Proceedings of the 25th Annual International Conference of the IEEE EMBS
Cancun, Mexico • September 17-21, 2003

On finding convolutional code generators for translation initiation of Escherichia Coli K-12

L. Ponnala[1], D. L. Bitzer[2], M. A. Vouk[2]
[1]Department of Electrical and Computer Engineering, North Carolina State University, NC, USA
[2]Department of Computer Science, North Carolina State University, NC, USA

Abstract— **Using error-control coding theory, the translation of mRNA into amino acids can be functionally paralleled to the decoding of noisy, convolutionally encoded data parity streams. The ribosome is modeled as a table-based convolutional decoder. This work attempts to find plausible convolutional code generators for the Escherichia Coli K-12 translation initiation. The g-mask is chosen from the exposed part of the 16s rRNA. The generators are calculated from the g-mask, using an algorithmic approach. The most plausible generators are chosen based on their ability to produce encoded sequences which provide a clear distinction between the translated and non-translated sequences.**

Keywords— **E.Coli, mRNA translation, convolutional code model.**

I. Introduction

Viewing protein synthesis as an information-processing system allows nucleotide sequences to be analyzed as messages [1]. May's model defines a genetic channel as the DNA replication, transcription and translation process during which errors may be introduced, detected and possibly corrected [3]. The transfer of biological information can be modeled as a communication channel, with the DNA sequence as input, and the polypeptide amino acid sequence as the output. As in error-protected information channels, redundancy occurs within genomic sequences. The DNA encoded message, in its double-helix form, is doubly redundant. Hence, we could consider un-replicated DNA as a half-rate systematic code, i.e., for each item (base) there is a corresponding parity item (complementary base). At the nucleic-acid level, messenger RNA (mRNA) sequences are mapped to amino-acids by grouping three nucleic acid bases together to form a codon. A codon, or three base nucleic acid vector, is mapped to a single amino acid information item. This process which occurs during translation, could be viewed as the decoding of a rate one-third code, i.e., each amino acid is encoded using three "parity" items, or it could be viewed as a code of different (non-systematic) rate depending on how much error detection capability is assumed to be present.

It is also known that leader regions of the messenger RNA (mRNA), and other prokaryotic regulatory regions contain consensus sequences which in some way signal or control translation. Examples are Shine-Dalgarno (SD) sequence in bacterial mRNA, and the TATA box and the Pribnow box in double helix DNA [5]. Specifically, the SD sequence always occurs before the start codon (usually AUG) of a translating mRNA sequence in E-coli. Our hypothesis is that, if there is a method in place that checks for the validity of the leader sequence (which includes SD), the ribosome would somehow have a way of recognizing it. Assuming there is a validating relationship among the leader sequence bases and/or codons, and assuming that the ribosome has an exposed region which is in contact with the mRNA leader for validation purposes, one may conjecture that the leader sequence may have embedded in it (or may be modeled as) an error detecting code such as a block code or a convolutional code. In Escherichia-Coli (E-coli), the exposed part of the 16s ribosomal RNA (rRNA) binds with the mRNA leader sequence during the initiation of translation. The same exposed part appears to remain in contact with mRNA (in addition to P and A sites and possibly other ribosomal regions) during the translation process.

The specific form of coding theory we apply in the present analysis is called table-driven convolutional coding. A brief overview of table-based convolutional coding [2],[9] is given in the Appendix. The mathematics of coding is carried out over a finite field, also referred to as Galois Field (GF), using a set of discrete source symbols [8]. The bases are mapped to the field of five, as follows: Inosine (I) = 0, Adenine (A) = 1, Guanine (G) = 2, Cytosine (C) = 3, Uracil (U) = 4. This calls for arithmetic operations such as addition and multiplication to be carried out in GF(5). The assignment of numerical values to the nucleotides is consistent with the bonding characteristics of the base pairs. In digital communication theory, checking the validity of an encoded message corresponds to matching of the message, piecewise, to a syndrome former (a sequence of symbols also known as the g-mask). A correct message is expected to yield zero syndrome when matched against the syndrome former. The syndrome former symbols (which will be the same symbols used to encode the sequence, A, U, G , C and I in our case) are "multiplied" (in this case base five) with corresponding portion of the encoded sequence, and these products are then "added" (base five) to form one symbol of the syndrome. The syndrome former is then moved downstream by a code-specific distance and the process is repeated. If the resulting syndrome is zero, the message is assumed to be error-free. However, the syndrome can contain any of the 5 operational symbols. Also, it must be noted that if the number of "errors" exceeds the error-detecting capability of the code, the message may be flagged as correct even when it is not. It is interesting to note that

0-7803-7789-3/03/$17.00 ©2003 IEEE

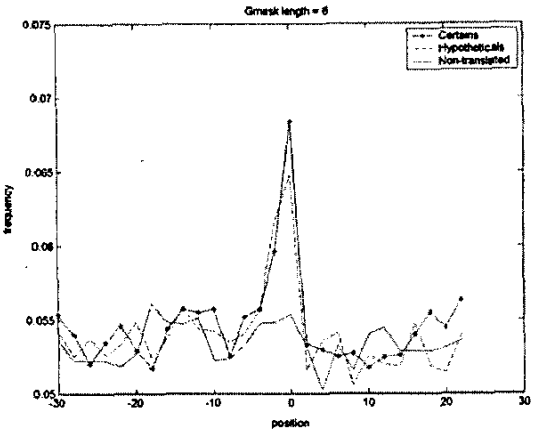

Fig. 1. Frequency of most-occurring 2-symbol distance patterns, by position in leader sequence

this reminds us very much of what happens in the case of cancers where erroneous genetic code is allowed to translate. To demonstrate the validity of the convolutional code hypothesis, the results of sliding a g-mask of length 6 over the mRNA leader sequences are shown in Fig. 1.

The genetic g-mask is formed from subsets of contiguous bases on the exposed part of the 16s rRNA. There are 8 possible g-masks of length 6 that can be formed from the 13-base long exposed part. Each g-mask is applied on three different sets of sequences: "certains" or translated sequences, "hypotheticals" or hypothetically-translated sequences and non-translated sequences. At every position, we measure how well a g-mask decodes the leader sequence at that position by calculating the syndrome at that position for all participating sequences (1000 in our case). Then we consider groups of adjacent syndrome values as a "distance pattern" [4]. And finally, at each position, we find the most frequent "distance pattern" and compute its relative frequency. If the pattern were truly random we might expect that relative frequency to be around 0.04. In Figure 1, relative frequency is shown on the vertical axis. The larger the frequency, the more the g-mask favors the sequence. The horizontal axis indicates position in the mRNA leader sequence, with zero corresponding to the location of the first base of the start codon (usually AUG). A clear peak is observed around position zero in the translated and hypothetically-translated sequences, the peak being more prominent in the former case. This indicates that the convolutional coding model may have the ability to distinguish between the three types of sequences

In order to fully understand the implications of the model based on the convolutional coding theory, we need to be able to back-track from the syndrome former (which appears to be physically expressed as SD sequence in the case of bacteria), to the codes that actually can produce that syndrome former. In general, this is a many to one relationship. We would expect that to be the case since the DNA information engine may employ errors to control translation efficiency and to "lock-in" into the translation frame [11]. We do assume that in the present study, the most important feature of a good code is its ability to distinguish the translated sequences from the hypothetical and non-translated sequences.

Genetic algorithms have been used with considerable success in constructing convolutional code models for translation initiation [6]. The optimal codes possess the following features: high similarity of the g-mask to the 3' end of the 16s rRNA, ability of the g-mask to recognize key regions on the mRNA leader such as the non-random and Shine-Dalgarno domains, and potential to detect valid and invalid leader sequences. The genetic algorithms-based method searches the space of all possible codes of a given description, to find optimal generators. In this work, instead of using search-based methods, we apply analytical methods to find good generators.

II. METHODOLOGY

We now describe a numerical algorithm for finding the generators of a convolutional code, when the g-mask is known. The algorithm is based on concepts of matrix theory and linear algebra, and is easy to compute.

The syndrome former and the generators of a convolutional code have the following relationship:

$$\mathbf{Ca} = \mathbf{0} \qquad (1)$$

C is the vector containing the generator coefficients, and **a** is the g-mask. An example of forming the above equation is described in the appendix. For a catastrophic code generator, **C** is not invertible [10]. This serves as a test for catastrophicity, and will hence be referred to as the "rank test". If r = k/n is the rate of the code, and L is the length of each generator, number of generator coefficients is (nL).

Equation (1) may be re-arranged in the form

$$\mathbf{Ac} = \mathbf{0} \qquad (2)$$

The matrix **A** now contains elements of the g-mask and **c** contains the generator coefficients. Given the g-mask coefficients, (2) may be solved to find the generator coefficients. There are (w+k) equations in (nL) unknowns. The number of "basis" solutions $\mathbf{g}_k$ is N = nL - (w+k). Since the coding process is carried out in GF(5), any linear combination of these N vectors, may be treated as a possible solution. A possible generator may be expressed as

$$\mathbf{p}_i = \sum_{k=1}^{N} \alpha_k \mathbf{g}_k , \qquad \alpha_k \in GF(5) \qquad (3)$$

The set of all possible generators will have 5^N-1 elements. Thus the search-space of possible generators grows

exponentially with the number of generators and the code constraint length.

For example, a rate 1/3 code having L = 5, would yield w = 6, and N = 8. The search-space of possible generators has 5^8-1 elements, using the above method. In comparison, if we were to search the space of all possible codes having rate 1/3 and L = 5, we would have 5^{15}-1 elements. Thus, the algorithmic approach described above narrows-down the search-space, since the g-mask is known beforehand.

III. Results

For this analysis, we use the GeneBank dataset for the Escherichia coli K-12 strain MG1655 [12]. The sequence was parsed, and the translated, hypothetically-translated and non-translated sequences were extracted. The number of sequences in each group was selected to be 1000.

We now try to find a code model having the following parameters: Rate = ½, L = 3, g-mask length = 6. The number of possible generators that can be obtained using a single g-mask is 4. There are 8 possible g-masks of length 6, but 2 of them are non-invertible. So, there result 24 possible generators. We require a generator that is non-catastrophic. So each possible solution is first examined, using the rank test for catastrophicity. The first non-catastrophic generator is retained for further analysis. At the end of this step, there remain 6 generators, one from each g-mask. It is worth noting that all the generators that are derived from a particular g-mask exhibit the same behavior when their parity streams are decoded. In other words, the syndrome-distance pattern obtained is identical for all generators derived from the same g-mask. It is therefore justified that selecting only one non-catasrophic generator per g-mask is a viable approach.

The set of translated, hypothetically-translated and non-translated sequences are used to produce the corresponding encoded parity streams. The g-masks are applied on all the obtained encoded data, and the average relative frequencies (referred to as just "frequency") of 2-symbol distance patterns is calculated at each position in the parity stream. It is found that the region -12 to -2 shows a good distinction between the three types of sequences (Fig. 2). This is in agreement with the biology of the system, since the region before the start codon is known to contain information that distinguishes the coding from non-coding sequences [5].

A plausible measure of fitness of each generator would be its ability to produce encoded sequences, that show a clear distinction in parity, based on whether they are coding or non-coding. In accordance with this, two possible fitness values can be assigned to each generator: one based on the ratio of translated to hypothetically-translated peak frequencies (Fitness1) and another based on the ratio of translated to non-translated peak frequencies (Fitness2), in

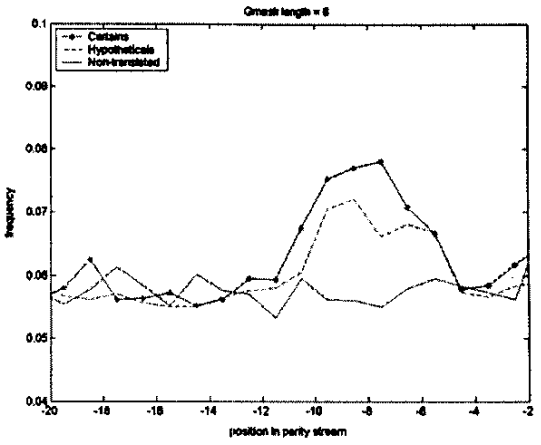

Fig. 2. Frequency of most-occurring 2-symbol distance patterns by position in parity stream, for rate ½ code

the range -12 to -2. An example showing the calculation of each fitness ratio follows. Let us consider the fifth generator. For the translated sequences, it has a peak frequency of 0.0786, and for the hypothetically-translated sequences, the peak frequency is 0.0781. The ratio of the two (i.e., Fitness1) is about 1, which is not unexpected. Using non-translated sequences, a peak frequency of 0.0559 is obtained. The ratio of the peak frequency of the translated sequences to the peak frequency of non-translated sequences (i.e., Fitness2) is about 1.4. If the distribution of the syndrome distance patterns is assumed to be random, an average relative frequency of 0.04 should result for non-translated sequences. The fitness graphs for the two cases are depicted in Fig. 3. The horizontal axis identifies each selected generator.

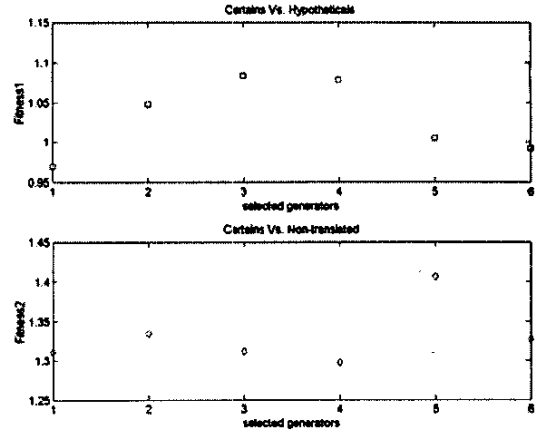

Fig. 3. Fitness of chosen generators, for rate ½ code

IV. DISCUSSION

It is observed from Fig. 3 that a generator that has a high value of Fitness1, does not necessarily have a high value of Fitness2. Generator 5, for instance, has a relatively low value of Fitness1, which means that it does not perform very well in distinguishing translated from hypothetically-translated sequences. But, the same generator has the highest value of Fitness2. In that sense, it can be considered optimal with regard to the distinction between translated and non-translated sequences.

Our current research is focused on determining whether the convolutional code generator has a meaning in the biological sense. For instance, a convolutional code generator could represent an enzyme, which recognizes specific patterns in DNA chains and repairs or breaks the DNA at those locations.

V. CONCLUSION

A sequence-based model for prokaryotic translation initiation has been presented using the theory of convolutional codes. We have devised a novel method for finding the generators of the convolutional code model, using table-based coding techniques. The performance of each generator has been evaluated based on its ability to produce a clear distinction between translated, hypothetically-translated and non-translated sequences. More efficient g-masks would produce consistent syndrome patterns, and could be constructed using techniques such as binding vector analysis [6]; research into this continues. Generators determined using such g-masks would represent the translation initiation process with greater accuracy.

REFERENCES

[1] Ramon Roman-Roldan, Pedro Bernaola-Galvan, and Jose L. Oliver, "Application of information theory to DNA sequence analysis: a review", *Pattern Recognition*, vol. 29, no. 7, pp. 1187-1194, 1996.

[2] Bitzer, D.L., Vouk, M.A., "A table-driven (feedback) decoder.", *Proceedings of the Tenth Annual International Phoenix Conference on Computers and Communications*, 1991 pp 385 - 392

[3] Elebeoba E. May, "Comparative Analysis of Information Based Models for Initiating Protein Translation in Escherichia coli K-12," M.S. Thesis, NCSU, December 1998.

[4] May, E.E., Vouk, M.A., Bitzer, D.L., Rosnick, D.I., "The ribosome as a table-driven convolutional decoder for the Escherichia coli K-12 translation initiation system," *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 4, pp. 2466 - 2469, 2000.

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

[6] Elebeoba E. May, ".Analysis of Coding Theory Based Models for Initiating Protein Translation in Prokaryotic Organisms", PhD thesis, North Carolina State University, Raleigh, NC, 2002.

[7] May, E.E., Vouk, M.A., Bitzer, D.L., Rosnick, D.I., "Constructing Optimal Convolutional Code Models for Prokaryotic Translation Initiation," *Proceedings of the 2nd Joint EMBS-BMES Conference*, October 2002

[8] Shu Lin, Daniel J. Costello Jr, *Error Control Coding: Fundamentals and Applications*, Prentice Hall, 1982

[9] Dholakia A., Vouk M.A., and Bitzer D.L., "Table based decoding of rate one-half convolutional codes, " *IEEE Transactions on Communications*, Vol. 43(2-4), pp. 681-686, 1995.

[10] Bitzer D.L., A. Dholakia, H. Koorapaty, and M.A. Vouk, "On Locally Invertible Rate-1/n Convolutional Encoders," *IEEE Transactions on Information Theory*, Vol 44 (1), pp. 420-422, January 1998.

[11] David I. Rosnick, "Free energy periodicity and memory model for genetic coding", PhD thesis, North Carolina State University, Raleigh, NC, 2001

[12] The complete genome of Escherichia Coli ftp://ftp.ncbi.nih.gov/genbank/genomes/Bacteria/Escherichia_coli_K12/U00096.gbk

APPENDIX

A. Table-based convolutional coding

A rate k/n convolutional code will have n generators, each of which operate on L input symbols at a time. The length L, which is also the length of each generator, is called the constraint length. Thus, an n-bit encoded block at time t depends on the k-bit information block at time t, and on m previous information blocks. The encoded data is also referred to as parity. Each generator is applied on the input data, and the encoded symbols are calculated. Then the generators are moved by k positions to the right, and the procedure is repeated. For a rate ½ code, with L = 3, assuming the generators to be **G** = [1 2 4, 1 3 2], input data **u** = [0 0 1 1 0 2 0 0], we get the following parity: **v** = [4 2 1 0 3 4 4 0 4 1 2 2].

The syndrome-former, or g-mask enables one to check if the received parity stream is correct or not. The g-mask is and-ed with the received parity bits and the result is summed modulo-5, giving the syndrome. The procedure is repeated after shifting the g-mask n positions along the parity stream. The g-mask for the above code is [1 4 3 3 2 1]. In the absence of any errors, it can be verified that the syndrome is all zero. The window length w is defined as $w = n(L-k)/(n-k)$. The length of the g-mask = $w+n$

B. Constructing matrix C in (1)

The g-mask coefficients are contained in matrix **a**. The matrix **C** is formed as follows: Consider the rate ½ code, L = 3. First, choose a data vector having just one non-zero element, **u** = [0 0 0 0 1 0 0 0 0];
Apply the generators and obtain the parity stream.
v = [0 0 0 0 4 2 2 3 1 1 0 0 0 0];
C will have $(w+k)$ rows and $(w+n)$ columns. Shift elements from the parity stream into the rows of the C matrix, n at a time. This yields

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 0 & 0 & 4 & 2 \\ 0 & 0 & 4 & 2 & 2 & 3 \\ 4 & 2 & 2 & 3 & 1 & 1 \\ 2 & 3 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

C. Catastrophic codes

A convolutional code generator is said to be catastrophic if a finite number of errors in the parity stream produced by it, result in an infinite number of decoding errors. Catastrophicity is a very undesirable property of any generator. In genetic coding systems, a catastrophic generator would not be able to recognize individual mRNA leader sequences. Invertibility of **C** guarantees non-catastrophicity.