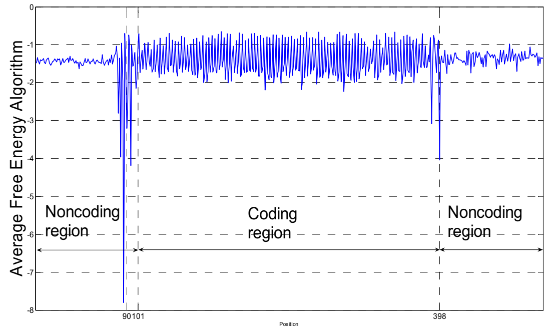

**Figure 11**: Model IV: Free Energy Metric

# 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 application of ideas and techniques from information theory, communications, and coding theory to model and analyze gene expression and gene and regulatory sequence identification. Four new different methods for regulatory elements identification are developed and investigated. Simulation results certify the correctness and accuracy of these methods in detecting regulatory sequences. Moreover, as these methods are "interestingly" capable of distinguishing coding from noncoding regions, they can be utilized to identify genes in the given mRNA sequence.

# 6 References

[1] Mohammad Al Bataineh, Maria Alonso, Siyun Wang, Guillermo Atkin and Wei Zhang, "Ribosome Binding Model Using a Codebook and Exponential Metric," IEEE EIT 2007 Proceedings, Chicago, IL, USA, May 17 – 20, 2007.

[2] Mohammad Al Bataineh, Maria Alonso, Siyun Wang, Guillermo Atkin and Wei Zhang "An Optimized Ribosome Binding Model Using Communication Theory Concepts,"; In: Proceedings of 2007 International Conference for Bioinformatics and Computational Biology, Las Vegas, June 25 – 27, 2007.

[3] E. E. May, M. A. Vouk, D. L. Blitzer, and D. I. Rosnick, "Coding theory based models for protein translation initiation in prokaryotic organisms," BioSystems, vol. 76, pp. 249–260, August-October 2004.

[4] Z. Dawy, F. Gonzalez, J. Hagenauer, and J. C. Mueller, "Modeling and analysis of gene expression mechanisms: a communication theory approach," proceedings of the IEEE International Conference on Communications (ICC), May 2005.

[5] Z. Dawy, B. Goebel, J. Hagenauer, et al., "Gene mapping and marker clustering using Shannon's mutual information," IEEE Transactions on Computational Biology and Bioinformatics, vol. 3, no. 1, pp. 47–56, January-March 2006.

[6] "DNA as Digital Data - Communication Theory and Molecular Biology," IEEE Engineering in Medicine and Biology, vol. 25, no. 1, January/February 2006.

[7] E. May, M. Vouk, D. Bitzer, and D. Rosnick. An errorcorrecting code framework for genetic sequence analysis. Journal of the Franklin Institute, 34:89–109, January-March 2004.

[8] 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. Natl. Acad. Sci., vol. 84, pp. 4762– 4766, 1987.

[9] S. Lin and D. J. Costello, Jr., Error Control Coding. 2nd Edition: Prentice-Hall, 2004.

[10] J. G. Proakis, Digital Communications, 5th ed. New York: McGraw- Hill, 2007.