31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Effect of Mutations on the Detection of Translational Signals Based on a Communications Theory Approach

Mohammad Al Bataineh, Maria Alonso, Lun Huang, G. E. Atkin, Nick Menhart

Abstract—**Gene and regulatory sequence identification is the first step in the functional annotation of any genome. Identification and annotation of such elements in the genome 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 and coding theory for modeling, identification and analysis of genomic regulatory elements and biological sequences. The last 13 bases sequence in the 16S rRNA molecule was used as a test sequence and was detected using the proposed models. Results show that the proposed models are not only able to identify this regulatory element (RE) in the mRNA sequence, but also can help identify coding from noncoding regions. The models described in this work where used to study the effect of mutations in the last 13 bases sequence of the 16S rRNA molecule. The obtained results showed total agreement with published investigations on mutations which further certify the biological relevance of the proposed models.**

I. Introduction

Communications and information theory have proved to provide powerful tools for the analysis of genomic regulatory elements and biological sequences [1]-[6]. An up-to-date summary of current research can be found in [7]. 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 cell functions like oxygen transport. Gene expression codes for the expression of specific proteins that carry out and regulate such processes. Protein gene expression takes place in two steps: transcription and translation.

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 previous work [1], [2] we have improved Dawy's model presented in [5] for this detection/recognition system by designing a one-dimensional variable-length codebook (see Table I) and an exponentially weighted metric. The codebook uses a variable codeword length N between 2 and 13 using the Watson-Crick complement of the last 13 bases sequence of the 3' end of the 16S rRNA molecule. Hence, we obtain (13-N+1) codewords; $\overline{c}_i = [s_1, s_2 \ldots s_{i+N-1}]$; $i \in [1, 13\text{-}N+1]$ where $\overline{s} = [s_1, s_2 \ldots s_{13}]$ = [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. 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 [8] is adopted to calculate the energy function. This function represents a free energy distance metric in kcal/mol instead of minimum distance (see Tables II). Our algorithm assigns weights to the doublets such that the total energy of the codeword increases with a match and decreases with a mismatch. Hence, this value gets emphasized or de-emphasized when consecutive number of matches or mismatches occurs. The proposed modeling of ribosome decoding is summarized in Algorithm I.

Algorithm I: Exponentially-Weighted Free Energy Metric Based Ribosome (EWFERD) Decoding

Given: Codebook C with L codewords of length N and a subsequence S of length N from the received noisy mRNA sequence. Notation: c_n^k is the n[th] symbol of codeword k, s_n is the n[th] symbol of S, E_k is the exponentially weighted free energy metric when codeword k is used (E_k is initialized to *0,* $0 \le k \le L$), and *Energy(a,b)* is the energy dissipated on binding with the nucleotide doublets *ab* (see Table II, e.g. the energy dissipated by binding with AC is −1.8 kcal/mol). w_k is the weight applied to the doublet in the k^{th} position. σ and $\tilde{\sigma}$ are the numbers of consecutive matches or

This work was supported in part by a grant from the Pritzker Institute, Illinois Institute of Technology.

Mohammad Al Bataineh, Ph.D. Candidate; Maria Alonso, Ph.D. Candidate; L. Huang, PhD Candidate, and G. E. 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, lhuang13@iit.edu, atkin@iit.edu).

N. Menhart, Ph.D., Associate Professor; are with the Department of Biological, Chemical, Physical Sciences, Illinois Institute of Technology, Chicago, IL 60616; 312-567-3123; 312-567-3494 fax (e-mail: menhart@iit.edu).

mismatches respectively, and ρ is an offset variable updated at each step:

EWFERD Algorithm

for $k = 1 \ldots L$ *do*
  Initialize $\sigma_0 = 0, \tilde{\sigma}_0 = 0, \rho_0 = 0, w_1 = a;$
  for $n = 1 \ldots N-1$ *do*
    if $c_n^k c_{n+1}^k$ *and* $s_n s_{n+1}$ *are matching then*
      Increment $\sigma_n = \sigma_{n-1} + 1;$
      set $\tilde{\sigma}_n = 0;$
      $w_n = \rho_n + a^{\sigma_n};$
    else
      Increment $\tilde{\sigma}_n = \tilde{\sigma}_{n-1} + 1;$
      set $\sigma_n = 0;$
      $v = w_1 - (a^{\tilde{\sigma}_n + 1} - a^{\tilde{\sigma}_n});$
      if $n >= 2$
        $v = w_{n-1} - (a^{\tilde{\sigma}_n + 1} - a^{\tilde{\sigma}_n});$
      end
      $w_n = \max(0, v);$
      if $\rho_{n-1} \le a$
        $\rho_n = 0;$
      else
        $\rho_n = \max\left\{w_{n-1} - \left(a^{\tilde{\sigma}_n + 1} - a^{\tilde{\sigma}_n}\right)\right\};$
      end of if
    end of if
    $E_n = E_{n-1} + w_n . Energy(c_n^k c_{n+1}^k);$
  end of for
  $E_k = E_n$
end of for
Calculate the $\min(E_k), 0 \le k \le L.$

The parameter a is a constant that will control the exponential growth of the free energy E_k.

TABLE II: CODEBOOK

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

TABLE I: ENERGY DOUBLETS

Cl	Codeword	C5	GAGGU
C1	UAAGG	C6	AGGUG
C2	AAGGA	C7	GGUGA
C3	AGGAG	C8	GUGAU
C4	GGAGG	C9	UGAUC

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 Shine-Dalgarno in the genes under study.

For the analysis, sequences of the complete genome of the prokaryotic bacteria E. coli strain MG1655. Fig. 1 shows average results for the detection of the SD, start and stop codons with and without using the weighting algorithm (EWFERD) described before. It can be observed that the proposed algorithm is not only able to identify the Shine-Dalgarno (dip at position 90) and the start codon (dip at position 101) and the stop codon (dip at position 398) at their exact corresponding positions, but also with a much better resolution being controlled by the choice of the parameter a. Moreover, these results support the arguments for the importance of the 16S rRNA in translation.

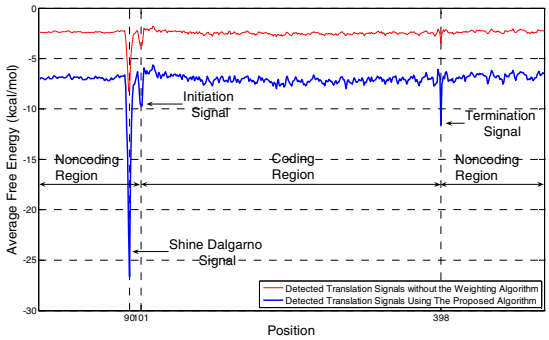

Fig. 1. Detection of translation signals

II. NEW MODELS BASED ON COMMUNICATIONS THEORY

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:

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

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:

$$w(n) = \begin{cases} a^{\sigma} & , \quad if \ match \\ 0 & , if \ mismatch \end{cases}, \quad (2)$$

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 $\overline{W}$:

$$\overline{W} = [w(1), w(2), ..., w(L-N+1)], \quad (3)$$

where L is the length of the mRNA sequence.

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

III. Simulation Results and Mutation Analysis

In this section, we show results of applying the four models described briefly in section II 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) [9]. 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 noncoding 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 noncoding 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

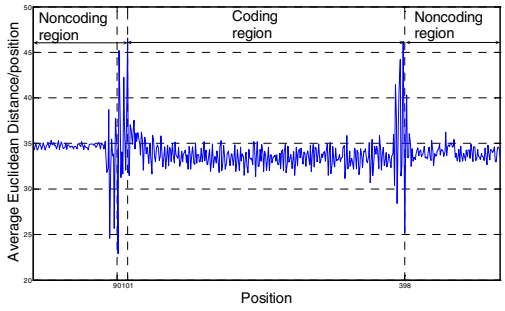

Fig. 2. Model II: Euclidean Distance Metric

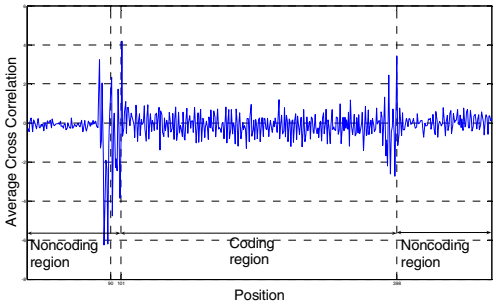

Fig. 3. Model III: Cross Correlation (Matched Filter)

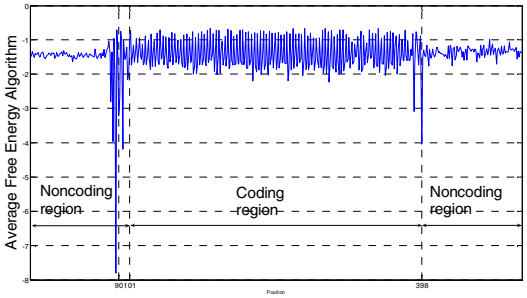

Fig. 4. Model V: Free Energy Metric

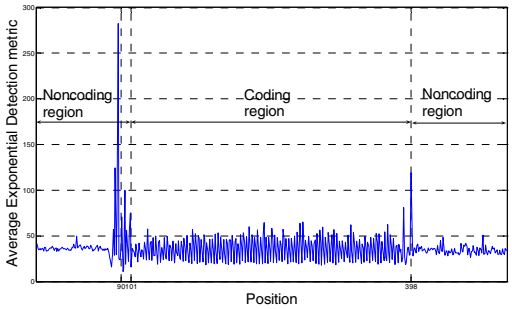

Fig. 5. Model IV: Exponential Detection

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 [10], results in a reduction in the level of protein synthesis. This mutation was tested using Model IV (The Exponential Detection Metric). 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.

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 [11]. The results of both mutations are lethal for the organism in the sense that the production of proteins stop. Figures 7 and 8 show a complete loss of the Shine-Dalgarno (SD) signal (at position 90). Hence, it can be inferred that the translation will never take place.

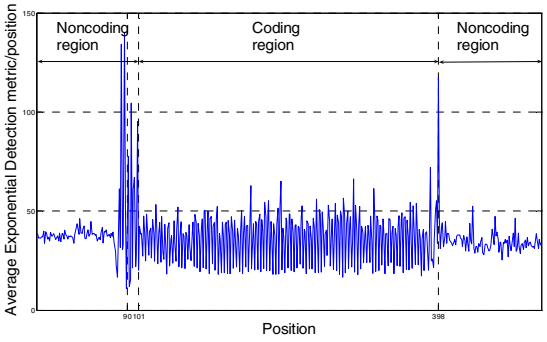

Fig. 6: Jacob Mutation using Exponential Detection

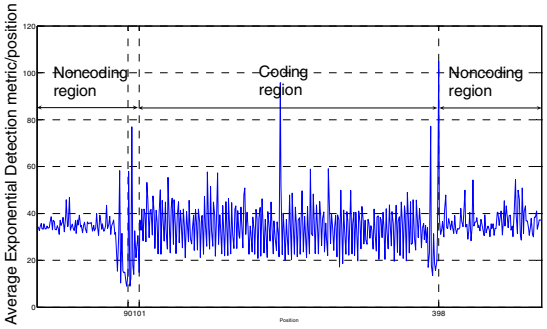

Fig. 7: Hui Mutation using Exponential Detection

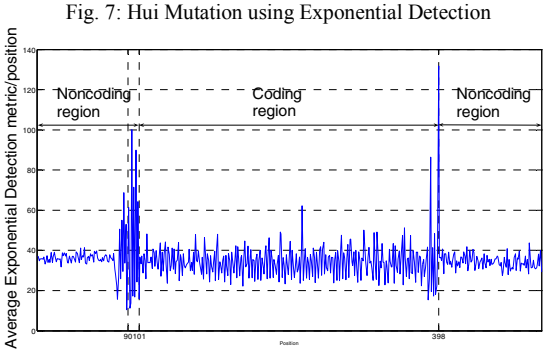

Fig. 8: De Boer Mutation using Exponential Detection

We are in the process of evaluating the developed models to other organisms such as Salmonella typhimurium LT2, Bacillus subtilis, and Staphylococcus aureus Mu50.

IV. 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 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, 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] Mohammad Al Bataineh, Maria Alonso, Lun Huang, Nick Menhart, and Guillermo Atkin, "Gene Expression Analysis using Communications, Coding and Information Theory Based Models,"; In: Proceedings of 2009 International Conference for Bioinformatics and Computational Biology, Las Vegas, June 13 – 16, 2009.

[4] 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.

[5] 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.

[6] 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.

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

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

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

[10] 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.

[11] 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.

[12] J. Shine and L. Dalgarno, "The 3' terminal sequence of Escherichia coli 16S ribosomal RNA:complementarity to nonsense triplets and ribosome binding sites," Proc. Natl. Acad. Sci., vol. 71, pp. 1342–1346, 1974.

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