Coding Theory Based Models for Protein Translation Initiation in Prokaryotic Organisms

Elebeoba E. May [a,*], Mladen A. Vouk [b], Donald L. Bitzer [b], David I. Rosnick [b]

[a] *Computational Biology Department, Sandia National Laboratories, Albuquerque, NM 87185 USA*

[b] *Computer Science Department, North Carolina State University, Raleigh, NC 27695 USA*

Abstract

Our research explores the feasibility of using communication theory, error control (EC) coding theory specifically, for quantitatively modeling the protein translation initiation mechanism. The messenger RNA (mRNA) of *Escherichia coli* K-12 is modeled as a noisy (errored), encoded signal and the ribosome as a minimum Hamming distance decoder, where the 16S ribosomal RNA (rRNA) serves as a template for generating a set of valid codewords (the codebook). We tested the E. coli based coding models on 5' untranslated leader sequences of prokaryotic organisms of varying taxonomical relation to *E. coli* including: *Salmonella typhimurium* LT2, *Bacillus subtilis*, and *Staphylococcus aureus* Mu50. The model identified regions on the 5' untranslated leader where the minimum Hamming distance values of translated mRNA sub-sequences and non-translated genomic sequences differ the most. These regions correspond to the Shine-Dalgarno domain and the non-random domain. Applying the EC coding-based models to *B. subtilis*, and *S. aureus* Mu50 yielded results similar to those for *E. coli K-12*. Contrary to our expectations, the behavior of *S. typhimurium* LT2, the more taxonomically related to *E. coli*, resembled that of the non-translated sequence group.

Key words:
Coding Theory, Translation Initiation, Information Theory, Information processing

* Corresponding author
Email address: eemay@sandia.gov (Elebeoba E. May).

1 Introduction

A fundamental challenge for all communication systems, engineered or living, is the problem of 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 engineered systems. Living systems also successfully transmit their biological information through genetic processes such as replication, transcription, and translation, where the genome of an organism is the contents of the transmission.

The study of the information processing capabilities of living systems began in the 1970s (Roman-Roldan et al., 1996; Sarkar et al., 1978; Fowler, 1979) and was revived in the later part of the 1980s, due to the increase in genomic data which spurred a renewed interest in the use of information theory in the study of genomics. Information measures, such as entropy, have been used in recognition of DNA patterns, classification of genetic sequences, and other computational studies of genetic processes (Roman-Roldan et al., 1996; Palaniappan and Jernigan, 1984; Almagor, 1985; Schneider, 1991b,a; Altschul, 1991; Salamon and Konopka, 1992; Oliver et al., 1993; DeLaVega et al., 1996; Schneider and Mastronarde, 1996; Strait and Dewey, 1996; Pavesi et al., 1997; Loewenstern and Yianilos, 1997; Schneider, 1997, 1999). Applying techniques from Coding Theory, a subfield of Information Theory, is a logical next step in the study of the information processing mechanisms of genetic systems.

Application of channel coding theory to genetic data dates back to the late 1950s (Hayes, 1998; Golomb, 1962) with the mapping of the genetic code (the codon to amino acid mapping). Since then coding theoretic methods have been used for frame determination, motif classification, oligo-nucleotide chip design, and DNA computing (Arques and Michel, 1997; Stambuk, 1998, 1999a,b; Loewenstern and Yianilos, 1997; Sengupta and Tompa, 2002; Kari et al., 1999). In addition to the application of coding theoretic methods to computational biology problems, researchers, such as Hubert Yockey who performed fundamental investigations of error correcting coding properties of genetic systems, have explored the error control coding properties of genetic sequences and systems (Yockey, 1992; Liebovitch et al., 1996; May et al., 1999, 2000; MacDonaill, 2002; Rosen and Moore, 2003). P. Bermel, D. Bitzer, M. Vouk, and E. Eni. (Bermel et al., unpublished) investigated table-based convolutional code models for Escherichia coli promotors. Based on the information content of the promoters, Bermel et al. approximated a 1/9 coding rate for the E. coli promoter and devised a 1/5 binary convolutional code model for the region. Beyond the work of Bermel et al. and May et al.'s investigation of block and convolutional code models for translation initiation, there is little known research into the development of channel coding models for genetic processes.

Informational analysis of genetic sequences has provided significant insight into parallels between the genetic process and information processing systems used in the field of communication engineering. Of particular interest are the results from Schneider et al. (Schneider, 1997; Schneider et al., 1986) and Eigen (Eigen, 1993). 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.

The first observation is mutations or errors are present within the genome of an organism. Analogous to an error-producing channel used by an engineering system to transmit information to a receiver, genetic processes such as replication can introduce errors into the genome of an organism. Mutations or variations in a genomic sequence can also be caused by external forces and can be passed down from parent to offspring. Some of these "errors" may be part of an organism's survival mechanism.

A second observation is that there exists sets of acceptable information strings or sequences that are functionally equivalent within a genetic system. For instance, ribosomal binding sites (translation initiation sites) appear to evolve to functional requirements rather than to genetic sequences that produce the strongest binding site (Schneider, 1997). Viable mutants, or imperfect sequences, have error rates near an error threshold assuring the organism's evolutionary flexibility (Eigen, 1993). Whether these variations are inherited or newly developed errors, genetic systems (macro-molecules, like the ribosome, that interact with nucleic acid sequences) still recognize a set of sequences that are similar but nonidentical. In a communication system, the decoder recognizes a set of similar but nonidentical group of information sequences or codewords. It will even recognize variations of this set, within the code's error detecting/correcting threshold. If survival and evolution of an organism necessitates errors, then, similar to an engineering communication system, there must exist a genetic error correction mechanism (Battail, 1997).

The error control mechanism employed by an engineering communication system is constructed using principles from the field of Coding Theory, specifically channel (or error-control) coding theory. Error control is accomplished by introducing redundancy into the original information sequence through a well-defined encoding algorithm (Sweeney, 1991; Lin and Costello, 1983; Dholakia, 1994). Similar to an error-control encoded information sequence, redundancy occurs naturally within RNA and DNA sequences (Lewin, 1995) in the form of tandem repeats and "extra" genomic information that in the

past was considered "junk DNA."

A final observation is that the ribosome maps, or decodes, a fixed length nucleic acid signal (codon) to specific information (amino acid). This parallels the behavior of a decoder in a communication system. For an (n, k) code, the decoder takes a n-symbol vector and maps it to k symbols of information, where k is less than n. From these observations, we can theorize that the transmission of genetic information can be viewed as a biological, cellular communication system that employs some method of coding to recognize valid information regions and to correct for "transmission" errors. Given that messenger RNA is viewed as a noisy encoded signal, the principal hypothesis of our work is that it is feasible to use principles of error control coding theory to interpret and model genetic regulatory processes such as the regulation of translation initiation. In addition to the translation initiation process, other areas where an error-control coding model would be applicable include the regulatory processes that govern DNA replication and the transcription of DNA into mRNA. As in translation, both processes are regulated by regions on the genome such as promoter and enhancer sequences.

In the following section, we give a brief overview of error-control coding methods, block coding specifically, and associated decoding methods. Section 3 discusses the relationship between the block coding method and the genetic process, and presents the methodology for forming a coding based genetic decoder. The preliminary results of applying the genetic decoder to *E. coli* K-12, *S. typhimurium* LT2, *B. subtilis*, and *S. aureus* Mu50 are presented in Section 4 and implications of the models are analyzed and discussed in Section 5. In the final section of this paper we discuss possible extensions to our research, based on the results of the block coding model for translation initiation.

2 Theoretical Background

The mathematics of coding is carried out using a set of discrete source symbols over a mathematical construct known as a finite field (Sweeney, 1991). In block encoding, a n-symbol encoded block at time i depends on the k-symbol information block at time i (Dholakia, 1994). Block codes are referred to as (n, k) codes. A codeword (or correct set of symbols) is the output of the block encoder for a given input data block (Sweeney, 1991).

2.1 Systematic Zero Parity Check Encoding

There are several ways to produce codewords from a k-symbol information sequence. A systematic code is a code which contains the k information symbols at the beginning of the codeword. The information symbols are then followed by $n-k$ parity symbols. Parity symbols are extra symbols added to the information sequence pre-transmission. Addition of redundancy via parity symbols aids in the detection and correction of errors that occur in the transmitted message. The value of the $n-k$ parity symbols is determined by the selected encoding method. In *binary* codes, where each symbol is a bit and can be represented as a 0 or 1, each of the final $n-k$ bits are set such that the parity bit is a linear combination of the information bits (Sweeney, 1991; Lin and Costello, 1983; Dholakia, 1994). The code is even if the modulo two sum of the information and parity bits is zero. The coding is odd otherwise.

A codeword is generated for every possible k-symbol information sequence. The codebook is the set of all codewords generated by the encoder. If a transmitted n-symbol sequence does not map to a codeword, we assume one or more symbols have been corrupted. The decoding task is to find the most likely changes in the received n-symbol sequence that will result in a valid codeword.

2.2 Minimum Distance Decoding

A decoder provides a strategy for selecting the transmitted codeword for a given received sequence. There are various decoding methods. One method, maximum likelihood decoding, compares the received sequence with every possible codeword sequence in the codebook and selects the most likely sequence. Decoding involves two steps. First the decoder checks whether the sequence corresponds to a codeword. A distance metric is used to determine how close the received sequence is to the codewords in the codebook. Second, if the decoder is an error correcting decoder, then it must identify the error pattern and use the error pattern to correct the received sequence to the most probable codeword transmitted. In this paper, we are only concerned with the decoder's ability to perform error detection.

The method tested in this work is called minimum distance decoding. The distance between codewords, $d(a, b)$ is the number of differences between codeword a and codeword b; this is called the Hamming distance (Sweeney, 1991). Although the Hamming distance is widely used in coding theory (and used in this work) there are other distance metrics for codes including the Lee and Triangular metrics (Duckworth, 1998). Use of the Lee or Triangular distance

measures should produce average results similar to those generated using the Hamming distance. Since the Lee and Triangular metrics define distance in a non-binary (match or non-match) sense, we suspect that the base five mapping of the RNA sequences will have a greater impact on Lee or Triangular metric-based decoding results than the current Hamming metric-based model.

For a received sequence r, the minimum Hamming distance, $dmin$ of r is the minimum of $d(r, S_c)$, where S_c is the codebook. In minimum distance decoding, we decode r to the codeword for which $d(r, S_c)$ is the least. If the minimum distance computation results in the same distance value for more than one codeword, although an error is detected, it is not correctable because of the degeneracy of the mapping. Minimizing the distance is the optimum decoding approach for a channel in which symbol errors are independent (memoryless channel) (Sweeney, 1991). The systematic-zero parity encoding concept and the minimum distance decoding concept are employed in our block coding model for the translation initiation system.

3 Methods

One does not know the exact mechanism employed by the genetic system to introduce redundancy or encode genetic information. By analyzing key elements involved in protein translation initiation, we hope to gain insight into a possible encoding and corresponding decoding model that quantitatively describes the behavior of the ribosome during translation initiation in prokaryotic organisms. The key biological elements considered in forming the coding model are: the 3' end of the 16S ribosomal RNA, the common features of bacterial ribosomal binding sites (such as the existence and location of the Shine-Dalgarno sequence), and RNA/DNA base-pairing principles.

3.1 Messenger RNA as a Block Encoded Sequence

If it is assumed that genetic information in DNA is encoded in a manner equivalent to block encoding, then the received message, the mRNA, can be viewed as a received parity sequence of a block encoded data stream. The RNA bases must be mapped to a numeric representation. The genetic coding alphabet must correspond to a finite field. In binary codes, the finite field consists of 0 and 1, modulo two addition and multiplication. For the genomic code, we know that four bases are found in mRNA: adenine, guanine, cytosine, and uracil. A fifth base, inosine, is found in transfer RNA (tRNA) which participates mainly in the elongation phase of translation. During elongation, inosine can wobble pair with more than one of the mRNA bases. These biological characteristics

are used to define an alphabet on the field of five. The RNA bases are mapped to the field of five as follows: Inosine(I) = 0, Adenine(A) = 1, Guanine(G) = 2, Cytosine(C) = 3, and Uracil(U) = 4. Multiplication and addition are modulo five operations (Bitzer et al., 1992). The RNA bases are mapped such that in modulo five addition the sum of bases that form hydrogen pairs is zero (as a simplification, we ignore all other chemical bonds that influence translation initiation). These definitions are used to construct the block code model for the the translation initiation process. Although multiplication is not used for the current block code model, it is used in our work on table-based convolutional code models for translation initiation (May, 2002; May et al., 2002).

3.2 The Genetic Codebook for Translation Initiation

In evaluating mRNA as a block encoded sequence, an (n,k) systematic zero parity code was developed based on the 16S ribosomal RNA. The reasoning behind the developed code is as follows. The 3' end of the 16S rRNA is directly involved in binding the messenger RNA during the initiation phase of protein translation (Lewin, 1995). We use the last thirteen bases of the 16S rRNA in forming our codewords. The last thirteen bases are used since the hexamer complementary to the Shine-Dalgarno sequence is found in this region of the 16S rRNA. The Shine-Dalgarno sequence is a series of nucleic acid bases on the 5' untranslated region of prokaryotic mRNA. The Shine-Dalgarno sequence helps attract the ribosome to the initiation site by forming Watson-Crick bonds with the 16S ribosomal RNA, part of the 30S ribosomal subunit (Watson et al., 1987; Lewin, 1995). Specifically, the last thirteen bases of the 16S rRNA that interact with the Shine-Dalgarno domain and other sequences on the 5' untranslated mRNA leader, are (Lewin, 1995):

$$3'AUUCCUCCACUAG...5' \tag{1}$$

Since our received sequence, the mRNA, contains the nucleotide sequence which base pairs with the 16S rRNA, we use the Watson-Crick complement of the thirteen base sequence in forming our codewords. The complement of the 3' end of the 16S rRNA is:

$$5'UAAGGAGGUGAUC...3' \tag{2}$$

We select our $n-k$ parity symbols from all (n-k)-base sub-sequences of the thirteen base complement in Equation 2. For instance, if we desire a (5,2) code, we would select our parity symbols from all three-base nucleotide sub-sequences of the thirteen base 16S complement. Table 1 shows these sub-sequences and their summation values. The three base parity sub-sequences

Table 1
Three-base Parity Bits derived from 16s rRNA.

Parity Bases	Sum of Parity Bases
U A A	1
A A G	4
A G G	0
G G A	0
G A G	0
G G U	3
G U G	3
U G A	2
G A U	2
A U C	3

are selected so that the following equation is satisfied:

$$\sum_{1}^{k} u_{genetic} + \sum_{1}^{n-k} ParityBases \;=\; 0 \tag{3}$$

where $u_{genetic}$ is the k-base information vector and $ParityBases$ is the $n-k$ base parity vector. To illustrate, if we define the information sequence as: $u_{genetic} = (C \;\; A)$. The numerical representation is $u = (3 \;\; 1)$. We select a set of parity symbols from Table 1 such that $u_1 + u_2 + \sum_1^3 ParityBases = 0$. Hence (U A A) is selected as our parity bases. The resulting codeword is: $Codeword = (3 \; 1 \; 4 \; 1 \; 1)$. The equivalent genetic codeword is: $Codeword_{genetic} = (C \;\; A \;\; U \;\; A \;\; A)$. We generate codewords for all possible k-base genetic information vectors. For a (5,2) code our information vectors would be drawn from every possible two-base RNA sequence; there are sixteen such sequences. A codeword is produced, as previously illustrated, for each possible two-base RNA sequence. If the resulting codeword satisfies Equation 3, then it is included in the codeword list (the codebook) otherwise it is excluded.

3.2.1 Multiple Codewords

In the preceding example, parity base selection is straightforward. Since (U A A) is the only three-base parity set that sums to one, it is the only choice. But, consider the next example: $u_{genetic} = (G \;\; A)$. The numerical representation is $u = (2 \;\; 1)$. The sum of the parity bases needs to be two in order to achieve a zero parity codeword. From Table 1 we can choose either (U G A) or (G A

U). Therefore, there are two possible codewords: $Codeword^1 = (2\ 1\ 4\ 2\ 1)$ or $Codeword^2 = (2\ 1\ 2\ 1\ 4)$. The equivalent genetic codewords are: $Codeword^1_{genetic} = (G\ A\ U\ G\ A)$ and $Codeword^2_{genetic} = (G\ A\ G\ A\ U)$. Since there are two codewords, we must decide which codeword to select. In order to make this decision, let us consider the protein translation model.

Several factors influence translation of mRNA sequences, including: initiation codon, presence and location of the Shine-Dalgarno sequence, spacing between the initiation codon and the Shine-Dalgarno domain, the second codon following the initiator codon, and possibly other nucleotides in the -20 to +13 region (including the non-random domain) of the mRNA leader region (Gold and Stormo, 1987). The exact bases of the 3' end of the 16S rRNA to which the mRNA leader region binds are not always fixed. What is known is that during translation initiation, the last thirteen bases of the 16S rRNA bind to the mRNA leader sequence in a region known as the ribosome binding site (Lewin, 1995; Watson et al., 1987). Several bases on the exposed part of the 16S rRNA are candidates for binding with the mRNA. To model this biological possibility, we include both codewords, (G A U G A) and (G A G A U), in our list of valid codewords.

3.3 Minimum Distance Decoder for Model Verification

A minimum Hamming distance decoder, based on the systematic, zero-parity check encoding methodology, was designed to verify the block coding model. The analysis sequence is composed of: the thirty bases of the mRNA leader sequence preceding the initiation (start) signal, the initiation signal (usually AUG), and twenty-seven bases from the coding region immediately following the initiation signal:

$$[b_{-30}\ \ b_{-29}\ \ ...\ \ b_{-1}\ \ A\ \ U\ \ G\ \ b_{+3}\ \ ...\ \ b_{+29}] \tag{4}$$

Bases numbered -30 to -1 are part of the leader region of the mRNA. Bases numbered +3 to +29 are part of the coding region of the mRNA. The A of the AUG initiation signal is position zero in the sequence. The analysis sequence is the portion of the mRNA that we evaluate in order to determine any significant differences between the coding characteristics of translated and non-translated mRNA sequences. The ribosome covers approximately thirty bases of the mRNA at a time (Lewin, 1995). Therefore, a sixty base analysis sequence should be sufficient in representing the region of the mRNA that interacts with the small subunit of the ribosome during the initiation process.

The received sequence is an n-base subset of the analysis sequence. For instance, assuming a (5,2) code, the first two received sequences would be:

$$r_{-30} = [b_{-30} \;\; b_{-29} \;\; b_{-28} \;\; b_{-27} \;\; b_{-26}] \tag{5}$$

$$r_{-29} = [b_{-29} \;\; b_{-28} \;\; b_{-27} \;\; b_{-26} \;\; b_{-25}] \tag{6}$$

The received sequence is referenced with respect to its positional distance from the first base of the start signal. The minimum Hamming distance of a received sequence is defined as:

$$dmin_p = min[d(r_p, S_c)] \tag{7}$$

where p is position relative to the initiation codon and $S(c$ represents the codebook, the set of all codewords in our codeword list.

The decoding process normally corrects the received sequence to the codeword with the lowest minimum distance value and recovers the transmitted information sequence, u. Since our objective is to analyze the coding model, the minimum distance is recorded for each received sequence in the analysis stream. This distance is used to evaluate how well the block coding model captures the biological aspects of the initiation process. The *E. coli* K-12 strain MG1655 sequence data was used to test a (5,2) and (8,2) block code model for prokaryotic translation initiation.

4 Results

The complete *E. coli* K-12 genome file, *ecoli.gbk*, was downloaded from the NIH ftp site (ncbi.nlm.nih.gov). Using a PERL script program, the following information was extracted from the GenBank files: 1) Complete DNA sequence of *E. coli* K-12 strain MG1655, accession number U000096; 2)Location of each known gene in the coding strand; 3)Location of each known gene in the complement strand. Using the information in the GenBank file, we separated known and possible reading frames into three data groups:

Translated Sequences (2,917 sequences): Open reading frames which GenBank indicates as sequences that translate into protein.

Hypothetical Sequences (1,372 sequences): Open reading frames which GenBank indicates as hypothetically translated sequences. According to the GenBank documentation, GeneMark software was used to predict open reading frames.

Non-Translated Sequences (21,039 sequences): Open reading frames which do not appear on the list of translating or hypothetically translating sequences in the GenBank genome file. 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 GenBank sequence is the DNA sequence of the *E. coli* K-12. The prokaryotic mRNA transcript does not undergo modification, unlike eukaryotic mRNAs, prior to translation by the ribosome. Therefore we can use the DNA sequence of the E.coli as our mRNA sequence. In our experiment, uracil (U) replaces thymine (T) in the GenBank sequences since uracil appears in RNA.

4.1 Analysis Method

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

$$\begin{bmatrix} dmin_{-30}^{1} & dmin_{-29}^{1} & ... & dmin_{PosValid}^{1} \\ \vdots & \vdots & \vdots & \vdots \\ dmin_{-30}^{nseq} & dmin_{-29}^{nseq} & ... & dmin_{PosValid}^{nseq} \end{bmatrix} \tag{8}$$

where $nseq$ is the number of analysis sequences in the group and $PosValid$ is the last valid comparison (or decoding) position on the sequence. For a (n, k) code,

$$PosValid \;=\; InitPos + NumValid - 1 \tag{9}$$

and

$$NumValid \;=\; SeqLength - n + 1 \tag{10}$$

In this work, $InitPos = -30$ and $SeqLength = 60$; therefore $NumValid = 56$ for the (5,2) block code model and $NumValid = 53$ for the (8,2) model. The corresponding values for $PosValid$ are $PosValid = +25$ and $PosValid = +22$, respectively.

To extract the information signal from the noise contained in our model, we take the average $dmin$ value by position for each sequence group. This produces one signal, $dmin_{SequenceGroup}$, that describes the minimum distance char-

acteristic of each sequence group, where

$$dmin_{(SequenceGroup)}(p) = \frac{1}{nseq} \sum_{i=1}^{nseq} dmin^{i}_{SequenceGroup}(p) \tag{11}$$

The range for the index p is $p = -30..PosValid$, where p represents position relative to the initiation codon. Averaging 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 $dmin_{SequenceGroup}$ vector indicate stronger hydrogen bond formations between the 16S ribosomal RNA and the messenger RNA.

4.2 (5,2) and (8,2) Block Code Model

For the (5,2) code, there were a total of thirty-three codewords. In this model, for each two base information sequence a three base parity sub-sequence was selected. Figure 1 shows the resulting mean or average minimum Hamming distance by position for the (5,2) block code model. The horizontal axis is the

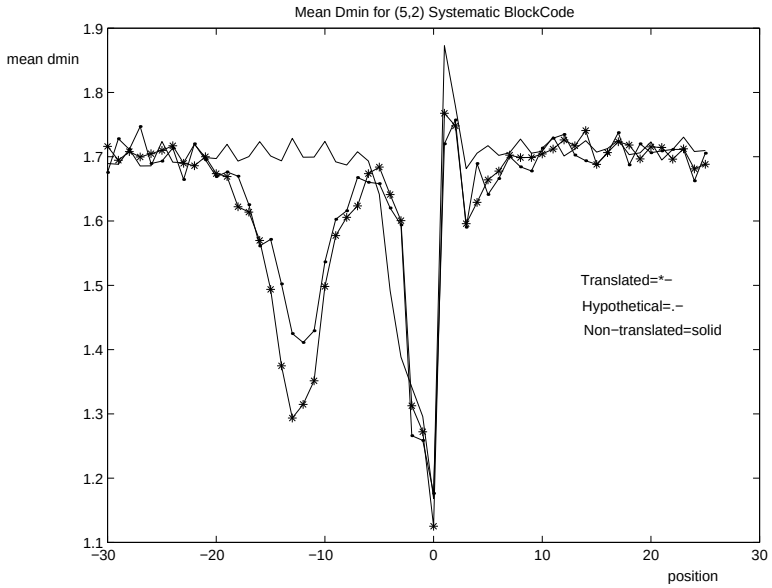

Fig. 1. Results of Minimum Distance Block Decoding Model for (5,2) Code

position relative to the first base of the initiation codon. Zero on the horizontal axis corresponds to the alignment of the first base of a codeword with 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 three sequence data groups (translated sequences, hypothetical translated sequences, and non-translated sequences).

For the (8,2) code, there were a total of twenty-six codewords. For each two base information sequence a six base parity sub-sequence was selected. Figure 2 shows the resulting mean minimum Hamming distance by position for the (8,2) block code model. As in Figure 1 the horizontal axis corresponds

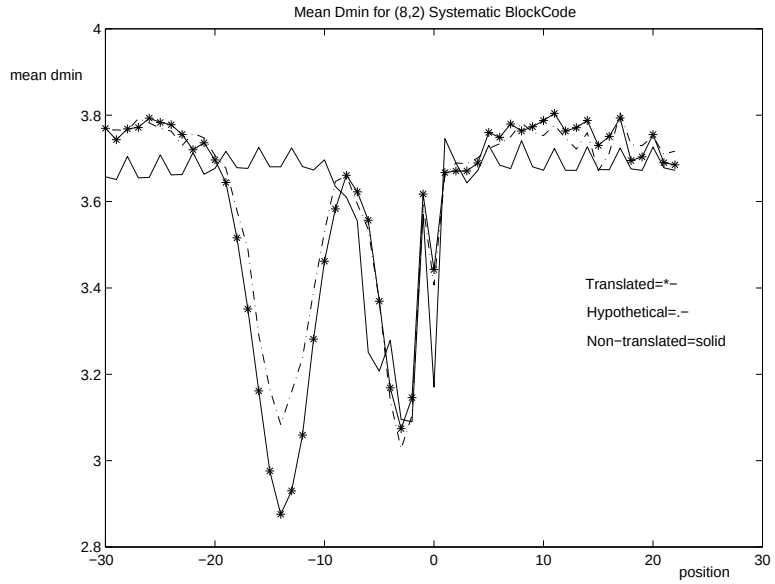

Fig. 2. Results of Minimum Distance Block Decoding Model for (8,2) Code

to the position relative to the first base of the start codon, which is position zero. The vertical axis is the mean of the aligned minimum Hamming distance values of the leader sequences in each of the three sequence groups.

4.3 Application of Block Code Model to Non-model Prokaryotic Organisms

The block coding models produced using the *E. coli* K-12 genome were tested on data sets from three prokaryotic organisms: *Salmonella typhimurium* LT2, *Bacillus subtilis*, and *Staphylococcus aureus* Mu50. *E. coli* and *S. typhimurium* share a common taxonomical lineage as do *B. subtilis* and *S. aureus*. The sequence data used for testing the model on prokaryotic organisms were compiled and processed using the web-based GenBank Information Retrieval Tool developed by Cheng and Chandra of the North Carolina State University Scientific Data Management Center (Chandra, 2002).

Figure 3 shows the resulting mean minimum distance by position for the (5,2) block code model. Figure 4 shows the resulting mean minimum distance by position for the (8,2) block code model. The horizontal and vertical axes for Figure 3 and Figure 4 correspond, as in previous figures, to position and average minimum Hamming distance values, respectively.

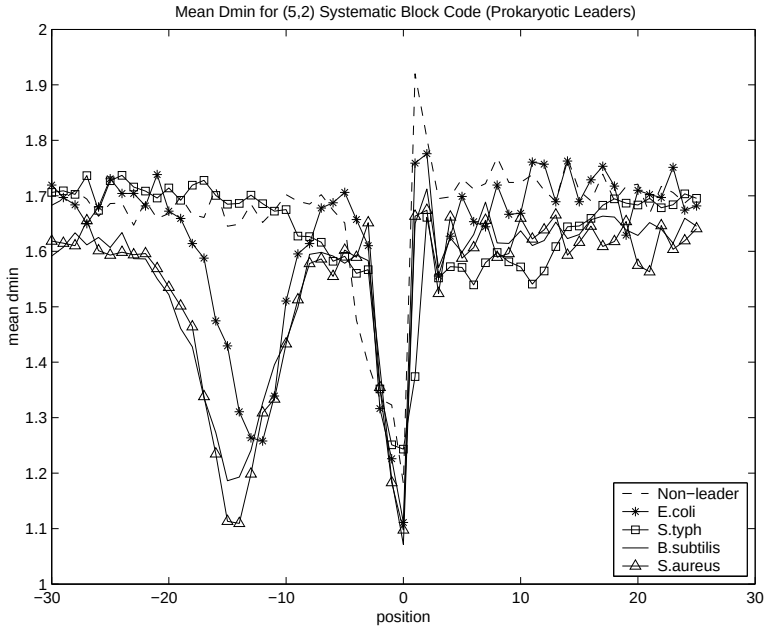

Fig. 3. Results of Minimum Distance Block Decoding Model for (5,2) Code for Prokaryotic Data Set

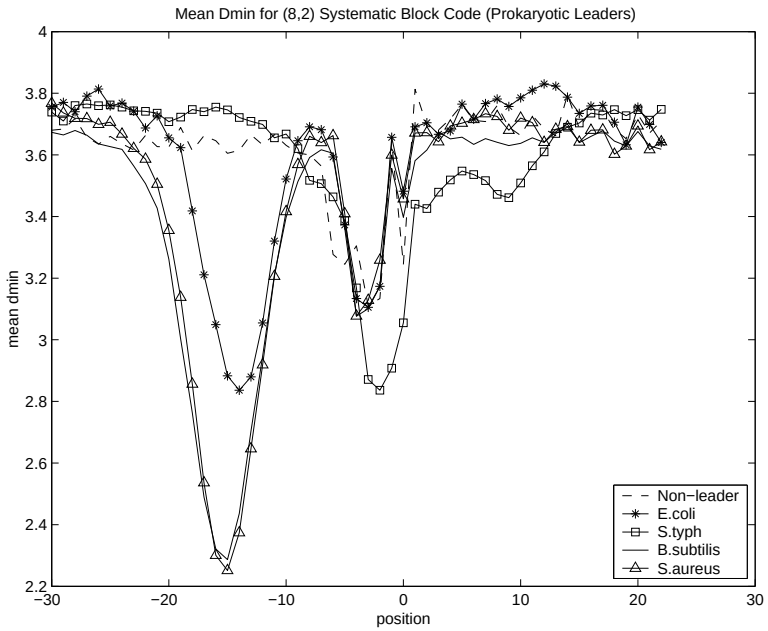

Fig. 4. Results of Minimum Distance Block Decoding Model for (8,2) Code for Prokaryotic Data Set

5 Discussion

Three criteria were used to analyze the effectiveness of the block code model: 1)Distinction between translated and non-translated sequence groups; 2)Indication and recognition of the open reading frame construct; 3)Recognition of

biologically significant regions within the mRNA leader sequence. The results of the (5,2) and (8,2) block code models show a significant difference between the translated, hypothetical and the non-translated group. As Figure 1 and Figure 2 illustrate the -15 to 0 region (-20 to 0 for the (8,2) model) contains large synchronization signals which can be used to distinguish between translated and non-translated sequence groups (Criteria 1). Although the current models do not produce strong evidence of frame synchronization patterns within the protein coding region, they do indicate a strong synchronization pattern at the initiation codon (Criteria 2). Both the (5,2) and (8,2) block code model recognize biologically important regions on the mRNA leader (Criteria 3). A minimum distance trough occurs between the -15 and -10 regions for both models. These regions contain the non-random domain and the Shine-Dalgarno domain, key regions in the translation initiation process (Gold and Stormo, 1987).

5.1 Response of Prokaryotic Organisms to Block Code Model

The behavior of *B. subtilis* and *S. aureus* are similar to the behavior of *E. coli*. *B. subtilis* and *S. aureus* have larger average *dmin* values than the *E. coli* model organism for both the (5,2) and (8,2) models. They differ significantly from the non-leader (non-translated) sequence groups and recognize biologically significant regions such as the Shine-Dalgarno domain, the non-random domain, and the initiation codon. Surprisingly, *S. typhimurium*, the most taxonomically related to *E. coli*, differed significantly from *E. coli* and the other prokaryotic organisms. In fact in both the (5,2) and (8,2) models, *S. typhimurium* behaves more like the non-leader sequence group, reaching its global minimum within the protein coding region of the mRNA analysis sequence. The cause and implication of this difference in behavior continue to be investigated.

6 Conclusion

Both the (5,2) and (8,2) models distinguish translated sequence groups and non-translated sequence groups. They both also indicate the existence of key regions within the mRNA leader sequence. 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 (Gold and Stormo, 1987). We were able to successfully apply the model to other prokaryotic organisms, including *B. subtilis* and *S. aureus*. The unexpected response of *S. typhimurium* to the model raises interesting questions regard-

ing the genetic, taxonomical relatedness of *E. coli* and *S. typhimurium* that warrant further computational and biological investigation.

The results of our work suggest that it is possible to design a coding based heuristic for distinguishing between protein coding and non-protein coding genomic sequences by "decoding" the mRNA leader region. The block code model used in this work is a sliding block code. We evaluated overlapping information, hence mimicking a convolutional code. A convolutional code more accurately depicts the behavior of the ribosome as a decoder that incorporates memory in its translation (or decoding) decisions. In our current work we explore this memory based coding model for genetic regulatory processes. We investigate evolutionary computing and algebraic methods for constructing coding models of translation initiation sites and for analyzing general EC coding properties of regulatory sequences.

The success of this work can lead to the development of improved methods for identifying the precise location of translation initiation start sites, an area that is receiving greater computational research attention (Frishman et al., 1999; Tompa, 1999; Hannenhalli et al., 1999; Suzek et al., 2001; Walker et al., 2002; Zien et al., 2000; Yada et al., 2001; Besemer et al., 2001). 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.

Acknowledgments

This work was supported in part by a National Science Foundation Graduate Fellowship and the Ford Foundation Dissertation Fellowship for Minorities.

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

References

Almagor, H., 1985. Nucleotide distribution and the recognition of coding regions in DNA sequences: an information theory approach. Journal of Theoretical Biology 117, 127–136.

Altschul, S. F., 1991. Amino Acid substitution matrices from an information theoretic perspective. Journal of Molecular Biology 219, 555–565.

Arques, D. G., Michel, C. J., 1997. A code in the protein coding genes. BioSystems 44, 107–134.

Battail, G., November 1997. Does information theory explain biological evolution? Europhysics Letters 40 (3), 343–348.

Besemer, J., Lomsadze, A., Borodovsky, M., 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 29 (12), 2607–2618.

Bitzer, D. L., Vouk, M. A., Dholakia, A., 1992. Genetic Coding Considered as a Convolutional Code, north Carolina State University, Raleigh.

Chandra, S., December 2002. Service-based Support for Scientific Workflows. Master's thesis, NCSU.

DeLaVega, F. M., Cerpa, C., Guarneros, G., 1996. A mutual information analysis of tRNA sequence and modification patterns distinctive of species and phylogenetic domain. In: Pacific Symposium on Biocomputing. pp. 710–711.

Dholakia, A., 1994. Introduction to Convolutional Codes with Applications. Kluwer Academic Publishers, Norwell, Massachusetts.

Duckworth, W. M., 1998. Code, Designs, and Distance. Ph.D. thesis, University of North Carolina - Chapel Hill, Chapel Hill, NC.

Eigen, M., 1993. The origin of genetic information: viruses as models . Gene 135, 37–47.

Fowler, T. B., 1979. Computation as a thermodynamic process applied to biological systems . International Journal of Bio-Medical Computing 10 (6), 477–489.

Frishman, D., Mironov, A., Gelfand, M., 1999. Starts of bacterial genes: estimating the reliability of computer predictions. Gene 234 (2), 257–65.

Gold, L., Stormo, G., 1987. Translational Initiation. In: Escherichia coli and Salmonella typhimurium, Cellular and Molecular Biology. pp. 1302–1307.

Golomb, S. W., 1962. Efficient coding for the desoxyribonucleic channel. Proc. of Symposia in Applied Mathematics 14, 87–100.

Hannenhalli, S. S., Hayes, W. S., Hatzigeorgiou, A. G., Fickett, J. W., 1999. Bacterial start site prediction. Nucleic Acids Res. 27 (17), 3577–3582.

Hayes, B., 1998. The Invention of the Genetic Code. American Scientist 86 (1), 8–14.

Kari, L., Kari, J., Landweber, L. F., 1999. Reversible molecular computation in ciliates. In: Jewels are Forever. pp. 353–363.

Lewin, B., 1995. Genes V. Oxford University Press, New York, NY.

Liebovitch, L. S., Tao, Y., Todorov, A., Levine, L., 1996. Is there an Error

Correcting Code in DNA? Biophysical Journal 71, 1539–1544.
Lin, S., Costello, D. J., 1983. Error Control Coding: Fundamentals and Applications. Prentice-Hall, Inc., Englewood Cliffs, NJ.
Loewenstern, D., Yianilos, P. N., 1997. Significantly lower entropy estimates for natural DNA sequences. In: Proceedings of the Data Compression Conference.
MacDonaill, D., 2002. A Parity Code Interpretation of Nucleotide Alphabet Composition. Chem Communic , 2062–2063.
May, E. E., May 2002. Analysis of Coding Theory Based Models for Initiating Protein Translation in Prokaryotic Organisms. Ph.D. thesis, North Carolina State University, Raleigh, NC.
May, E. E., Vouk, M. A., Bitzer, D. L., Rosnick, D. I., 1999. Coding Model for Translation in E. coli K-12 . In: First Joint Conference of EMBS-BMES.
May, E. E., Vouk, M. A., Bitzer, D. L., Rosnick, D. I., 2000. The Ribosome as a Table-Driven Convolutional Decoder for the Escherichia coli K-12 Translation Initiation System . In: World Congress on Medical Physics and Biomedical Engineering Conference.
May, E. E., Vouk, M. A., Bitzer, D. L., Rosnick, D. I., 2002. Constructing Optimal Convolutional Code Models for Prokaryotic Translation Initiation. In: Second Joint EMBS-BMES Conference 2002.
Oliver, J. L., Bernaola-Galvan, P., Guerrero-Garcia, J., Roman-Roldan, R., 1993. Entropic profiles of DNA sequences through chaos-game-derived images . Journal of Theoretical Biology 160, 457–470.
Palaniappan, K., Jernigan, M. E., 1984. Pattern analysis of biological sequences . In: Proceedings of the 1984 IEEE International Conference on Systems, Man, and Cybernetics.
Pavesi, A., Iaco, B. D., Granero, M. I., Porati, A., 1997. On the informational content of overlapping genes in prokaryotic and eukaryotic viruses . Journal of Molecular Evolution 44 (6), 625–631.
Roman-Roldan, R., Bernaola-Galvan, P., Oliver, J. L., 1996. Application of information theory to DNA sequence analysis: a review. Pattern Recognition 29 (7), 1187–1194.
Rosen, G., Moore, J., 2003. Investigation of Coding Structure in DNA. In: ICASSP 2003.
Salamon, P., Konopka, A. K., 1992. A maximum entropy principle for the distribution of local complexity in naturally occuring nucleotide sequences. Computers and Chemistry 16 (2), 117–124.
Sarkar, R., Roy, A. B., Sarkar, P. K., 1978. Topological Information Content of Genetic Molecules – I. Mathematical Biosciences 39, 299–312.
Schneider, T. D., 1991a. Theory of Molecular Machines. I. Channel Capacity of Molecular Machines. Journal of Theoretical Biology 148, 83–123.
Schneider, T. D., 1991b. Theory of Molecular Machines. II. Energy Dissipation from Molecular Machines. Journal of Theoretical Biology 148, 125–137.
Schneider, T. D., 1997. Information content of individual genetic sequences. Journal of Theoretical Biology 189, 427–441.

Schneider, T. D., 1999. Measuring molecular information. Journal of Theoretical Biology 201, 87–92.

Schneider, T. D., Mastronarde, D. N., 1996. Fast multiple alignment of ungapped DNA sequences using information theory and a relaxation method. Discrete Applied Mathematics 71, 259–268.

Schneider, T. D., Stormo, G. D., Gold, L., Dhrenfeucht, A., 1986. Information Content of Binding Sites on Nucelotide Sequences. Journal of Molecular Biology 188, 415–431.

Sengupta, R., Tompa, M., 2002. Quality Control in Manufacturing Oligo Arrays: A Combinatorial Design Approach. Journal of Computational Biology 9 (1), 1–22.

Stambuk, N., 1998. On the genetic origin of complementary protein coding. Croatica Chemica ACTA 71 (3), 573–589.

Stambuk, N., 1999a. On circular coding properties of gene and protein sequences. Croatica Chemica ACTA 72 (4), 999–1008.

Stambuk, N., 1999b. Symbolic Cantor Algorithm (SCA): A method for analysis of gene and protein coding. Periodicum Biologorum 101 (4), 355–361.

Strait, B. J., Dewey, T. G., 1996. The Shannon information entropy of protein sequences . Biophysical Journal 71, 148–155.

Suzek, B. E., Ermolaeva, M. D., Schreiber, M., Salzberg, S. L., 2001. A probabilistic method for identifying start codons in bacterial genomes. Bioinformatics 17 (12), 1123–1130.

Sweeney, P., 1991. Error Control Coding an Introduction. Prentice Hall, New York, NY.

Tompa, M., 1999. An Exact Method for Finding Short Motifs in Sequences, with Application to the Ribosome Binding Site Problem. In: ISMB 1999.

Walker, M., Pavlovic, V., Kasif, S., 2002. A comparative genomic method for computational identification of prokaryotic translation initiation sites. Nucleic Acids Res. 30 (14), 3181–3191.

Watson, J., Hopkins, N., Roberts, J., Steitz, J., Weiner, A., 1987. Molecular Biology of the Gene. The Benjamin Cummings Publishing Company, Inc., Menlo Park, CA.

Yada, T., Totoki, Y., Takagi, T., Nakai, K., 2001. A novel bacterial gene-finding system with improved accuracy in locating start Codons. DNA Res. 8 (3), 97–106.

Yockey, H., 1992. Information Theory and Molecular Biology. Cambridge University Press, NY, NY.

Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T., Muller, K. R., 2000. Engineering support vector machine kernels that recognize translation initiation sites. Bioinformatics 16 (9), 799–807.