Proceedings of the Second Joint EMBS/BMES Conference
Houston, TX, USA • October 23-26, 2002

CONSTRUCTING OPTIMAL CONVOLUTIONAL CODE MODELS FOR PROKARYOTIC TRANSLATION INITIATION

Elebeoba E. May[1*], Mladen A. Vouk[1,2], Donald L. Bitzer[2], and David I. Rosnick[2]
[1]Department of Electrical and Computer Engineering, North Carolina State University, NC, USA
(*email: eueni1@eos.ncsu.edu)
[2]Department of Computer Science, North Carolina State University, NC, USA

Abstract— **Rapid advances in both genomic data acquisition and computational technology have encouraged the development and use of engineering methods in the field of bioinformatics and computational genomics. Several researchers are encouraging the use of error-correction coding in analyzing genetic data [1][2]. Using information theory, coding theory specifically, the translation of messenger RNA (mRNA) into amino acid sequences is functionally paralleled to the decoding of noisy, convolutionally encoded parity streams. The ribosome is modeled as a table-based convolutional decoder. This work presents a genetic algorithms (GAs) method for the design of optimal table-based convolutional coding models for prokaryotic translation initiation sites using *Escherichia coli* K-12 as the model organism. We explore and compare several categories of error-control codes, including: horizontal, vertical, equal and unequal error protection (UEP) codes. Results show that UEP code models recognize the non-random and Shine-Dalgarno domain of mRNA leaders better than equal error protection models. Codes whose decoding masks (gmasks) have high similarity to the 3' end of the 16S ribosomal RNA (rRNA) were discovered. Additional results are presented.**

Keywords— **Coding Theory, Translation Initiation, Genetic Algorithms**

I. Introduction

Previous research [3][4] showed that error control coding can be used to describe the translation initiation region in an average sense. Present work aims to develop a set of convolutional coding models that specifically describe individual leader sequences. This work investigates horizontal and vertical code models. Horizontal codes correspond to a coding model in which a single encoder produces the encoded genetic sequence for a single receiver. The vertical coding model corresponds to a multiple receiver model, where the message is encoded such that specific regions within the sequence are recognized by specific receivers.

Channel codes can be generally defined as pattern recognition systems [5]. The codewords produced by the code are patterns the system wants to recognize. A "good" code will recognize valid patterns with a probability of one and all non-system patterns with a probability of zero. This work defines a "good" code based on how well the code recognizes the "patterns" or RNA bases that form the leader region. Unlike block code design, which has proven construction methods, "good" convolutional codes are designed using search techniques [6][7] and in recent years genetic algorithms (GAs) [8][7]. Genetic algorithms are numerical optimization techniques based on a generalized view of the theory of evolution, natural selection, and genetics [9]. Usually, the definition of a good convolutional code is based on memory length, error detecting, and error correcting capabilities. For "genetic" convolutional codes the most important feature of a good code is how well it distinguishes errors from non-errors, non-ribosome binding sites from ribosome binding sites. The construction of genetic convolutional code models using genetic algorithms is the first step towards designing efficient translation initiation sites for transgenic protein production.

II. Methods

A genetic algorithm is used to search for table-based convolutional codes whose gmasks recognize individual mRNA leader sequences. The GAs search space consists of all possible $(n = 3, k = 1, m = 4)$ convolutional codes or individuals. The fitness of each individual in the population (a set of potential solutions) is based on the syndrome values produced when the code's gmasks are applied to the mRNA parity sequence. A syndrome value of zero indicates that no errors within the code's error detection capability occurred. Random selection and target sampling rates are used to select highly fit individuals for reproduction. New populations are created using parameterized uniform crossover. Mutation is used to preserve population diversity and elitism ensures that the most fit solution is not discarded. The GA searches for the optimal horizontal equal-weight (equal error protection) and motif-based (UEP) codes for each sequence. To construct vertical code models, the GA searches for the optimal convolutional code in each position of the leader region.

III. Results

Messenger RNA leader sequences from *E. coli* K-12 strain MG1655 (downloaded from the NIH ftp site: ncbi.nlm.nih.gov and parsed by Rosnick [10]) were used as training sequences for constructing the best candidate code model. The syndrome distance vector for each code model is calculated and indicates how well the associated decoder recognizes the subsequence at hand. If the genetic algorithm found the perfect code - the convolutional coding system that produced the exact sequence - then the syndrome distance vector would be the all zero vector and the fitness value would be one. Figure 1 shows the average syndrome distance value for the optimal codes discovered using the *E. coli* training set. In Figure 1 the horizontal axis is position relative to the first base in the initiation codon and the vertical axis is the average syndrome distance value. For the horizontal codes the individual syndrome distance values for each code model are averaged over 266 models. The vertical code model is the average syndrome value for each of the 48 positional models.

0-7803-7612-9/02/$17.00 © 2002 IEEE

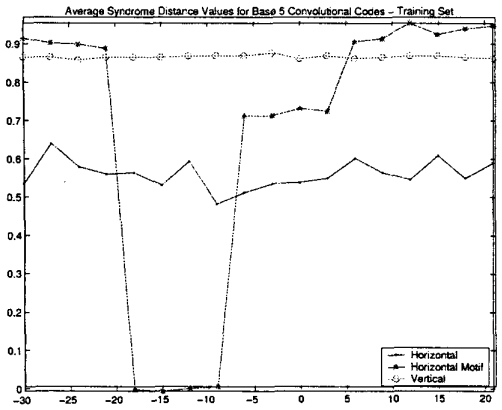

Fig. 1. Average Syndrome Distance of Table-Based Convolutional Code Models for Translation Initiation

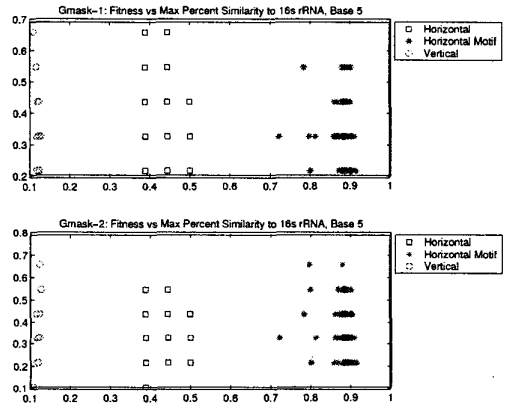

Fig. 2. Individual Fitness versus Individual Similarity Values for Code Models; Gmask1(top) and Gmask2(bottom)

If the ribosome functionally parallels a table-based decoder, the gmask of the code models may resemble the exposed part of the 16S rRNA. Convolutional codes with high similarity to the last thirteen bases of the 16S rRNA and low syndrome distance (i.e. high fitness) values would, from a biological perspective, be more plausible models for translation initiation. Figure 2 depicts the relationship between each code model's fitness score and the model's similarity to the last thirteen bases of the 16S rRNA. The horizontal axis in Figure 2 is fitness and the vertical axis is percent similarity.

IV. Discussion

As Figure 1 illustrates, when compared to the horizontal code models, the average syndrome distance for the vertical code models does not indicate any regions of significant activity. The lowest average syndrome distance value for the equal weight horizontal code models occurs at position -9 while the motif-based horizontal code models have approximately zero average syndrome distance values from position -18 to position -9. These positions correspond to the non-random domain and the Shine-Dalgarno domain, key regions in the translation initiation process.

In Figure 2, for gmask1, the equal weight horizontal code model and the vertical code model achieve the highest percent similarity to the 16S rRNA. But, for gmask2 the motif-based horizontal code model and the vertical code model achieve the highest percent similarity scores. In all code model groups, there exist individuals with high similarity values and relatively high fitness values.

V. Conclusion

The motif-based horizontal code results suggest that UEP or nested codes are used by the genetic system to protect vital information. Although the motif-based codes recognize the non-random and Shine-Dalgarno domain better than the other two models, preliminary testing suggests that equal weight horizontal code models are more effective error detectors. The existence of individual gmask values with high fitness values and high similarity to the exposed portion of the 16S rRNA lends credibility to the supposition that the ribosome's behavior is equivalent to a table-based decoder, where the 16S rRNA functions as the system's decoding mask. The results of the sequence-based convolutional code models are encouraging. Further investigation continues.

Acknowledgments

This research supported in part by a NSF Graduate Fellowship, a NSF MGE Grant, and a Ford Foundation Dissertation Fellowship for Minorities.

References

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

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

[3] 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.

[4] Elebeoba E. May, Mladen A. Vouk, Donald L. Bitzer, and David I. Rosnick, "Coding Theory Based Maximum-Likelihood Classification of Translation Initiation Regions in Escherichia coli K-12 ," in *2000 Biomedical Engineering Society Annual Meeting.*, 2000.

[5] Yu. G. Savchenko and A. A. Svistel'nik, "An Approach to Pattern Recognition Systems," *Engineering Cybernetics*, , no. 2, pp. 144–146, 1968.

[6] Tadashi Wadayama, Koichiro Wakasugi, and Masao Kasahara, "An 8-Dimensional Trellis-Coded 8-PSK with Non-Zero Crossing Constraint," *IEICE Trans. Fundamentals*, vol. E77-A, no. 8, pp. 1274–1280, August 1994.

[7] Robert Kotrys and Piotr Remlein, "The Genetic Algorithm used in search of the good TCM codes," in *4th Inetrnational Workshop on Systems, Signals and Image Processing, IWSSIP'97.*, 1997, pp. 53–57.

[8] Tiffany M. Barnes, "Using Genetic Algorithms to Find the Best Generators for Half-Rate Convolutional Coding," North Carolina State University, Raleigh, NC, 1994.

[9] David A. Coley, *An Introduction to Genetic Algorithms for Scientists and Engineers*, World Scientific Publishing Co. Pte. Ltd., Singapore, 1999.

[10] David I. Rosnick, *Free Energy Periodicity and Memory Model for E. coli Codings*, Ph.D. thesis, North Carolina State University, Raleigh, NC, 2001.