A CODING THEORY FRAMEWORK FOR GENETIC SEQUENCE ANALYSIS

E.E. May * †

Sandia National Laboratories
Computational Biology Department
P. O. Box 5800, Albuquerque, NM 87185

M.A. Vouk, D.L. Bitzer, and D.I. Rosnick

North Carolina State University
Department of Computer Science
Box 8206, Raleigh, NC 27695

ABSTRACT

This work gives a brief overview of information theory based approaches to genetic sequence and system analysis. The state of research with regard to error-correction coding theory methods for evaluating the genetic translation initiation system is explored. We present research results of interest in the area of error-control coding methods for modeling the translation initiation system of *Escherichia coli* K-12.

1. INTRODUCTION

Roman-Roldan et al. suggest that living beings can be characterized by their information processing ability and hence information based analysis can be used in their study [1]. Viewing protein synthesis as an information processing system allows nucleotide sequences to be analyzed as messages without considering the physical-chemical elements for information processing [1]. Transfer of biological information can be modeled as a communication channel with the DNA sequence as the input and the amino acid sequence which forms protein as the channel output [1]. The communication channel view proposed by Roman-Roldan et al. differs from the initial model presented by May et al. [2]. May et al.'s initial model defines the messenger RNA (mRNA) as the output of the communication channel and incorporates a decoder that translates the mRNA into protein forming amino acid chains. Roman-Roldan et al. designate the process of mapping codons to amino acids as the transmission channel through which DNA is transmitted and protein is received. May et al.'s initial model defines the genetic channel as the DNA replication and transcription process during which errors are introduced into the nucleotide sequence [2]. Both May et al. and Roman-Roldan et al. assume the transmission channel to be stationary and memoryless. Schneider et al. and Eigen also evaluate genetic processes based on the systems information processing ability.

*This work performed while at North Carolina State University and supported in part by a NSF Graduate Fellowship, NSF MGE Grant, and Ford Foundation Dissertation Fellowship for Minorities.

†Email: eemay@sandia.gov

1.1. Information Theory in Binding Site Analysis

Schneider et al. [3] analyze *E. coli* binding site (region on DNA and RNA sequences to which macromolecules bind) groups using two information based measures derived from the Shannon entropy, $H = -\sum_{i=1}^{M} p_i \log_2 p_i$ (where p_i is the probability of each symbol i): (1) $R_{sequence}$ - Measure of the information in the binding site sequence patterns; (2) $R_{frequency}$ - Amount of information needed to locate the binding site, given that the binding site occurs with a certain frequency in the genome. $R_{sequence}$ and $R_{frequency}$ serve as quantitative tools for studying how proteins locate their respective binding sites among non-binding site sequences. Schneider et al.'s information based evaluation of binding sites led to two notable discoveries [4]: (1) The consensus sequence (or the most "perfect" sequence) is improbable; (2) There exists an evolutionary relationship between changes or variations of specific control points and the overall cellular control mechanism. This suggests that the genetic translation system (most likely the genetic system as a whole) permits, if not requires, some degree of error. Therefore it must provide some method of error detection and error correction.

1.2. Information Theory and Genetic Evolution

Eigen [5] evaluates evolution based on a living system's informational capacity. He asserts that if reproduction is the foundation for information conservation and if reproduction causes natural selection then there must exist an error threshold for reproduction. Above and below said error threshold, information is lost [5]. Only near the error threshold of reproduction will there exist a large population of viable variations or mutants. This mutant distribution or "quasi-species" have a defined consensus sequence. Their sequences are similar but non-identical [5]. The mutants that form this set of quasi-species are the ones which survive, hence resulting in evolutionary flexibility.

Eigen further suggests that the genetic information, DNA, has error correcting capabilities and that the complementary interactions found in the DNA molecule provide for

an encodable alphabet. The information space concept, or sequence space, developed by Eigen maps nucleic acid sequences to a discrete point space [5]. The distance between the points (sequences) in the sequence space is equal to the number of positions in which the sequences differ from one another [5]. Eigen's sequence space can be paralleled to a decoding sphere that is composed of n-symbol sequences that are located around an n-symbol codeword [6]. The sequence distance concept is equivalent to the Hamming distance concept in coding theory [6]. Eigen and Schneider's work leads us towards a coding theory framework for the analysis of genetic information.

2. CODING THEORY AND BIOLOGICAL INFORMATION PROCESSING

Battail [7] argues, similar to Eigen, that for Dawkins' model of evolution to be tractable, error-correction coding must be present in the genetic replication process. According to Battail, proof-reading, a result of the error avoidance mechanism suggested by genome replication literature, does not correct errors present in the original genetic message. Only a genetic error correction mechanism can guarantee reliable message regeneration in the presence of errors or mutations due to thermal noise, radioactivity, and cosmic rays [7]. The survival of an organism necessitates the existence of a reliable information replication process. Therefore error-correcting codes must be used in replication or in another process of information regeneration that precedes replication [7]. Battail also suggests that genetic information undergoes nested encoding, where the result of a previous encoding process is combined with new information and encoded again. The more important genetic information is assumed to be in the primary coded message [7].

Battail's nested coding model mirrors coding theory's concept of concatenated codes [6]. Based on Battail and Eigen's works, the initial communication view of the genetic system proposed by May et al. [2] is modified as follows: (1) The replication process represents the error-introducing channel; (2) Assuming a nested genetic encoder, the genetic decoding process occurs over three levels: transcription, translation initiation, and translation elongation plus termination. Figure 1 depicts May et al.'s final coding theory view of translation initiation. Battail makes a plea for increased research for the purpose of identifying the error-correcting process proposed [7]. Though there is little known research into error-correcting models for genetic processes [2, 8, 9, 10], there is some research into coding theory based approaches to analyzing genetic sequences [11, 12, 13, 14].

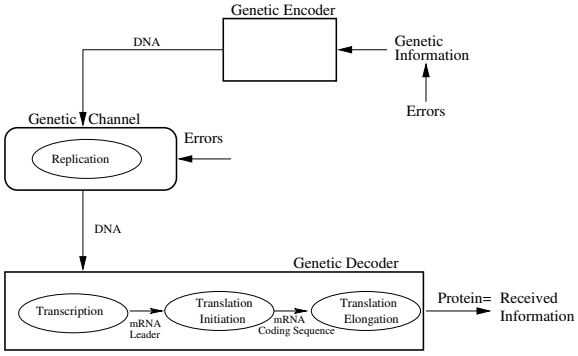

Fig. 1. Modified Coding Theory View of the Central Dogma of Genetics

2.1. Coding Theory and DNA Computing [11]

Kari et al. use circular codes to define heuristics for constructing codewords for DNA computing applications. In DNA computing, the information storage capability of DNA is combined with laboratory techniques that manipulate the DNA to perform computations [11]. A key step in DNA computing is encoding the problem in the DNA strand. The challenge is to find codewords for encoding that do not form undesirable bonds with itself or other codewords used or produced during the computational process. Kari et al. used coding theory to define rules for constructing "good" codewords for DNA computing.

2.2. Coding Theory in Reading Frame Identification

Arques et al. statistically analyzed the results of 12,288 autocorrelation functions of protein coding sequences. Based on the results of the autocorrelation analysis, they identified three sets of circular codes X_0, X_1, X_2 which can be used to distinguish the three possible reading frames in a protein coding sequence [12]. A set of codons X is a circular code, or a code without commas, if the code is able to be read in only one frame without a designated initiation signal [12]. Crick et al. originally introduced the concept of codes without commas in the alphabet A, C, G, T. It was later successfully addressed and extracted over the alphabet R, Y, N [12]. Arques et al. define a circular code over the A, C, G, T alphabet. They were able to use the three sets of circular codes to retrieve the correct reading frame for a given protein sequence in a thirteen base window. They have used their coding based model to analyze Kozak's scanning mechanism for eukaryotic translation initiation and other models of translation [12].

2.3. Coding Theory Based Sequence Analysis

Stambuk also explored circular coding properties of nucleic acid sequences [13] [15]. His approach was based on the

combinatorial necklace model which asks: "How many different necklaces of length m can be made from a bead of q given colors [16, 13]." Using $q = [A, C, G, T]$ and $q = [R = Purine, Y = Pyrimidine, N = R \ or \ Y]$, Stambuk applied the necklace model to genetic sequence analysis, enabling the use of coding theory arithmetic in the analysis of the genetic code [13]. Although Stambuk did not use error control coding in his analysis, his work provided important insight into the structure of DNA sequences [13].

3. CHANNEL CODE MODELS FOR TRANSLATION INITIATION

Although one does not know the exact mechanism employed by the genetic decoder, by analyzing key elements involved in initiating protein translation, it is hoped that we will gain insight into possible decoding schemes used in the initiation of translation in prokaryotic organisms. The key elements considered 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. A block coding model and convolutional coding model for the translation initiation system [2][8][9] were explored. Assuming an encoding method, the corresponding decoding algorithm was designed using the 16S ribosomal RNA.

3.1. Block Code Model [8]

In the block code model, the genetic encoder is modeled as an (n, k) block code whose output is a systematic zero parity check code [17] [2]. Codewords of length $n = 5$ and $n = 8$ were developed based on the last thirteen bases of the 3' end of 16S ribosomal RNA (which contains the hexamer complementary to the Shine-Dalgarno sequence [18]) and the proposed encoder model. The model employed a minimum distance decoder to verify the block coding model for translation initiation.

The *E. coli* K-12 strain MG1655 sequence data (downloaded from the NIH ftp site: ncbi.nlm.nih.gov) was used to test the model. Figure 2 shows the resulting mean minimum distance by position for the (5,2) block code model. The smaller the value on the vertical axis, the stronger the bond formed between the ribosome and the mRNA. Zero on the horizontal axis corresponds to the alignment of the first base of a codeword with the first base of the initiation codon.

As Figure 2 illustrates there is a significant difference among the translated, hypothetical and the non-translated sequence groups. For the translated and hypothetically translated sequence groups, a minimum distance trough occurs between the -15 and -10 regions. 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.

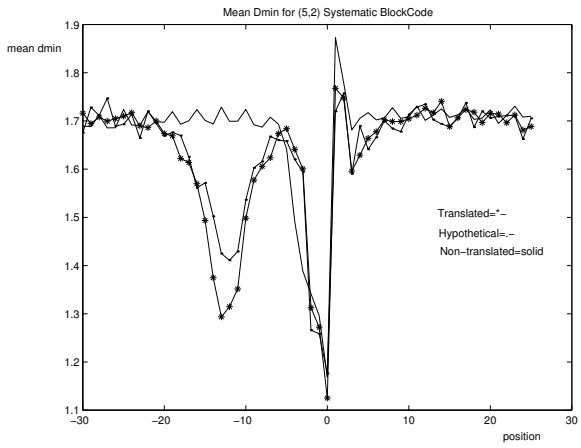

Fig. 2. Results of Minimum Distance Block Decoding Model for (5,2) Code

3.2. Convolutional Code Model [9]

The second error-correcting coding model investigated was based on the principle hypothesis that the messenger RNA (mRNA) sequence can be viewed as a noisy, convolutionally encoded signal. The ribosome was functionally paralleled to a table-based convolutional decoder. The 16S ribosomal RNA (rRNA) sequence was used to form decoding masks for table-based decoding.

Convolutional coding produces encoded blocks based on present and past information bits or blocks. The modeling assumption is that genetic operations such as initiation and translation may involve "decisions" which are based on immediate past and immediate future information. This would allow error correction and other related functions. The convolutional code model viewed the ribosome as a mechanism with memory, which differs from Schneider's idea of macromolecular machines without memory [10]. Evaluating the messenger RNA as convolutionally encoded data allowed the model to capture the inter-relatedness between the bases in a mRNA sequence.

Figure 3 shows the frequency of the most frequent distance pattern among all possible two-symbol distance patterns $d_i d_j$, where distance values range from zero to four. The horizontal axis indicates position, with zero corresponding to the alignment of the coding 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).

As shown in Figure 3, the convolutional code model was able to distinguish between translated and non-translated

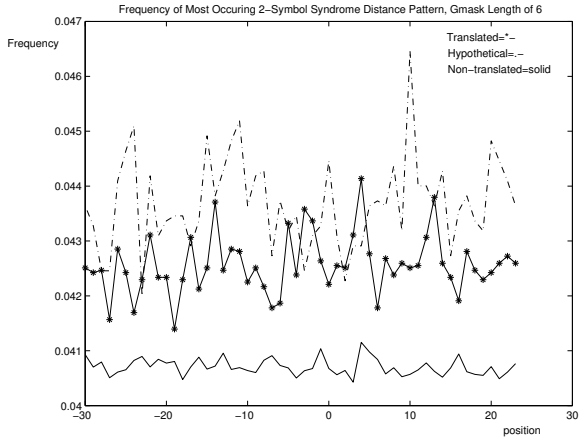

Fig. 3. Frequency of Two-Pattern Syndrome Distance Values

sequence groups. The distinction among hypothetical and translated groups is also evident. The convolutional code model indicated 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 [18].

3.3. Analysis of Coding-Based Models

Three issues were critical to analyzing the effectiveness of each error-control model for translation initiation: (1) Recognition of regions within the mRNA leader sequence; (2) Distinction between translated and non-translated sequence groups; (3) Indication and recognition of the open reading frame construct. Both models distinguished translated sequence groups from the non-translated sequence group. They both also indicated the existence of key regions within the mRNA leader sequence. The block code model seemed to recognize the ribosomal binding site (the location of the Shine-Dalgarno sequence) more readily than the convolutional code model. The block code model also indicated the existence of a reading frame synchronization construct more so than the convolutional code model. Additional results for longer block codes and results for the longer gmask(twelve-base masks) are presented in [2].

4. CONCLUSION

The results of the error-control coding models suggest that it is 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. Results also imply that genetic systems may use methods which are functionally parallel to channel coding techniques to protect and detect genetic signals. The successful development and implementation of a channel coding model for the translation initiation system can lead to the development of powerful methods for identifying and manipulating protein coding sequences within a genome as well as further our understanding of translation regulatory mechanisms.

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

[3] Thomas D. Schneider, Gary D. Stormo, Larry Gold, and Andzej Dhrenfeucht, "Information Content of Binding Sites on Nucelotide Sequences," *Journal of Molecular Biology*, vol. 188, pp. 415–431, 1986.

[4] Thomas D. Schneider, "Information content of individual genetic sequences ," *Journal of Theoretical Biology*, vol. 189, pp. 427–441, 1997.

[5] Manfred Eigen, "The origin of genetic information: viruses as models ," *Gene*, vol. 135, pp. 37–47, 1993.

[6] Richard E. Blahut, *Theory and Practice of Error Control Codes*, Addison-Wesley Publishing Company, Inc., Reading, MA, 1983.

[7] G. Battail, "Does information theory explain biological evolution? ," *Europhysics Letters*, vol. 40, no. 3, pp. 343–348, November 1997.

[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] Elebeoba E. May, Mladen A. Vouk, Donald L. Bitzer, and David I. Rosnick, "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.*, 2000.

[10] Thomas D. Schneider, "Theory of Molecular Machines. I. Channel Capacity of Molecular Machines," *Journal of Theoretical Biology*, vol. 148, pp. 83–123, 1991.

[11] Lila Kari, Rob Kitto, and Gabriel Thierrin, "Codes, Involutions and DNA Encodings," University of Western Ontario, London, Ontario, Canada. Submitted.

[12] Didier G. Arques and Christian J. Michel, "A code in the protein coding genes," *BioSystems*, vol. 44, pp. 107–134, 1997.

[13] Nikola Stambuk, "On circular coding properties of gene and protein sequences," *Croatica Chemica ACTA*, vol. 72, no. 4, pp. 999–1008, 1999.

[14] Nikola Stambuk, "On the genetic origin of complementary protein coding ," *Croatica Chemica ACTA*, vol. 71, no. 3, pp. 573–589, 1998.

[15] Nikola Stambuk, "Symbolic Cantor Algorithm (SCA): A method for analysis of gene and protein coding ," *Periodicum Biologorum*, vol. 101, no. 4, pp. 355–361, 1999.

[16] Elwyn R. Berlekamp, *Algebraic Coding Theory*, McGraw-Hill Book Company, New York, NY, 1968.

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

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