The Ribosome as a Table-Driven Convolutional Decoder for the Escherichia coli K-12 Translation Initiation System

Elebeoba E. May*, Mladen A. Vouk*, Donald L. Bitzer*, and David I. Rosnick

Abstract— **Redundancy occurs naturally within RNA and DNA sequences [1]. The existence of tandem repeats, and sequences such as the Shine-Dalgarno sequence, the Pribnow box and the TATA box, leads us to believe that cellular communication systems use some method of coding to recognize valid information regions within a nucleotide sequence and correct for "transmission" errors such as mutations.**

In this paper we use principles of convolutional coding theory to analyze the translation initiation process. The principle hypothesis is that the messenger RNA (mRNA) sequence can be viewed as a noisy, convolutionaly encoded signal. The ribosome is functionally paralleled to a table-driven convolutional decoder. The 16s ribosomal RNA (rRNA) sequence is used to form decoding masks for table-driven decoding. The results of applying this method to Escherichia coli K-12 strain MG1655 are presented.

Keywords— **Coding Theory, Table-driven Codes, Convolutional Codes, Translation Initiation**

I. Introduction

Increased availability of genomic data continues to encourage the development of computational tools and methods for the analysis of such data. Some computational methods for genomic data analysis are based on engineering and mathematical constructs such as neural networks [2], Hidden Markov Models [3] [4], fractals [5], and Fourier transforms [6].

We believe that the translation of the messenger RNA (mRNA) sequence into chains of protein-forming aminoacids is analogous to the decoding of an information sequence [7] [8]. In a communication system, coding techniques are used to compensate for errors that occur during transmission of information [9]. Error control is accomplished through redundancy.

Convolutional coding produces encoded blocks based on present and past information. It seems reasonable to assume that genetic operations such as initiation and translation may involve "decisions" which are based on the immediate neighborhood of a codon. This would allow error correction and other related functions. Considering messenger RNA as convolutionally encoded data, allows the model to capture the inter-relatedness between the bases in a mRNA sequence.

In the sections which follow, we give a brief overview of convolutional coding and table-driven decoding. We also discuss the relationship to the genetic process, and present the methodology for forming a table-driven genetic decoder. The preliminary results of applying the genetic decoder to E. coli K-12, and possible implications of the model, are analyzed and discussed.

II. Theoretical Background

The mathematics of coding is carried out over a finite field, using a set of discrete source symbols [9]. In convolutional encoding, a n-bit encoded block at time i depends on the k-bit information block at time i and on m previous information blocks [10]. Hence, a convolutional encoder requires memory. Convolutional codes are referred to as (n, k, m) codes or (n, k) codes.

A codeword (or correct set of symbols) is the output of the convolutional encoder for a given input data block [11]. A decoder provides a strategy for selecting an estimated codeword for each possible received sequence. There are various decoding methods. One method, maximum likelihood decoding, compares the received sequence with every possible code sequence that the encoding system could have produced and selects the most likely sequence.

A. Table-Driven Encoders and Decoders

The existence of a one to one mapping between data bits and parity bits (parity bits are encoded bits) is the basis for table-driven encoding and decoding [12]. A set of w-bit data blocks must correspond uniquely to a set of w-bit parity blocks. For an (n, k, m) code

$$w = n\frac{L-k}{n-k} \tag{1}$$

where

$$L = m + 1 \tag{2}$$

The methodology for table-driven encoding is described in [12].

A way to recognize a correct sequence is to form the so called syndrome. The syndrome vector is zero if there are no detectable errors in the parity stream (i.e exact match between overlapping bits); otherwise, for binary data, the syndrome value is one. The number of syndrome values in a syndrome vector is equivalent to the number of overlaps used to determine the vector. Syndrome vectors can also be used to correct the encoded data bits using the table look-up method [12] [13] [14] [15].

A.1 G-Mask

The syndrome vector, which is used to detect errors in the parity stream, can be generated by repeated applica-

*Department of Electrical and Computer Engineering
North Carolina State University Raleigh,NC
E. E. May, corresponding author: euenil@eos.ncsu.edu

tion of the decoding table to the parity stream. The g-mask provides an efficient method for syndrome vector generation. The values that comprise the g-mask are based on the codewords of the encoding system. The g-mask is $w+n$ bits long. Generation of the g-mask, given a decoding table, is described in [12].

Once the g-mask has been constructed, it can be used to calculate the syndrome vector for the parity stream. To calculate the syndrome vector using the g-mask:

- The g-mask is ANDed with the first $w+n$ parity bits.
- The result is exclusive-ORed to produce a syndrome value.
- The received parity stream is shifted by n bits.
- The process is repeated until all syndrome values of the syndrome vector are produced. Each shift results in one syndrome value.

Based on the value of the syndrome vector, the received parity sequences can be used to estimate the transmitted sequence or used to detect errors in the transmission. The concept of a decoding mask, the g-mask, is employed in our convolutional coding model for the translation-initiation system.

III. Methods

The messenger RNA is considered as a received parity sequence of a convolutional encoded data stream [16]. The coding alphabet must be derived from a finite field as in the binary code. Using base pairing, wobble pairing, and translation initiation information [1] the RNA bases were mapped to the field of five as follows: Inosine(I) = 0 Adenine(A) = 1 Guanine(G) = 2 Cytosine(C) = 3 Uracil(U) = 4. Multiplication and addition are modulo five. The RNA bases are defined so that in modulo five addition the sum of bases that pair is zero. These definitions are used to construct the convolutional code model for the the protein translation initiation process.

A. Messenger RNA as a Convolutionally Encoded Sequence

This work uses the syndrome developed for table-driven decoding to check whether a messenger RNA protein translation initiation region can be interpreted using covolutional coding model.

The formation of bonds between the mRNA and the 16s rRNA significantly influence the initiation of protein translation. When a base on the mRNA pairs with a base on the 16s rRNA, hydrogen bonds are formed. The greater the number of consecutive pairings formed between these two RNA molecules, the greater the probability of translation initiation. Every time the 16s ribosomal subunit attaches to the mRNA, a bonding pattern is formed. The bonding pattern that results in a positive signal is the bonding pattern with high numbers of consecutive hydrogen bonds. This process of locating regions on the mRNA which form high numbers of consecutive hydrogen bonds can be paralleled to locating parity blocks which produce zero syndrome vectors for a received parity stream.

In order to use the table-driven decoding model, we must define biological coding constructs which are analogous to the following coding concepts: the decoding mask, syndrome, and interpretation of syndrome values.

A.1 Genetic G-Mask Based on 16s rRNA

The g-mask selects which bits are included in the exclusive-OR operation. For binary data, the bits in the decoding window associated with the g-mask are the bits used to determine the syndrome vector.

For the genetic model, the genetic g-mask is derived from the 16s rRNA sequence:

$$3' AUUCCUCCACUAG...5'$$

The equivalent field-five mapping is:

$$3'...1\ 4\ 4\ 3\ 3\ 4\ 3\ 3\ 1\ 3\ 4\ 1\ 2...5'$$

The genetic g-mask is formed from subsets of contiguous bases of the 16s rRNA. The subsets indicate which $n+w$-base region is being included in the exclusive-OR operation of the ribosome. Selecting subsets of the 16s rRNA corresponds to base pairing between regions of the 3' end of the 16s rRNA and regions within the mRNA sequence. Assuming a coding model with n=2, $k=1$, $L=3$, and w=4, the length of the genetic g-mask is $w+n$ or six. A g-mask for the translation initiation system can be selected from a table of eight possible six-base genetic g-mask values derived from the 16s rRNA [7].

For the chosen g-mask, the syndrome values of a stream of mRNA codons can be calculated. The received mRNA parity (or codon) sequence includes the last thirty bases of the leader region, the initiation codon, and the first nine codons of the translated region:

$$mRNA = [b_{-30}\ b_{-29}\ ...\ b_{-1}\ A\ U\ G\ b_{+3}\ ...\ b_{+29}] \quad (3)$$

with,

$$b_i = [A,\ G,\ C,\ U] \quad (4)$$

A.2 Syndrome Calculation

The syndrome value for a given mRNA is calculated by ANDing the received codon bases with the genetic g-mask and exclusive-ORing the result. Multiplication (AND) and addition (XOR) are modulo-five. For example [7], given the following mRNA sequence

$$mRNA = [AUGUGAUCUC]$$

and the following six-base g-mask (which is in essence an element of the Shine-Dalgarno sequence)

$$g-mask = [CACUAG]$$

the first three syndrome values are calculated as follows:
A U G U G A U C U C ⇐ mRNA
C A C U A G ⇐ g-mask
The numerical equivalences are:
1 4 2 4 2 1 4 3 4 3

3 1 3 4 1 2

3+4+1+1+2+2 = s_1 = 3
Shift by n=2:
G U G A U C U C ⇐ *mRNA*
C A C U A G ⇐ g-mask
The numerical equivalences are:
2 4 2 1 4 3 4 3
3 1 3 4 1 2

1+4+1+4+4+1 = s_2 = 0 Note that this is an exact pairing match between the six-base mRNA sequence and the g-mask
Shift by n=2 again:
G A U C U C ⇐ *mRNA*
C A C U A G ⇐ $g - mask$
The numerical equivalences are:
2 1 4 3 4 3
3 1 3 4 1 2

1+1+2+2+4+1 = s_3 = 1
This work looks for a correlation between syndrome values and the position of the genetic g-mask relative to the translation initiation codon.

A.3 Distance Value Derivations

In binary table-driven decoding, a syndrome value of zero indicates that there are no detectable errors within the parity stream. For the translation initiation system, it would be ideal if syndrome values could be used to determine the presence or absence of valid ribosome binding sites. The presence of a valid ribosome binding site would indicate a valid translation initiation site.

In the example in the preceding section our syndrome vector S was $[3, 0, 1]$. The zero syndrome value occurred when an exact complement of the six-base genetic g-mask appeared in the decoding window. Theoretically, a zero syndrome value should occur when an exact complement to the genetic gmask is present in the decoding window. But experiments indicate that the genetic g-mask match value (the syndrome value resulting from the presence of an exact complement to the genetic g-mask in the decoding window) does not always result in a zero syndrome value [7].

Since various g-masks yield different mask match values, syndrome values are normalized by transforming each syndrome value to represent the distance of the syndrome value from the genetic g-mask match value. For example, if the genetic g-mask match value is three and the resulting syndrome value is four then the normalized syndrome value or distance representation is one because $3 + 1 = 4$. Hence the normalization equation for a syndrome value s, given a genetic g-mask match value mm, is as follows:

$$distance = [(s + 5) - mm] \bmod 5 \qquad (5)$$

The algorithm and table for conversion from syndrome value to distance value is presented in [7] for different values of mm. These normalized distance values are used to evaluate the convolutional coding model of the translation-initiation system.

IV. Results

The Escherichia coli K-12 strain MG1655 sequence data (downloaded for the NIH ftp site: ncbi.nlm.nih.gov) was used to test the model for two genetic g-mask lengths. The extracted mRNA sequence data was divided into three groups: translated, hypothetical translated, and non-translated. The translated sequences are sequences in the E.coli K-12 genome which GenBank indicates as sequences which translate into proteins. The hypothetical-translated sequences are sequences which GenBank indicates are hypothetically translated into proteins. Non-translated sequences are open reading frames which do not appear on either the translated or hypothetical translated list for GenBank. Figure 1 shows the frequency of the most frequent distance pattern among all possible two-symbol distance patterns $d_i d_j$, where i and j range from zero to four. The horizontal axis indicates position, with zero corresponding to the alignment of the g-mask with the first base of the initiation codon. The vertical axis indicates frequency (0.04 corresponds to four percent). The expected frequency of occurrence for a random, two-symbol distance pattern is four percent. Patterns with frequency of occurrence values greater than four percent are considered significant. The greater the frequency of occurrence the greater the significance of the pattern.

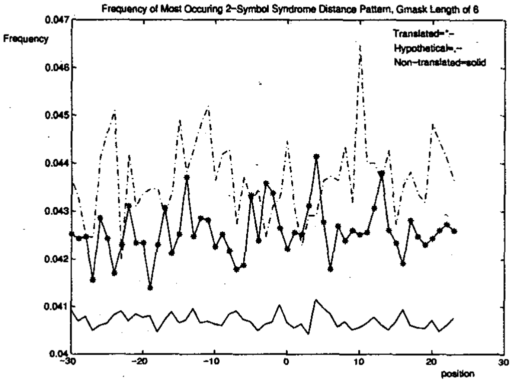

Fig. 1. Frequency of Two-Pattern Syndrome Distance Values

V. Discussion

Figure 1 indicates that the hypothetical group contained the greatest frequency of occurrence values, followed by the translated and the non-translated group. Prior to the zeroth position (position of the initiation codon), the highest frequency value for a given distance pattern occurs around -14 for translated regions. There is a distinction in the frequency of occurrence of two-symbol patterns between the translated/hypothetical group and the non-translated group. The two-pattern frequency analysis for the syndrome values is used as a preliminary indicator to test whether syndrome values can correlate to information. The results indicate that the translated group contains more two-pattern syndrome values than the non-translated. The distance in pattern frequency percentages between translated and non-translated my vary for greater pattern lengths; further research into this continues.

The results for the longer g-mask (twelve-base masks) and additional analyses are presented in [7].

VI. Conclusion

The convolutional model appears to distinguish between translated and non-translated sequences. The distinction between hypothetical and translated groups is also evident. In the convolutional model, the higher frequency present in the hypothetical group (when compared to the translated group) may be a result of biasing introduced through hypothetical sequence identification methods, which are based on finding statistically significant patterns within possible reading frames. The convolutional code model indicates greater information, or occurrence of significant activity, in the area spanning the -15 to 0 region. The Shine-Dalgarno sequence is located within this region [1].

The preliminary results of our work suggest that it may be possible to design a convolutional coding based heuristic for distinguishing between protein coding and non-protein coding genomic sequences by "decoding" the mRNA leader region. We are presently researching methods for developing g-masks which produce consistent syndrome patterns that distinguish translated and non-translated sequences with high accuracy. The success of this work can lead to the development of methods for identifying protein coding sequences within a genome as well as further our understanding of the translation regulatory mechanisms.

Acknowledgments

This research is supported in part by an NSF Graduate Fellowship Grant.

References

[1] Benjamin Lewin, *Genes V*, Oxford University Press, New York, NY, 1995.

[2] Edward C. Uberbacher and Richard J. Mural, "Locating Protein-Coding Regions in Human DNA Sequences by a Multiple Sensor-Neural Network Approach ", *Proceedings of the National Acadamy of Science, USA*, vol. 88, pp. 11261–11265, December 1991.

[3] John Henderson, Steven Salzberg, and Kenneth H. Fasman, "Finding Genes in DNA with a Hidden Markov Model", *Journal of Computational Biology*, pp. 127–1441, 1997.

[4] A. Krogh, I. Mian, and D. Haussler, "A Hidden Markov Model that Finds Genes in E. Coli DNA ", *Nucleic Acids Research*, vol. 22, pp. 4768–4778, December 1994.

[5] A. Arneodo, Y. d'Aubenton Carafa, E. Bacry, P. V. Graves, J. F. Muzy, and C. Thermes, "Wavelet based fractal analysis of DNA sequences", *Physica D*, pp. 1–30, 1996.

[6] Irena Cosic, *The Resonant Recognition Model of Macromolecular Bioactivity: Theory and Applications*, Birkhauser Verlag, Basel, Switzerland, 1997.

[7] Elebeoba E. May, "Comparative Analysis of Information Based Models for Initiating Protein Translation in Escherichia coli K-12", Master's thesis, NCSU, December 1998.

[8] Elebeoba E. May, Mladen A. Vouk, Donald L. Bitzer, and David I. Rosnick, "Coding Model for Translation in E. coli K-12 ", in *First Joint Conference of EMBS-BMES.*, 1999.

[9] Peter Sweeney, *Error Control Coding an Introduction*, Prentice Hall, New York, NY, 1991.

[10] Ajay Dholakia, *Introduction to Convolutional Codes with Applications*, Kluwer Academic Publishers, Norwell, Massachusetts, 1994.

[11] Shu Lin and Daniel J. Costello Jr., *Error Control Coding: Fundamentals and Applications*, Prentice-Hall, Inc., Englewood Cliffs, NJ, 1983.

[12] Donald L. Bitzer and Mladen A. Vouk, "A Table-Driven (Feedback) Decoder", in *Tenth Annual International Phoenix Conference on Computers and Communications*, 1991, pp. 385–392.

[13] A. Dholakia, D. L. Bitzer, and M. A. Vouk, "Table based decoding of rate one-half convolutional codes", *IEEE Trans. on Communications*, vol. 43, pp. 681–686, 1995.

[14] D. L. Bitzer, A. Dholakia, H. Koorapaty, and M. A. Vouk, "On Locally Invertible Rate-1/n Convolutional Encoders", *IEEE Trans. on Information Theory*, vol. 44, pp. 420–422, 1998.

[15] D. L. Bitzer, M. A. Vouk, A. Dholakia, E. Gonzalez, L. F. Wang, V. Srinivasan, T. M. Lee, S. Lo, and H. Koorapaty, "System and method for Enoding and Decoding of Convolutionally Encoded Data", January 1995, United States Patent Number 5,381,425.

[16] Donald L. Bitzer, Mladen A. Vouk, and Ajay Dholakia, "Genetic Coding Considered as a Convolutional Code", North Carolina State University, Raleigh, 1992.