DNA Sequences Generated by $\mathbb{Z}_4$-linear Codes

Andréa Santos Leite Rocha  
Departamento de Telemática  
Universidade Estadual de Campinas  
São Paulo, Brazil 13083-970  
andrea@dt.fee.unicamp.br

Luzinete Cristina Bonani Faria  
Departamento de Telemática  
Universidade Estadual de Campinas  
São Paulo, Brazil 13083-970  
luzinete@dt.fee.unicamp.br

João Henrique Kleinschmidt  
Departamento de Telemática  
Universidade Estadual de Campinas  
São Paulo, Brazil 13083-970  
joaohk@gmail.com

Reginaldo Palazzo Jr  
Departamento de Telemática  
Universidade Estadual de Campinas  
São Paulo, Brazil 13083-970  
palazzo@dt.fee.unicamp.br

Márcio Castro Silva-Filho  
Departamento de Genética  
Universidade São Paulo  
São Paulo, Brazil 13400-970  
mdcsilva@esalq.usp.br

Abstract—One of the puzzling problems in mathematical biology is to show the existence of any form of error-correcting code in the DNA structure. Here we propose a model for the biological coding system similar to that of a digital communication system. This model consists of an encoder (a mapper and a BCH code over $\mathbb{Z}_4$) and a modulator (genetic code). Here we show that DNA sequences including proteins and targeting sequences from different species with 63, 255, and 1023 nucleotides long were identified as codewords of $\mathbb{Z}_4$-linear codes.

I. INTRODUCTION

Due to the scientific and technological relevance and the enormous amount of genomic sequences available and data banks, the research in genomic coding has been focused on: proposing systematic procedures to determine the coding and non-coding regions in the DNA structure; characterizing the error-correcting code structure in DNA sequences; modeling the biological system to that of a layered communication network.

Schneider, [1], has proposed some algorithmic procedures to determine the coding and non-coding regions in the DNA structure. Under the coding theory point of view, Liebovitch, [2], proposes a procedure capable of determining whether a type of error-correcting code is or is not present in the DNA sequence. Rosen, [3], proposes a method for detecting linear block codes and so explaining the insertions and deletions in the DNA sequences. May et. al., [4], proposes the use of block and convolutional codes in the translation initialization process in prokaryotic organisms. Battail, [5], argues on the existence of nested codes in the DNA, due to the fact that the human genome is greater than the necessary to specify the individual characteristics. Under the communication network point of view, [6] and [7] have considered the information contained in the genome is structured and organized in packets or frames. Each packet contains identification, coding and non-coding regions, targeting and stop sequences, and so on. A question always present in the majority of the works related with genomic coding, is the following: Is there any form of error-correcting code in the DNA structure? This paper answers the question positively based on the fact that if the genome consists of exons, introns, promoter, repetitive DNAs, internal signaling, etc and that each one of these sequences may be reproduced by a specific code, then the genome consists of concatenated codes.

II. CODING MODEL PROPOSAL

One possible interpretation of Shannon’s channel coding theorem, regarding the flow of information from the source to the sink is that the mutual information of the discrete channel (consisting of the modulator, channel, and the demodulator) be as close as possible to the entropy of the source. To achieve this goal, an error correcting code has to be used. Therefore, the transmitter in a digital communication system model consists of two cascade blocks, one block associated with an encoder and the other block associated with a modulator (signal constellation).

Here we make use of this model characterized as follows: the encoder consists of a mapper, a set of transformations taking an element from the set of nucleotides corresponding to the bases adenine (A), cytosine (C), guanine (G), and thymine (T) or uracil (U), denoted by $N = \{A,C,G,T/U\}$, to an element of the set $\mathbb{Z}_4 = \{0,1,2,3\}$, and an encoder of a BCH code; the modulator (signal constellation) consists of the genetic code where each codon is associated with a signal in the signal constellation. The codeword at the encoder output is directly related to the mature mRNA, whereas the output of the modulator is related to the protein. Although quite obvious in the biological context, the process of matching each codon in the mature mRNA strand with its corresponding complementary codon from the genetic code by the tRNA, in a digital communication system context this property is quite important and it is called matched mapping [8]. This property in addition to implying that the underlying algebraic structure of the encoder and the modulator are...
the same up to an isomorphism, it guarantees the least complexity. The class of codes satisfying this property is known as geometrically uniform codes [9]. An important subclass of the geometrically uniform codes is the G-linear codes [10]-[12], where G denotes an algebraic group.

In Fig. 1, the encoder consists of a mapper and an encoder of a linear block code. The modulator consists of the genetic code, tRNA and the ribosome. The genetic code may be viewed as a signal constellation, where each codon is considered as a signal in the signal constellation, code may be viewed as a signal constellation, where each of the genetic code, tRNA and the ribosome. The genetic encoder of a linear block code. The modulator consists of codes [10]-[12], where G denotes an algebraic group. subclass of the geometrically uniform codes is the G-linear codes [10]-[12].

In Fig. 1, the encoder consists of a mapper and an encoder of a linear block code. The modulator consists of the genetic code, tRNA and the ribosome. The genetic code may be viewed as a signal constellation, where each codon is considered as a signal in the signal constellation, code may be viewed as a signal constellation, where each of the genetic code, tRNA and the ribosome. The genetic encoder of a linear block code. The modulator consists of codes [10]-[12], where G denotes an algebraic group.

In Fig. 1, the encoder consists of a mapper and an encoder of a linear block code. The modulator consists of the genetic code, tRNA and the ribosome. The genetic code may be viewed as a signal constellation, where each codon is considered as a signal in the signal constellation, code may be viewed as a signal constellation, where each of the genetic code, tRNA and the ribosome. The genetic encoder of a linear block code. The modulator consists of codes [10]-[12], where G denotes an algebraic group.

In Fig. 1, the encoder consists of a mapper and an encoder of a linear block code. The modulator consists of the genetic code, tRNA and the ribosome. The genetic code may be viewed as a signal constellation, where each codon is considered as a signal in the signal constellation, code may be viewed as a signal constellation, where each of the genetic code, tRNA and the ribosome. The genetic encoder of a linear block code. The modulator consists of codes [10]-[12], where G denotes an algebraic group.

In Fig. 1, the encoder consists of a mapper and an encoder of a linear block code. The modulator consists of the genetic code, tRNA and the ribosome. The genetic code may be viewed as a signal constellation, where each codon is considered as a signal in the signal constellation, code may be viewed as a signal constellation, where each of the genetic code, tRNA and the ribosome. The genetic encoder of a linear block code. The modulator consists of codes [10]-[12], where G denotes an algebraic group.

In Fig. 1, the encoder consists of a mapper and an encoder of a linear block code. The modulator consists of the genetic code, tRNA and the ribosome. The genetic code may be viewed as a signal constellation, where each codon is considered as a signal in the signal constellation, code may be viewed as a signal constellation, where each of the genetic code, tRNA and the ribosome. The genetic encoder of a linear block code. The modulator consists of codes [10]-[12], where G denotes an algebraic group.
determined from the generator polynomial \( g(x) \), together with the parity-check matrix \( H \), which is obtained from the polynomial \( b(x) = (x^n - 1)/g(x) \). We note that, for each primitive polynomial used in the generation of the ring, \( GR(4, r) \) leads to a different generator polynomial \( g(x) \). We have to consider this fact when looking for a new code.

The error correction capability of a code is related to the number of codewords, which, in the case under consideration, equals \( 4^k \), where \( k = n - r \). For a given value of \( n \), smaller values of \( r \) lead to more codewords and, therefore, greater computational complexity in generating all \( 4^k \) codewords. In order to overcome this problem, which is classified as NP-complete, instead of generating all the codewords and then comparing them with a given DNA sequence, we first consider the DNA sequence, under the action of each of the eight permutations related to the polynomial \( g(x) \). Hence, to determine whether each of the eight possibilities is in fact a codeword, we use the linear, as a codeword. Hence, to determine whether each one of these possibilities is a codeword, we use the relation \( vH^T = 0 \), where \( v \) is a possible codeword (\( v = uG \), where \( u \) is the input information sequence and \( G \) is the generator matrix) and \( H^T \) is the transpose of the parity-check matrix. To analyze the DNA sequence, which differs by one nucleotide from the codeword, we consider three other possibilities for nucleotides in each position in the sequence, for each permutation, and again we use the relation \( vH^T = 0 \) to verify whether \( v \) is a codeword.

V. DNA Sequence Generation Algorithm

Input data: 1) \( seq \) = original DNA sequence in nucleotides (NCBI); 2) \( n = 2^r - 1 \); and 3) \( d_H = 2t + 1 \).

- Step 1 - Generate all primitive polynomial (PP) with degree \( r \) to be used in the Galois ring and Galois field extensions;
- Step 2 - Select one PP at a time from Step 2, and find the group of units of \( GR(4, r) \), denoted by \( GR^*(4, r) \);
- Step 3 - Find the generator and parity-check polynomials of the BCH code by knowing the minimum distance and the primitive element derived in Step 2. As a consequence the generator and the parity-check matrices as well as its transposes are determined;
- Step 4 - From the mapping \( N \rightarrow Z_4 \), convert the \( seq \) with elements in \( N \) into the corresponding sequence with elements in \( Z_4 \);
- Step 5 - Verify by use of the syndrome \( s \) if each one of the converted DNA sequences is a codeword:
  - If \( s = 0 \), then store the sequence;
  - If \( s \neq 0 \) implies that it may exist up to \( t \) differences of nucleotides, so one has to consider the \( n \) combinations \( t \) by taking the three other possibilities of nucleotides in each one of the combinations of the DNA sequence. Verify if each one of these combinations is a codeword: if so, store it; otherwise disregard it;
- Step 6 - From the mapping \( Z_4 \rightarrow N \) convert each stored sequence in Step 6 with elements in \( Z_4 \) into the corresponding sequence with elements in \( N \). Compare each one of these sequences with the \( seq \) and show the position where the nucleotides differ;
- Step 7 - Go to Step 1. Select another PP and verify if all PP have already been used: if not, repeat Steps 2 to 5 for each PP from Step 1; otherwise, go to Step 8.
- Step 8 - End.

VI. Results

The results may be described as follows:

- By employing the reciprocal mapping, \( Z_4 \rightarrow N \), we can create eight equal codewords in terms of nucleotides and amino acids;
- The results allow a new approach for the classification of DNA sequences from a mathematical point of view;
- Due to space limitation, Table I shows some of the many sequences which were reproduced by the \( Z_4 \)-linear codes (BCH code over ring) and classified as nonlinear cyclic sequences;
- The first three entries in Table I show DNA sequences whose generator (primitive) polynomials are the same. This means that all of these sequences can be identified as codewords that belong to the same code, although differing with respect to its input information sequence \( u \). Therefore, distinguishing among these sequences requires referring to the input information sequence;
- Another relevant property deals with the degree of primitive and generator polynomials. For a given DNA sequence length \( n \), we note a dependency between the error-correcting codes that were capable of reproducing the corresponding DNA sequences and the primitive polynomials;
- Among all codes with a minimum distance \( d > 2 \), only certain codes with \( d = 3 \) were capable of reproducing the corresponding DNA sequences. This fact implies that the degrees of the primitive polynomials \( r \) and of the generator polynomial \( (n - k) \) are equal. Any redundancy is associated with the degree of either one of these polynomials. Hence, a low redundancy implies a high-rate code as well as high entropy (substantial information flow);
- In the present study, every DNA sequence identified as a codeword of a \( G \)-linear code is related to a primitive/generator polynomial and to a labeling, suggesting the existence of an intrinsic geometric property that is associated with each DNA sequence;
- Due to space limitation, Fig. 3 and Fig. 4 show some targeting sequences which were reproduced by the \( Z_4 \)-linear. Under the biological point of view we have noticed that in the sequences \( B. napus \) and \( N. tabacum \) occurred silent mutations, since a change of nucleotide does not lead to another amino acid. On the other hand, in the sequences \( A. thaliana \) and \( M. martensii \) the amino acid changes occurred within
the same hydrophilic class, whereas in the sequence *P. dominulus* the change occurred within the same hydrophilic class.

- The *transversion mutations* (change of one purine/pyrimidine for another) occurred in the sequences *B. napus* (*C* → *T*), *N. tabacum* (*T* → *C*) and *M. martensi* (*G* → *A*), whereas the *transversion mutations* (change of a purine for a pyrimidine or vice-versa) occurred in the sequences *A. thaliana* (*A* → *C*) and *P. dominulus* (*T* → *A*).

- What has been observed is that there is always a single nucleotide difference between the original DNA sequence and the codeword generated by a BCH code.

In the biological context this mismatch is known as a *single nucleotide polymorphism* (SNP). Hence, one possible interpretation is that either the codeword generated by a BCH code is an SNP with respect to the corresponding original DNA sequence or the other way around;

- However, since this mismatch is within the error-correction capability of the code, it follows that the modified Berlekamp-Massey decoding algorithm for rings is capable of providing the position and the magnitude to e added in this position to fully reproduce the original DNA sequence.

- Although we have considered in this study DNA sequences of length *n* = 2<sup>*r* </sup>− 1, this is not a restriction since a class of error-correcting codes with variable codeword lengths may be used.

VII. CONCLUSIONS

This paper shows that the DNA sequences such as proteins (P) and targeting sequences (TS) are identified as codewords of *Z*<sub>4</sub>-linear codes.

ACKNOWLEDGMENT

The authors would like to thank FAPESP Grant 2007/56052-8, 2009/50837-9, and CNPq Grant 306617/2007-2.

REFERENCES


### TABLE I

DNA sequences generated by *Z*<sub>4</sub>-linear codes.
Fig. 3. Examples of DNA sequences generated by Z₄-linear codes

Z₄-linear codes

- Oaa: E V I E W
- Gaa: N V Y Y V R Y L T K N E A Y I I H K L N E
- Gnt: AAT GTT TAC TAT GTC AGA TAC TTA ACG AAA AAT GAG GCA TAT ATT AT
- Glb: 002 322 201 202 321 030 201 220 013 000 002 303 310 202 022 02
- Olb: 002 322 201 202 321 030 201 220 013 000 002 303 310 202 022 02
- Ont: AAT GTT TAC TAT GTC AGA TAC TTA ACG AAA AAT GAG GCA TAT ATT AT

Fig. 4. Example of a protein reproduced by the Z₄-linear code

A. thaliana - Pathogenesis related protein 4* - TS - GI number 186529758
- Oaa: M K I R L S
- Gnt: ATG AAG ATC AGA CTT AGC
- Glb: 023 003 022 032 010 122 220 222 120 322 021 212 312 210 02
- Olb: 023 003 022 032 010 122 220 222 120 322 021 212 312 210 02
- Ont: ATG AAG ATC AGA CTT AGC

B. napus - Malate dehydrogenase* - TS - GI number 899225
- Oaa: M K I S
- Gnt: ATG AAA ATT AGT
- Glb: 023 000 022 032
- Olb: 023 000 022 032
- Ont: ATG AAA ATT AGT

C. napus - Anti-epilepsy peptide precursor
- Oaa: LVLYVIAAGANA
- Gnt: CTG GTC CTC TAT GTC ATA GCC GCA GGA GCT AAT GCA
- Glb: 123 321 121 202 321 020 311 310 330 312 002 310
- Olb: 123 321 121 202 321 020 311 310 330 312 002 310
- Ont: CTG GTC CTC TAT GTC ATA GCC GCA GGA GCT AAT GCA

M. martensii - anti-epilepsy peptide precursor - TS - GI number 16742522
- Oaa: K L F L L L V I S A S
- Gnt: GC TTA ATT TGT CTC GTA ATT GTT CTT ACG ATC ATT CAT TTG TCT CAA GCT
- Glb: 22 220 300 300 003 220 302 300 012 000 311 121 000 021 112 302 322
- Olb: 22 220 300 300 003 220 302 300 012 000 311 121 000 021 112 302 322
- Ont: GC TTA ATT TGT CTC GTA ATT GTT CTT ACG ATC ATT CAT TTG TCT CAA GCT

Abbreviations:
- Oaa = original amino acid
- Gnt = original nucleotides
- Glb = generated labeling
- Gnt = generated nucleotides
- Gaa = generated amino acid

Fig. 4. Example of a protein reproduced by the Z₄-linear code