An Augmented Block Code Model for Protein Translation Using Free Energy Based Distance Decoders

Mohammad Al Bataineh, Lun Huang, Alicia Fuente Acedo, Guillermo Atkin, and Nick Menhart

Abstract—**Informational analysis of genetic sequences has revealed the existence of significant analogies between the genetic process and information processing systems used in the field of communications engineering. By analyzing key elements involved in the process of gene expression, we have developed several communications and coding theory based models for the process of translation [1-5]. A previous research investigated the use of coding theory based models that quantitatively describe the behavior of the ribosome during translation initiation in prokaryotic organisms [1]. In this paper we have investigated an augmented block code model with modified criteria and assumptions. We have also employed several minimum distance decoders to verify the proposed modified model based on the free energies involved in the binding between the ribosome and the mRNA sequence. The key biological elements considered in forming the investigated model are: the last 13 bases of the 3' end of the 16S rRNA molecule, the common features of bacterial ribosomal binding sites (such as the existence and location of the Shine-Dalgarno sequence), the energies involved in the rRNA-mRNA interaction, and RNA/DNA base-pairing principles. The model was tested on five different *E. coli* bacterial genomes. The obtained results prove the validity and significance of the model in clearly distinguishing four different test groups of gene predictions. Two of them are based on well known gene finder softwares (e.g. GeneMark [2] and Glimmer [3]).**

Index Terms—**communications, coding theory, Gene expression, translation initiation, block codes**

I. Introduction

The rapid advances in both genomic data acquisition and computational technology have encouraged development and use of engineering based methods in the field of genetic data analysis. Techniques from engineering fields such as information theory [4-6], communication [7], coding theory [8], signal processing [9], machine learning [10] are now being actively researched for use in gene and regulatory sequence identification. Novel use of techniques and principles from the latter fields are being used for examining and analyzing the genomic structure including coding and non-coding regions. In this paper, concepts and tools block codes are used for the analysis and understanding of the process of translation in gene expression. As a basic analogy, data information is encoded, transmitted and processed in communications, while DNA information is replicated, expressed and processed in genetics. The precision and robustness found in molecular biology motivates the quest to try to explain this behavior using concepts from communications and coding theory.

A fundamental challenge for all communication systems is achieving efficient, secure, and error-free communication over noisy channels. Information theoretic principals have been used to develop effective coding theory and cryptographic algorithms to successfully transmit information from a source to a receiver in transmission systems. Living organisms also successfully transmit their biological information through genetic processes such as replication, transcription, and translation, where the genome of an organism is the content of the transmission.

Informational analysis of genetic sequences has provided significant insight into parallels between the genetic process and information processing systems used in the field of communications engineering. Of particular interest are the results from Schneider et al. [5, 12] and Eigen [13]. Drawing from their work and previous work in protein annotation and gene identification, we make several key observations that lead one to hypothesize that similar to engineering, information-processing systems, the genetic system contains mechanisms to protect an organism from errors that occur within its genome.

By analyzing key elements involved in the process of protein gene expression, we have developed several communications and coding theory based models for the process of translation [14-18]. A previous research conducted by E. May investigated the use of coding theory based models that quantitatively describe the behavior of the ribosome during translation initiation in prokaryotic organisms [1]. In this work have reinvestigated a block code model with modified criteria and assumptions. We have also employed several distance based decoders to verify the proposed modified model. The key biological elements considered in

Manuscript received February 15, 2010.

Mohammad F. Al Bataineh, Ph.D. Candidate, Lun Huang, Ph.D. Candidate, Alicia Fuente Acedo, M.S. student, Guillermo E. Atkin, Ph.D., Senior Member IEEE, are with the Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago IL 60616; phone: 312-567-3417, fax: 312-567-8976; (e-mail: albamoh@iit.edu; lhuang13@iit.edu; afuentea@iit.edu; atkin@iit.edu; menhart@iit.edu;).

Nick Menhart, Ph.D., is with the Department of Biological, Chemical, Physical Sciences, Illinois Institute of Technology, Chicago, IL 60616; phone: 312-567-3123; fax: 312.567.3494 (e-mail: menhart@iit.edu).

978-1-4244-6875-1/10/$26.00 ©2010 IEEE

forming our model are: the 3' end of the 16S ribosomal RNA (16S rRNA), the common features of bacterial ribosomal binding sites (such as the existence and location of the Shine-Dalgarno sequence), the energies involved in the rRNA-mRNA interaction and RNA/DNA base-pairing principles. The model was tested on five different *E. coli* bacterial genomes. The obtained results prove the validity and significance of the model in clearly distinguishing four different test groups of gene predictions. Two of them are based on well known gene finder programs (e.g. GeneMark [2] and Glimmer [3]).

II. Block Code Model

In the investigated block code model, we verify the significance of using communications and coding theory specifically, for quantitatively modeling the protein translation initiation mechanism in the process of gene expression. Based on this model, the messenger RNA (mRNA) can be modeled as a noisy encoded signal and the ribosome as a minimum free energy distance decoder, where the 16S ribosomal RNA (16S rRNA molecule) serves as a template for generating a set of valid codewords (the codebook).

To define a coding alphabet for the investigated block code model, the principles of base pairing, wobble pairing, and translation initiation [2] were considered. The alphabet consists of inosine (I=0), adenine (A =1), guanine (G=2), cytosine (C=3), and uracil (U=4). The RNA bases are mapped such that in modulo-5 addition of the sum of bases that form hydrogen pairs is zero. The n-base codewords were selected based on the systematic (a systematic code contains the k information bases at the beginning of the codeword) zero-parity check encoding methodology, such that the following equation is satisfied

$$\sum_{i=1}^{k} u_i + \sum_{i=1}^{n-k} v_i = 0 \qquad (1)$$

where u_i is the i^{th} out of k information bases, and v_i is the i^{th} out of $(n-k)$ parity bases. The addition sign $(+)$ stands for modulo-5 addition. The k information bases are selected to include all the possible combinations of k number of bases (e.g for k=2, there are 4^2 possible information bases: AA, AG, AC, AU, …, UU). The parity bases are selected to include all the $(n-k)$-base subsequences of the Watson-Crick complement of the last 13 bases of the 3' end of 16S rRNA molecule given by UAAGGAGGUGAUC. For $(n,k)=(5,2)$, the parity bases are UAA, AAG, AGG, …, AUC). The reasoning behind the use of the last 13 bases of 16S rRNA to develop the codewords is due to the important role that this sequence plays in the binding between the ribosome and the mRNA sequence [19].

Based on the proposed criterion to select the parity bases in equation (1), the resulting number of codewords might be greater than 5^{k} (33 codewords in the case of (5,2) block code and 26 codewords in the case of (8,2) block code). This due to the multiple codewords assigned to the same set of information bases to account for the variation in the ribosome binding site structure (RBS). Due to this codeword multiplicity, we refer to the developed block code as an augmented block code.

The analysis sequence *(AS)* used is composed of q bases preceding the initiation codon (AUG or GUG or UUG) and $(q-3)$ bases from the coding region immediately following the initiation codon. Hence, the analysis sequence has the following form

$$AS = \left[b_{-q},\ b_{-(q-1)}, \dots, A, U, G, b_3, b_4, \dots, b_{q-3}\right] \qquad (2)$$

where each base is referenced with respect to its relative position from the initiation codon. The $2q$-bases-long analysis sequence is then split up into $(2q-n+1)$ subsequences each of length (n) bases which are then compared to the set of codewords to decide which codeword is the correct one for each received sequence (r_p). The devised minimum distance metric used for the j^{th} received sequence at the i^{th} alignment is defined as

$$d^{j}_{min_i} = min\left[d\left(r_p, C\right)\right] \qquad (3)$$

where p is a position relative to the initiation codon and C corresponds to the codebook. The minimum distance is recorded for each received sequence in the analysis stream based on different decoding strategies described in section III. This distance is used to evaluate how well the block coding model captures the biological aspects of the initiation process.

The block decoder stores the minimum distance information (MDI) for each sequence group in matrices of the form

$$MDI = \begin{bmatrix} d^1_{min_{-q}} & d^1_{min_{-(q-1)}} & d^1_{min_{(q-2)}} & \cdots & d^1_{min_{(q-n)}} \\ d^2_{min_{-q}} & d^2_{min_{-(q-1)}} & d^2_{min_{(q-2)}} & \cdots & d^2_{min_{(q-n)}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ d^M_{min_{-q}} & d^M_{min_{-(q-1)}} & d^M_{min_{(q-2)}} & \cdots & d^M_{min_{(q-n)}} \end{bmatrix} \qquad (4)$$

where M is the number of analysis sequences the group under test, and $(q-n)$ is the last valid comparison position in the analysis sequence.

To extract the information signal from the noise contained in the developed model, we take the average $\mathrm{d_{min}}$ in value by position for each sequence group. This produces one signal (MDI_{avg}) that describes the minimum distance characteristic of each sequence group

$$\mathrm{MDI_{avg}} = \frac{1}{\mathrm{M}}\sum \mathrm{MDI} \qquad (5)$$

The averaging used in equation (5) is a standard signal processing technique used to amplify a signal in the presence of noise. Averaging suppresses the noise in individual sequences and amplifies the common characteristics among all the sequences in a sequence group. Smaller distance values in the $\mathrm{MDI_{avg}}$ vector indicate stronger hydrogen bond formations between the 16S rRNA molecule and the mRNA sequence.

III. Minimum Distance Decoders

A decoder provides a strategy for selecting the transmitted codeword for a given received sequence. One method, maximum likelihood decoding, compares the received sequence with every possible codeword sequence in the codebook and selects the most likely sequence. In this paper, we have used the following three distance decoders (i.e. three

different strategies to calculate the minimum distance in equation (3)):

A. Minimum Hamming Distance Decoder

The Hamming distance between two strings of equal length is the number of positions at which the corresponding symbols are different. Based on this definition, the minimum Hamming distance is recorded for each received sequence (r_p) in the analysis base stream. The minimum Hamming Distance decoder strategy is described in Algorithm I.

Algorithm I Minimum Hamming Distance Decoding

Given: Codebook C with L codewords of length N and a received sequence 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, d_H^k $(0 \le k \le L)$ is the hamming distance when codeword k is used (d_H^0 is initialized to 0). The devised block decoder selects the codeword that results in the minimum hamming distance.

Minimum Hamming Distance

for $k = 1 \dots L$ **do**
  for $n = 1 \dots N$ **do**
    if $c_n^k \neq s_n$, **then**
      $d_H^k = d_H^{k-1} + 1$
    end if
  end for
end for

$d_{min} = \min(\boldsymbol{d_H})$, where $\boldsymbol{d_H} = [d_H^1, d_H^2, \dots, d_H^L]$

B. Minimum Free Energy Distance Decoder

The proposed minimum free energy distance decoding strategy is summarized in Algorithm II.

Algorithm II Free Energy Distance Decoding

Given: Codebook C with L codewords of length N and a received sequence 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 $(0 \le k \le L)$ is the free energy metric when codeword k is used (E_0 is initialized to 0), and $Energy(a,b)$ is the energy dissipated on binding with the nucleotide doublets ab. The devised block decoder selects the codeword that results in the minimum free energy distance.

Minimum Free Energy

for $k = 1 \dots L$ **do**
  for $n = 1 \dots N-1$ **do**
    if $c_n^k c_{n+1}^k = s_n s_{n+1}$, **then**
      $E_k = E_{k-1} + Energy(c_n^k c_{n+1}^k)$
    else
      $E_k = E_{k-1}$
    end if
  end for
end for

$E_{min} = \min(\boldsymbol{E})$, where $\boldsymbol{E} = [E_1, E_2, \dots, E_L]$

The parameter E_{min} is the minimum free energy distance. $Energy(ab)$ is calculated using Table I. The values refer to the free binding energy resulting from a bond between the listed dinucleotides and their complement. For example, the Watson–Crick bond between AA in the mRNA and UU in the 16S rRNA yields a free energy of −0.9 kcal/mol.

TABLE I
ENERGY TABLE (KCAL/MOL) [20]

Free Energy Doublets							
AA	-0.9	AG	-2.3	GA	-2.3	GG	-2.1
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

C. Minimum Exponentially-Weighted Free Energy Distance Decoder

The decoding strategy devised here represents an enhanced version of the one proposed in *B*. This strategy is summarized Algorithm III.

Algorithm III Exponentially-Weighted Free Energy Distance Ribosome Decoding

Given: Codebook C with L codewords of length N and a received sequence 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 $(0 \le k \le L)$ is the exponentially-weighted free energy when codeword k is used (E_0 is initialized to 0), and $Energy(a,b)$ is the energy dissipated on binding with the nucleotide doublets ab. 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 mismatches respectively, and ρ is an offset variable updated at each step. The parameter a is a constant that controls the exponential growth of the free energy E_k.

Minimum Exponentially-Weighted Free Energy

for $k = 1 \dots L$ **do**
  Initialize $\sigma_0 = 0, \tilde{\sigma} = 0, \rho_o = 0, w_1 = a$
  for $n = 1 \dots N-1$ **do**
    if $c_n^k c_{n+1}^k = s_n s_{n+1}$, **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 $n \ge 2$, **then**
        $v = w_{n-1} - (a^{\tilde{\sigma}_n+1} - a^{\tilde{\sigma}_n})$
      end if
      $w_n = max(0, v)$
      if $\rho_{n-1} \le a$, **then**
        $\rho_n = 0$
      else
        $\rho_n = \max\{w_{n-1} - (a^{\tilde{\sigma}_n+1} - a^{\tilde{\sigma}_n}), 0\}$
      end if
    end if
    $E_n = E_{n-1} + w_n \times Energy(c_n^k c_{n+1}^k)$
  end if
end for

$E_{min} = \min(\boldsymbol{E})$, where $\boldsymbol{E} = [E_1, E_2, \dots, E_L]$

$Energy(ab)$ is calculated using Table I.

IV. Analysis Data Preparation

The complete prokaryotic genome sequences required for the analysis in this paper were obtained from the National Center for Biotechnology Information (NCBI) [21]. Using MATLAB, we developed a toolbox to extract and manipulate the data required and put it in a format suitable to our analysis. This toolbox will be disseminated through our research lab website to be publically available. Using this toolbox, we extracted the following data from the NCBI for each tested genome: 1) the complete DNA sequence, 2) the exact locations of all known genes in the forward and reverse strands, 3) gene predictions obtained by GeneMark, 4) gene predictions obtained by Glimmer, and 5) the set of all possible open reading frames based on a pre-specified criteria. Based on this analysis, we were able to classify the tested data into four different groups (See Figure 1):

1. Actual Translated Sequences (4,423 sequences): Open reading frames which GenBank indicates as sequences that translate into proteins,
2. GeneMark Hypothetically Translated Sequences (905 sequences): Open reading frames which GeneMark indicates as genes but are actually not (GeneMark false positives),
3. Glimmer Hypothetically Translated Sequences (3110 sequences): Open reading frames which Glimmer indicates as genes but are actually not (Glimmer false positives),
4. Non-Translated Sequences (21,441 sequences): Open reading frames which do not appear on the list of Actually translated or hypothetically translated sequences. For this work, the open reading frame had to have: 1) A valid initiation codon; 2) A valid termination codon; 3) A sequence length greater than or equal to ninety-nine bases.

The sequence numbers given above for the four test groups are for *Salmonella Typhimurium LT2* genome. The following block diagrams give an illustrative description of the approach used to prepare the data required for analysis.

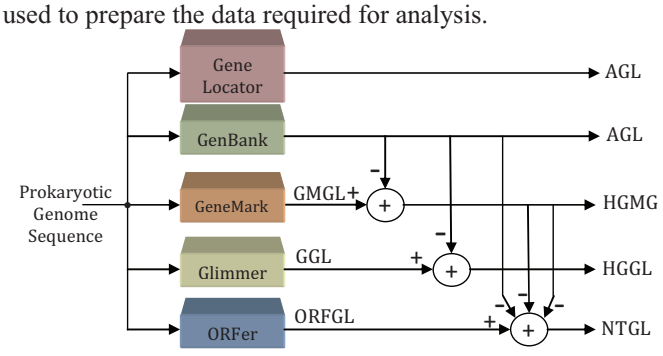


Figure 1. Schematic diagram of the analysis data preparation

The notations used is Figure 1 are: AGL (Actual Gene Locations) corresponds to group 1, HGMGL (GeneMark Gene Locations) corresponds to group 2, HGGL (Hypothetical Glimmer Gene Locations) corresponds to group 3, and NTGL (Non-Translated Gene Locations) corresponds to group 4. The intermediate parameter GMGL (GeneMark Gene Locations) corresponds to the gene predictions obtained by GeneMark, GGL (Glimmer Gene Locations) corresponds to the gene predictions obtained by Glimmer, and ORFGL (Open Reading Frame Gene Locations) corresponds to the set of all possible genes that are greater than 99 beases long and start with a valid start codon and end up with a valid stop codon.

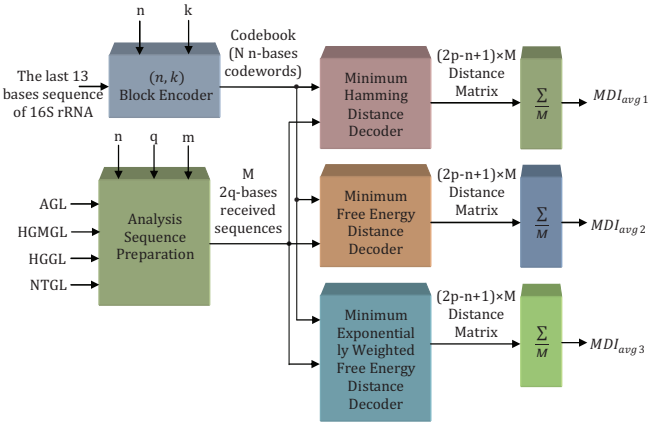


Figure 2. Schematic diagram of the proposed augmented block code model

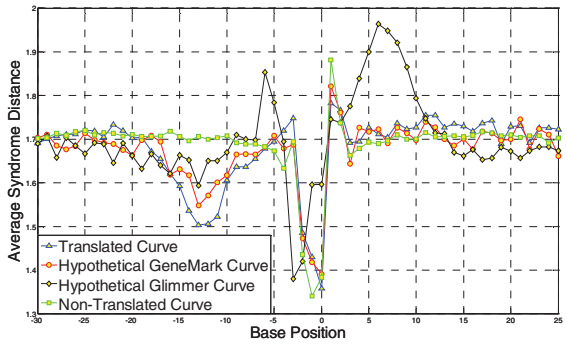
Figure 2 shows a schematic diagram of the proposed augmented block code model. The input parameters are 1) n and k: the block code parameters, 2) AGL, HGMGL, HGGL, and NTGL obtained in figure 1, and 3) the last 13 bases of the 3' end of 16S rRNA molecule. The parameter q is the number of bases upstream of the initiation codon where the beginning of our analysis tested sequences was selected, m is an input parameter to determine which of the four inputs (AGL/ HGMGL/ HGGL/ NTGL) to select, N is the number of codewords in the block codebook, M is the number of analysis sequences being processed (different for each input).

V. Simulation Results

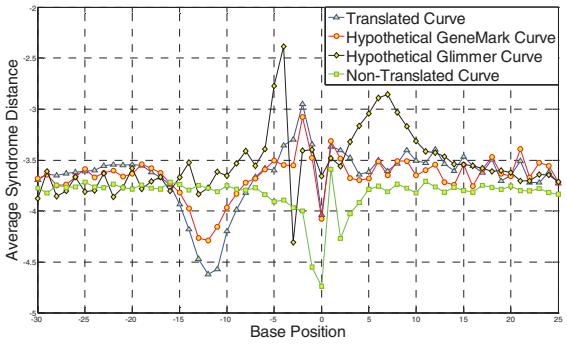
To test our model, we have applied the developed model to *Escherichia coli* K-12 MG1655 strain, *Escherichia coli* O157:H7 strain, and to other prokaryotic organisms of varying taxonomical relation to *E. coli* including: *Salmonella Typhimurium LT2*, *Bacillus Subtilis*, and *Staphylococcus Aureus Mu50* (The simulation results shown in Figures 3 and 4 are obtained for *Salmonella Typhimurium LT2*). The model was able to successfully identify and distinguish the regions on the 5' untranslated leader regions where the minimum free energy distance values of translated mRNA subsequences, hypothetically translated subsequences (obtained by GeneMark and Glimmer), and non-translated subsequences differ the most. These regions correspond to the Shine-Dalgarno domain and the non-random domain that exists in the genomic structure of the tested sequences.

Figures 3 and 4 clearly show significant differences between the translated, hypothetically translated and the non-translated sequence groups. The horizontal axis in figures 3 and 4 represent the position relative to the first base of the initiation codon. The vertical axis shows the mean of the aligned minimum distance values of the sequences in each of the four sequence groups. For the translated and hypothetically translated sequence groups, a minimum distance trough occurs

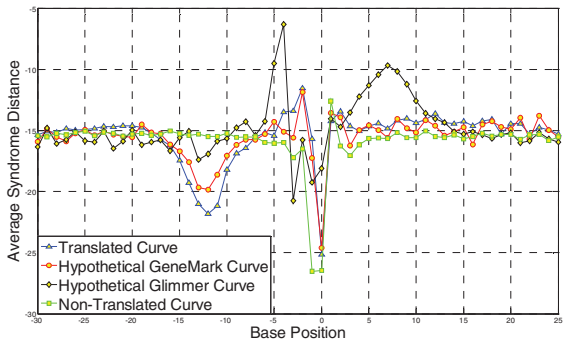
between the -15 and -10 regions. All the sequence groups in the (5,2) and (8,2) models achieve a global minimum distance value in the -5 to 0 region. The -15 to 0 region contains large synchronization signals which can be used to determine valid protein coding sequences or frames. There are also smaller synchronization signals outside the -15 to 0 region which seem to oscillate with a frequency of three (period-3 property). The results of the longer (8,2) block code model in Figure 4 illustrate the effect of two or more codons while the (5,2) block code model is affected by at most two codons.



(a)



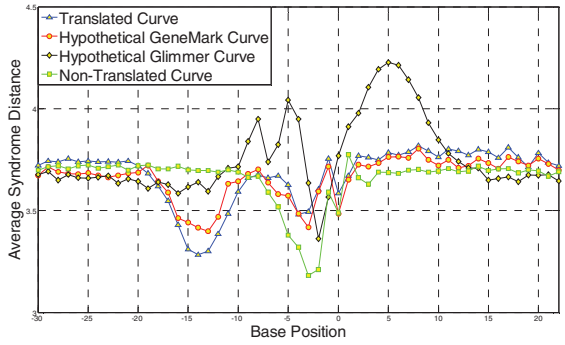
(b)



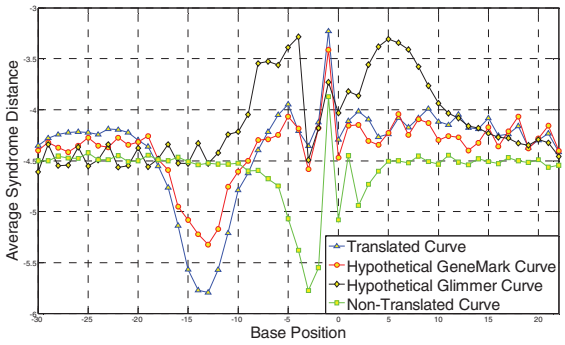
(c)

Figure 3. (5,2) block code model output for *Salmonella Typhimurium LT2* using (a) minimum Hamming distance decoder, (b) minimum free energy distance decoder, and (c) minimum exponentially-weighted free energy distance decoder

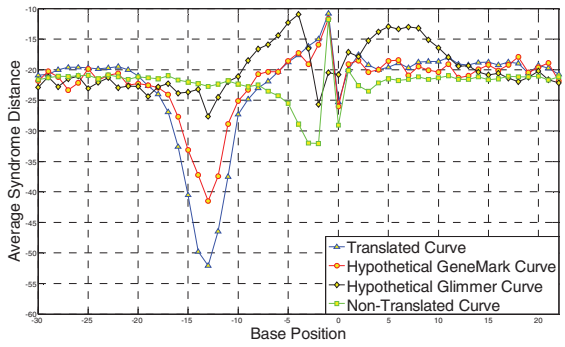
The simulation results obtained for the other prokaryotic organisms tested in this work are quite similar to the ones shown in figures 3 and 4. This further certifies the correctness and the biological relevance of the proposed block code model to identify and distinguish the four different test groups (translated, hypothetically translated, and non-translated test groups).



(d)



(e)



(f)

Figure 3. (8,2) block code model output for *Salmonella Typhimurium LT2* using (d) minimum Hamming distance decoder, (e) minimum free energy distance decoder, and (f) minimum exponentially-weighted free energy distance decoder

VI. Conclusion

An augmented block code model with a free energy based distance decoder can distinguish translated, hypothetically translated, and non-translated sequence groups. The hypothetically translated sequence groups represent the gene predictions false positives obtained by GeneMark and Glimmer gene finding softwares. The simulation results of the proposed model indicate the existence of key regions within the mRNA leader region. The block code model recognizes the ribosomal binding site (the location of the Shine-Dalgarno sequence) readily. The model also identifies the non-random domain, the region upstream of the Shine-Dalgarno domain believed to also affect translation initiation. We were able to successfully apply the model to several prokaryotic organisms,

including *Escherichia coli* K-12 MG165, *Escherichia coli* O157:H7, *Salmonella typhimurium LT2*, *Bacillus Subtilis*, and *Staphylococcus Aureus Mu50*. The results of this work suggest that it is possible to design a coding based algorithm for distinguishing between protein coding and non-protein coding genomic sequences by "decoding" the mRNA leader region. The success of this work can lead to the development of improved methods for identifying the precise location of translation initiation start sites. Additionally, design of effective coding-based models for genetic regulatory systems can potentially help researchers determine how to incorporate deliberate, sequence-controlled regulation into engineered proteins. Such a tool would be useful for designing regulatory sequences for transgenic organisms, as well as further our understanding of the translation regulatory mechanisms.

REFERENCES

[1] Elebeoba E. May, "Analysis of Coding Theory Based Models for Initiating Protein Translation in Prokaryotic Organisms," in *IEEE Transactions on Information Technology in BioMedicine*. vol. PhD thesis Raleigh, NC: North Carolina State University, March 2002.

[2] "GeneMark: A family of gene prediction programs developed at Georgia Institute of Technology, Atlanta, Georgia, USA. URL: http://exon.biology.gatech.edu/."

[3] "GLIMMER: Microbial Gene-Finding System, University of Maryland Institute for Advanced Computer Studies, Center for Bioinformatics & Computational Biology (CBCB), URL: http://www.cbcb.umd.edu/software/glimmer/."

[4] R. RomanRoldan, P. BernaolaGalvan, and J. L. Oliver, "Application of information theory to DNA sequence analysis: A review," *Pattern Recognition,* vol. 29, pp. 1187-1194, Jul 1996.

[5] T. D. Schneider, "Information content of individual genetic sequences," *J Theor Biol,* vol. 189, pp. 427-41, Dec 21 1997.

[6] G. Battail, "Should genetics get an information-theoretic education?," *Ieee Engineering in Medicine and Biology Magazine,* vol. 25, pp. 34-45, Jan-Feb 2006.

[7] 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.

[8] E. E. May, M. A. Vouk, D. L. Bitzer, and D. I. Rosnick, "An error-correcting code framework for genetic sequence analysis," *Journal of the Franklin Institute-Engineering and Applied Mathematics,* vol. 341, pp. 89-109, Jan-Mar 2004.

[9] V. Veljkovic, I. Cosic, B. Dimitrijevic, and D. Lalovic, "Is It Possible to Analyze DNA and Protein Sequences by the Methods of Digital Signal-Processing," *Ieee Transactions on Biomedical Engineering,* vol. 32, pp. 337-341, 1985.

[10] E. C. Uberbacher and R. J. Mural, "Locating protein-coding regions in human DNA sequences by a multiple sensor-neural network approach," *Proc Natl Acad Sci U S A,* vol. 88, pp. 11261-5, Dec 15 1991.

[11] J. Henderson, S. Salzberg, and K. H. Fasman, "Finding genes in DNA with a Hidden Markov Model," *Journal of Computational Biology,* vol. 4, pp. 127-141, Sum 1997.

[12] T. D. Schneider, G. D. Stormo, L. Gold, and A. Ehrenfeucht, "Information content of binding sites on nucleotide sequences," *J Mol Biol,* vol. 188, pp. 415-31, Apr 5 1986.

[13] M. Eigen, "The origin of genetic information: viruses as models," *Gene,* vol. 135, pp. 37-47, Dec 15 1993.

[14] M. Al Bataineh, M. Alonso, S. Wang, W. 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.

[15] M. Al Bataineh, M. Alonso, S. Wang, G. E. Atkin, and W. Zhang, "An Optimized Ribosome Binding Model Using Communication Theory Concepts," *Proceedings of 2007 International Conference for Bioinformatics and Computational Biology,* pp. 345-348, June 25 – 27, 2007.

[16] M. Al Bataineh, M. Alonso, L. Huang, G. E. Atkin, and N. Menhart, "Effect of mutations on the detection of translational signals based on a communications theory approach," *Conf Proc IEEE Eng Med Biol Soc,* vol. 1, pp. 3853-6, 2009.

[17] M. Al Bataineh, L. Huang, I. Muhamed, N. 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.

[18] M. Al Bataineh, L. Huang, M. Alonso, N. Menhart, and G. E. Atkin, "Analysis of Gene Translation Using a Communications Theory Approach," in *Advances in Computational Biology*: Springer, December, 2009.

[19] B. Lewin, "Genes V," *Oxford University Press, New York, NY,* 1995.

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

[21] "NCBI RefSeq bacterial genomes, URL: ftp://ftp.ncbi.nih.gov/genomes/Bacteria/."