VI International Conference on Computational Bioengineering
ICCB 2015
M. Cerrolaza and S.Oller (Eds)

IDENTIFYING mRNA SEQUENCES AND PROTEINS BY USE OF BCH CODES

M. Duarte-González[*] and R. Palazzo Jr.[*]

[*] School of Electrical and Computer Engineering (FEEC)
University of Campinas
FEEC/UNICAMP, Av. Albert Einstein, 400 Campinas, Brazil
e-mail: mduarte@dt.fee.unicamp.br, palazzo@dt.fee.unicamp.br

Key words: mRNA sequences, Proteins, BCH codes.

Abstract. In a recent work, a model of an intra-cellular transmission system of genetic information, similar to a digital communication system, has been proposed and narrow sense BCH error correcting codes over Z_4 and F_4 have been used for identifying a mathematical structure in DNA and mRNA sequences. In this work, for mRNA sequences, we use the proposed transmission system and extend its capability by considering not only narrow sense as well as non-narrow sense BCH codes. As a consequence, we are able to identify a mathematical structure for an increased number of mRNA sequences. For proteins, we establish an analogy between properties of error-correcting codes and proteins and propose a methodology for establishing a mathematical structure and for representing proteins by use of BCH codes over Z_{20} and F_4xF_5. The mapping from amino acids to Z_{20} is defined by using Dayhoff's matrix and the isometry between Z_4 and F_4. Consequently, some mRNA sequences and proteins from NCBI and PDB data banks, respectively, are identified.

1 INTRODUCTION

One of the great challenges of the scientific community on genomics and proteomics is to provide convincing arguments and proper hypothesis on the existence of a mathematical structure related to DNA, mRNA and proteins such that they may be formulated into an information and coding theory framework. This embedding can help to solve the question "How can information required for the proper functioning of a cell, an organism, or a specie be transmitted in a "hostile" environment?" [1] and contribute to the general understanding of biological communication mechanisms.

In [2], a model of an intra-cellular transmission system of genetic information, similar to a model of digital communication system, has been proposed and narrow sense BCH error-correcting codes over Z_4 and F_4 have been used for identifying a mathematical structure in DNA and mRNA sequences [3-5]. In this work, for mRNA sequences, we use the proposed transmission system and propose a procedure for considering all possible BCH codes. Therefore, we are able to identify a mathematical structure for an increased number of mRNA sequences.

In the case of proteins, there is a clear relation between properties of error-correcting codes and biologically functional proteins, as described next:

- Sequences over an alphabet of cardinality 20 with amino acid chains.

- Code's codewords with biologically functional proteins.
- An error correcting code *C* with a set of functional proteins.
- Correctable sequences for codeword *c* with proteins similar to a specific functional protein.

These links and the capability of chaperone molecules to detect errors [6, 7] justify the use of ECCs for modelling proteins and amino acid sequences. In this work, we aim to propose a methodology for representing or identifying proteins by use of BCH codes over Z_{20} and F_4xZ_5 (both with 20 elements). BCH codes are considered because its structure is well-known and they are relatively easy to design. We call attention to the use of the word error-correcting codes as an error control mechanism instead of stating that the biological information system explicitly corrects mutations.

In Section 2, we introduce the basic concepts in coding theory and further information can be found in [8]. In Section 3, we detail the methodology and procedures for identifying mRNA sequences and proteins by BCH codes. In section 4, we show and discuss the results when applying the methodology to some mRNA sequences and proteins. Finally, in Section 4 we draw the conclusions.

2 BASIC CODING THEORY CONCEPTS

Error Correcting Codes (ECCs) are always used for reliably transmitting and storing information, even if the communication channel is noisy; hence, the transmitted sequences may differ from the received ones. An *error-correcting code* (ECC) $\boldsymbol{C}$ is a subset of A^n, where *A* is the *alphabet* and any sequence of length *n* that belongs to the code is a *codeword*. A code $\boldsymbol{C}$ and its error detection and correction capabilities are specified by three parameters: the codeword length (n), the number of codewords in $\boldsymbol{C}$ ($|\boldsymbol{C}|$), and the minimum distance (d_c). In this work, we consider the *Hamming distance* as metric for two sequences $u=(u_1,...,u_n)$ and $v=(v_1,...,v_n)$ in A^n; and it counts the number of positions in which *u* and *v* differ*:*

$$d(u,v)=|\{i \ : \ u_i \neq v_i\}| \tag{1}$$

The Hamming minimum distance of the code (d_c) specifies the smallest number of positions by which any two different codewords differ and, therefore, the code can detect (d_c –1) at most d_c errors and correct $t=\lfloor (d_c-1)/2 \rfloor$ errors ($\lfloor \cdot \rfloor$ represents the floor operator).

In the next subsections we introduce the alphabets we use throughout this work and give a brief explanation on BCH codes over these alphabets.

2.1 Alphabets

The encoder in the transmission system receives the information message to be transmitted and, uniquely, maps it into one of the codewords. The receiver receives a sequence that can be different from the transmitted codeword and corrects it to obtain the transmitted codeword. In order to detect or correct errors, the ECC and the alphabet must have a well defined mathematical structure. In this work, we use four types of alphabets for designing ECCs, namely: Z_4 , Z_{20} , F_4 and F_4xZ_5; which are described as follows:

- **Integers module *m*** ($\boldsymbol{Z_m}=\{0,1,...,m-1\}$): It is a *ring* [8] with two binary operations: addition and multiplication modulo *m*. Let *a* and *b* be two elements in Z_m, then the operation addition modulo *m*, denoted by $(a+_m b)$, is obtained by reducing modulo *m*

the usual integer addition of a and b ($(a+b)_m$); and the operation multiplication modulo m, denoted by $(a \bullet_m b)$, is obtained by reducing modulo m the usual integer multiplication of a and b ($(a \bullet b)_m$).

Example. In Z_4 : $(2+_4 2)=(4)_4=0$, $(3+_4 3)=(6)_4=2$, $(2 \bullet_4 2)=(4)_4=0$ and $(3 \bullet_4 3)=(9)_4=1$ (We use the + symbol to represent both $+_4$ and $+_{20}$, the reader must identify the operation by the context).

The fundamental theorem of arithmetic and the Chinese Remainder Theorem [8] establish that the ring Z_m is isomorphic to a product of local rings:

$$m = p_1^{r_1} \cdot \ldots \cdot p_s^{r_s} \qquad (2)$$

$$Z_m = Z_{p_1}^{r_1} \oplus \ldots \oplus Z_{p_s}^{r_s}$$

where the p_i's are prime numbers and the r_i's are integer numbers greater than or equal to 0. Therefore, $Z_{20} = Z_4 x Z_5$ and the ring isomorphism is shown in Table 1.

Example. The isomorphism can be used to compute operations in Z_{20} : $(7+15)=((3,2)+(3,0))=(2,2)=2$ and $(8 \bullet 12)=((0,3) \bullet (0,2))=(0,1)=16$.

Table 1: Ring isomorphism between Z_{20} and $Z_4 x Z_5$

Z_{20}	0	1	2	3	4	5	6	7	8	9
$Z_4 x Z_5$	(0,0)	(1,1)	(2,2)	(3,3)	(0,4)	(1,0)	(2,1)	(3,2)	(0,3)	(1,4)

Z_{20}	10	11	12	13	14	15	16	17	18	19
$Z_4 x Z_5$	(2,0)	(3,1)	(0,2)	(1,3)	(2,4)	(3,0)	(0,1)	(1,2)	(2,3)	(3,4)

- **Galois field of 4 elements** ($F_4=\{0, 1, a, 1+a=b\}$): It is a *field* [8] and its two binary operations are defined according to Table 2.

  Example. $(1+a)=b$, $(a+a)=0$, $(a \bullet b)=1$ and $(a \bullet a)=b$

Table 2: Addition and multiplication operations in F_4

$a+b$	0	1	a	b
0	0	1	a	b
1	1	0	b	a
a	a	b	0	1
b	b	a	1	0

$a \bullet b$	0	1	a	b
0	0	0	0	0
1	0	1	a	b
a	0	a	b	1
b	0	b	1	a

2.2 BCH codes

BCH codes belong to the class of *cyclic linear error correcting codes* [8]. A code $\boldsymbol{C}$ is said to be cyclic if for any codeword $v=(v_1,\ldots,v_n)$ in $\boldsymbol{C}$, a cyclic shift of v (represented by $v^{(1)}=(v_n,v_1,\ldots,v_{n-1})$) also belongs to $\boldsymbol{C}$. From now on we make the following identification from sequences in A^n to polynomials in the residue class ring $R=A[x]/(x^n-1)$:

$$(u_0,u_1,\ldots,u_{n-1}) \in A^n \leftrightarrow u_0+u_1 x+\ldots+u_{n-1} x^{n-1} \in A[x]/(x^n-1)=R \qquad (3)$$

Since any cyclic code is an ideal in R [8], it follows that BCH codes are also ideals in R

and their construction is based on the unique factorization of the polynomial x^n-1 :

$$x^n-1 = f_1(x)...f_s(x) = (x-1)(x-\alpha)(x-\alpha^2)...(x-\alpha^{n-1}) \quad (4)$$

where the *fi's* are called *minimal polynomials* over A and α is a cyclic element of order n in the extension field or in the extension ring of A. α is said to be of order n when $\alpha^n=1$ and $\alpha^i\neq 1$ for $0<i<n$. Let us define the generated set by α as $G_n=\{1,\alpha,\alpha^2,...,\alpha^{n-1}\}$.

BCH codes are principal ideals generated by $g(x)$, see equation (5), where $g(x)$ is a polynomial over A constructed by the non repeated multiplication of some *fi's*, and it is called the *generator polynomial* of the code.

$$(g(x))=\{g(x)z(x) : z(x) \in A[x]/(x^n-1)\} \quad (5)$$

When A is F_4 or a field $Z_p=F_p$ (where p is a prime number), and $n=p^r-1$ for any positive integer r, the ring R is a principal ideal domain (all ideals in R are principal), the factorization shown in equation (4) is unique and α is a *primitive* element of order n of the ring $\Gamma=A[x]/(p(x))$, where $p(x)$ is a *primitive polynomial* with degree r and α satisfies $p(\alpha)=0$. Primitive polynomials are tabulated as shown in [8].

When A is Z_4 (a local ring) and $n=2^r-1$, for any integer r, the factorization shown in equation (4) is unique and α does exist [9] and it is computed according to the following procedure: 1) compute the ring extension $\Gamma=Z_4[x]/(p(x))$, where $p(x)$ is a primitive polynomial with degree r over Z_2 and let γ represent the element in Γ such that $p(\gamma)=0$; and 2) over Γ, γ is an element of order $n\cdot l$, where l is an integer greater than or equal to 1; so consider α as γ^l ($\alpha=\gamma^l$) and note that the order of α in Γ is n.

Using the above notation, the *primitive BCH code over A* of length n and generator polynomial $g(x)$, such that $\alpha^e, \alpha^{e+1},..., \alpha^{e+\delta-2}$ are roots of $g(x)$ (i.e. $g(\alpha^e)=0,...,g(\alpha^{e+\delta-2})=0$ over Γ), has a Hamming minimum distance (d_c) greater than or equal to δ.

The non-primitive BCH codes were introduced for the construction of such codes with lengths different from p^r-1 (or 2^r-1). Consider m satisfying $n=a\cdot m=p^r-1$ (or $n=a\cdot m=2^r-1$), i.e. m is a divisor of n. Then, the polynomial x^m-1 can be factored by substituting α^a by β ($\beta=\alpha^a$) in equation (4) and the generator polynomial is the non repeated multiplication of some *fi's*. Note that all *fi's* from equation (6) are in equation (4), however the converse is not true.

$$x^m-1 = f_1(x)...f_k(x) = (x-1)(x-\beta)(x-\beta^2)...(x-\beta^{m-1}) \quad (6)$$

Using the above notation, the *non-primitive BCH code over A* of length m and generator polynomial $g(x)$, such that $\beta^e, \beta^{e+1},..., \beta^{e+\delta-2}$ are roots of $g(x)$ (i.e. $g(\beta^e)=0,...,g(\beta^{e+\delta-2})=0$ over Γ), has Hamming minimum distance (d_c) greater than or equal to δ.

One subclass of the primitive and non-primitive BCH codes is formed by the narrow-sense BCH codes. A n*arrow-sense primitive (or non-primitive) BCH code over A* of length n (or m) and design distance δ is a BCH code such that $e=1$; i.e. the generator polynomial $g(x)$ has $\alpha, \alpha^2,..., \alpha^{\delta-1}$ (or $\beta, \beta^2,..., \beta^{\delta-1}$) as its roots.

Example. Construction of a narrow-sense primitive BCH code over Z_4 of length 7 and Hamming minimum distance $d_c\geq 3$.

Since $7=n=2^3-1$, it follows that a primitive polynomial of degree 3 over Z_2 is needed: $p(x)=x^3+x+1$.

Using the fact that $p(\gamma)=\gamma^3+\gamma+1=0$, we can compute the group generated by γ (see Table 3), where $\gamma^3=3\gamma+3$. Note that the order of γ is 14, then l must be equal to 2 to obtain α as an element of order 7 ($\alpha=\gamma^2$). Considering the generator polynomial as: $g(x)=(x-\alpha)(x-\alpha^2)(x-$

$\alpha^4)=x^3+2x^2+x+1$, we get $e=1$ and $\delta=3$.

Table 3: Multiplicative group generated by γ ($\gamma^3=3\gamma+3$)

1	1	1
γ	γ	
γ^2	γ^2	α
γ^3	$3+3\gamma$	
γ^4	$\gamma\cdot\gamma^3=3\gamma+3\gamma^2$	α^2

γ^5	$1+\gamma+3\gamma^2$	
γ^6	$1+2\gamma+\gamma^2$	α^3
γ^7	$3+2\gamma^2$	
γ^8	$2+\gamma$	α^4
γ^9	$2\gamma+\gamma^2$	

γ^{10}	$3+3\gamma+2\gamma^2$	α^5
γ^{11}	$2+\gamma+3\gamma^2$	
γ^{12}	$1+3\gamma+\gamma^2$	α^6
γ^{13}	$3+3\gamma^2$	
γ^{14}	1	α^7

3 METHODOLOGY

3.1 Identifying mRNA sequences by use of BCH codes

The mRNA sequences were obtained from the NCBI database and only sequences that satisfy the primitive and non-primitive length constraints were considered. In the case the alphabet is Z_4, the allowable lengths (n) are divisors of 2^r-1; and in the case alphabet is F_4, the allowable lengths are divisors of 4^r-1. Only alphabets with four elements were considered, since there are only four nucleotides ($\boldsymbol{N}=\{A,C,G,U\}$): adenine, cytosine, guanine and thymine.

Since the alphabet of the mRNA sequences must be converted into the alphabet of the BCH codes, and vice-versa, an association between the elements of the set $\boldsymbol{N}=\{A,C,G,U\}$ and the elements of the set Z_4 (or F_4) must be established. We call this association: a labeling. There are twenty four possible labeling, corresponding to the 24 permutations of $\boldsymbol{N}$. The labelings are shown in Table 4 and Table 5 for alphabets Z_4 and F_4, respectively.

When considering Z_4, according to [2, 4, 5], three subgroups of eight labelings were identified (labeling A, B and C) as shown in Table 4, hence equal results were obtained independent of the labeling used in that subgroup. When considering F_4, according to [3], all the 24 possible labelings led to the same result, therefore, it is enough to consider one arbitrary labeling for performing the procedure. In this work we consider all labelings, since those conclusions were valid only for narrow-sense BCH codes.

Table 4: Labelings and permutation subgroups from N to Z_4

Labeling **A**								Labeling **B**								Labeling **C**							
A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
0	1	3	2	0	3	1	2	0	1	2	3	0	3	2	1	0	2	1	3	0	2	3	1
A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
1	2	0	3	1	0	2	3	1	2	3	0	1	0	3	2	1	3	2	0	1	3	0	2
A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
2	3	1	0	2	1	3	0	2	3	0	1	2	1	0	3	2	0	3	1	2	0	1	3
A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
3	0	2	1	3	2	0	1	3	0	1	2	3	2	1	0	3	1	0	2	3	1	2	0

Table 5: Labelings and permutation from $\boldsymbol{N}$ to F_4

Labelings																							
A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
0	1	a	b	0	a	b	1	0	b	1	a	0	1	b	a	0	a	1	b	0	b	a	1
1				2				3				13				14				15			
A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
1	0	b	a	a	0	1	b	b	0	a	1	1	0	a	b	a	0	b	1	b	0	1	a
4				5				6				16				17				18			
A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
a	b	0	1	b	1	0	a	1	a	0	b	a	b	1	0	b	1	a	0	1	a	b	0
7				8				9				19				20				21			
A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T	A	C	G	T
b	a	1	0	1	b	a	0	a	1	b	0	b	a	0	1	1	b	0	a	a	1	0	b
10				11				12				22				23				24			

In order to identify mRNA sequences as codewords of primitive and/or non-primitive BCH codes, we apply the procedure described next:

- *Step 1:* Using the selected labeling, map the nucleotide sequence ($\boldsymbol{N}^n$) into a vector over Z_4 (or F_4).
- *Step 2:* Construct the ring (or field) extension $\Gamma=Z_4[x]/(p_i(x))$ ($\Gamma=F_4[x]/(p_i(x))$) by using a primitive polynomial $p_i(x)$ over Z_2 (or F_4).
- *Step 3:* Compute the minimal polynomials Z_4 (or F_4) that factorize x^n-1
- *Step 4:* Select the minimal polynomials that divide the translated sequence.
- *Step 5:* Select the elements in G_n that are roots of the minimal polynomials obtained from *Step 4*.
- *Step 6:* Verify the BCH bound, i.e. find the values e and δ and compute $g(x)$ as: $g(x)=\text{lcm}(f_e,\ldots,f_{e+\delta-2})$, where $\text{lcm}(\cdot)$ is the least common multiple operation and $f_e,\ldots,f_{e+\delta-2}$ are the minimal polynomials from *Step 4*, such that $f_i(\alpha^i)=0$.
- *Step 7:* Return the mathematical structure of BCH codes ($p(x)$ and $g(x)$), that have a design distance (δ) greater than or equal to 3.
- *Step 8:* Go to *Step 2* and choose another primitive polynomial $p_i(x)$ over Z_2 (or F_4).

3.2 Identifying proteins sequences by use of BCH codes

Since the proteins are sequences of amino acids and that there are 20 different amino acids, it follows that we must construct ECCs over alphabets with 20 elements. There are few results related to this problem, probably due to the few applications in engineering and information theory using alphabets with 20 elements. In [10], a methodology for designing codes over Z_m (integers module m) has been proposed. In the case $m = 20$ (Z_{20} is the integers modulo 20), the methodology uses the Chinese Remainder Theorem shown in equation (2) and consists in joining component-wise (by the Cartesian product) all codewords of two ECCs of equal length n over Z_4 and Z_5. The properties of the code over Z_{20} (C_{20}) are deduced from the

parameters of the codes over Z_4 and Z_5 (C_4 and C_5). C_{20} is cyclic if C_4 and C_5 are both cyclic codes, the length of C_{20} is n, the number of codewords is $|C_4|\cdot|C_5|=|C_{20}|$ and, if the minimum distance of the codes C_4 and C_5 are d_{c4} and d_{c5}, respectively, then the minimum distance of C_{20} is given by:

$$d_{c20} = \min\{d_{c4} , d_{c5}\} \tag{7}$$

The analysed proteins were obtained from the RCSB Protein Data Bank (PDB) and only proteins that satisfy the primitive and non-primitive length constraints for both Z_4 and Z_5 were considered. Therefore, n must be a divisor of both $2^{r1}-1$ and $5^{r2}-1$

In the case of mRNA, there are four nucleotides and 4! = 24 permutations or labelings. In the case of proteins, there are 20 amino acids and 20! = $24{,}33\text{x}10^{17}$ permutations or labelings from the set of amino acids $\boldsymbol{AA}$ = {K, Q, N, D, E, P, G, A, S, T, C, V, I, M, L, F, Y, W, H, R} to the set Z_{20}. Therefore, it is unfeasible to test every possible labeling for proteins. In [11], ideally the Dayhoff's mutation odds matrix is constrained to a circle and it expresses the idea that the amino acids which are close together exchange frequently. This representation reminds the traditional mathematical representation of the ring Z_{20}, as shown in Figure 1. Therefore, as labelings for proteins, we consider the labeling shown in Figure 1 and all its other 39 dihedral symmetries (rotations and reflections). In[12], the Dayhoff revised matrix is considered and the results coincides with [11] except for the exchange of amino acids R and H as indicated by an arrow in Figure 1. We also consider the labeling with the exchange of R and H and all its other 39 dihedral symmetries. In the total we consider 80 different labelings.

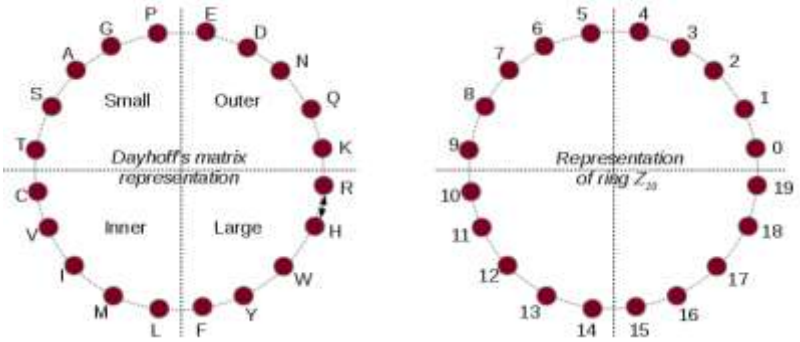

Figure 1: Left. Representation of Dayhoff's matrix according to [XXX].
Right - Graphical representation of the ring Z_{20}.

Another alphabet with 20 elements is $F_4\text{x}Z_5$. Similar to the code design procedure used for the alphabet Z_{20}, two ECCs (C_4 and C_5) with equal length n over F_4 and Z_5, respectively, are used to construct an ECC (C_{45}) with length n over $F_4\text{x}Z_5$. C_{45} is obtained by joining component-wise (Cartesian product) all codewords of C_4 and C_5. Again, the properties of C_{45} are deduced from the parameters of the codes C_4 and C_5. C_{45} is cyclic if C_4 and C_5 are both cyclic codes, the length of C_{45} is n, the number of codewords is $|C_4|\cdot|C_5|=|C_{45}|$ and if the minimum distance of the codes C_4 and C_5 are d_{c4} and d_{c5}, respectively, then the minimum distance of C_{45} is given by $d_{c45} = \min\{d_{c4} , d_{c5}\}$.

For the alphabet $F_4\text{x}Z_5$, 80 labelings were considered and they were obtained by using the isometry (an isometry is a map that preserve distance between elements) between Z_4 and F_4, see equation (8) [8].

$$\text{Isometry } Z_4 \rightarrow F_4 \ : \ \{0\rightarrow(0,0),\ 1\rightarrow(1,0),\ 2\rightarrow(1,1),\ 3\rightarrow(0,1)\} \tag{8}$$

Knowing the labelings from ***AA*** to Z_{20} and that $Z_{20}=Z_4 x Z_5$; we apply the isometry of equation (8) to the component Z_4 of the ring Z_{20} to obtain the labeling from ***AA*** to $F_4 x Z_5$. For example, considering the labeling expressed in Figure 1, the amino acid H is mapped to element 18 in Z_{20} or (2,3) in Z_4 xZ_5 and to element ((1,1),3) in $F_4 x Z_5$. Since we have used an isometry between Z_4 and F_4, then the Dayhoff matrix's idea (amino acids which are close together exchange frequently) is passed from alphabet Z_{20} to alphabet $F_4 x Z_5$.

In order to identify proteins as codewords of primitive and/or non-primitive BCH codes, we apply the following procedure:

- *Step 1:* Using the selected labeling, map the protein $(\boldsymbol{AA})^n$ into a vector over Z_{20} (or $F_4 x Z_5$).
- *Step 2:* Using the Chinese Remainder Theorem do the map from Z_{20} to $Z_4 x Z_5$, split the Z_{20} sequence into two sequences over Z_4 and Z_5 (or split the $F_4 x Z_5$ sequence into two sequences over F_4 and Z_5).
- *Step 3:* Apply twice the procedure shown in Section 3.2. One for identifying the Z_4 sequence as a codeword of a BCH code and the other for identifying the Z_5 sequence as a codeword of a BCH code (or one for identifying the F_4 sequence as a codeword of a BCH code and the other one for identifying the Z_5 sequence as a codeword of a BCH code).
- *Step 4:* Return the mathematical structure of both BCH codes C_4 and C_5, if their design distances (δ_4 and δ_5) are both greater than or equal to 3.

Table 6: mRNA sequences identified by non narrow-sense BCH codes over F_4
Abbreviations: Organism (Org), Eukaryotic cell (EC), Brassica napus (Bn), Arabidopsis thaliana (At), Nicotiana tabacum (Nt)

mRNA GI number	**Org Cell**	**Labeling δ**	**Primitive polynomial Generator Polynomial**	**SNP Position**	**Length (n)**
899225	Bn	1 – 12	$b+x+x^2+x^3$	UUC (F) → GUC (V)	63
	EC	4	$b+x+ax^2+x^4+ax^5+x^7$	1° codon	
899225	Bn	13 – 24	$a+x+x^2+x^3$	UUC (F) → GUC (V)	63
	EC	4	$a+x+bx^2+x^4+bx^5+x^7$	1° codon	
186509758	At	1 – 12	$a+x+x^2+x^3$	AGC (S) → AGU (S)	63
	EC	6	$a+bx+ax^2+x^3+x^5+bx^6+ax^8+x^{10}$	6° codon	
186509758	At	13 – 24	$b+x+x^2+x^3$	AGC (S) → AGU (S)	63
	EC	6	$b+ax+bx^2+x^3+x^5+ax^6+bx^8+x^{10}$	6° codon	
186509758	At	10 – 12	$a+x+x^2+x^3$	AGC (S) → AGU (S)	63
	EC	4	$1+x+x^2+ax^3+ax^4+x^5+x^6+x^7$	13° codon	
186509758	At	B3	$b+x+x^2+x^3$	UCA (S) → UCC (S)	63
	EC	19 – 21	$1+x+x^2+bx^3+bx^4+x^5+x^6+x^7$	13° codon	
632733	Nt	1 – 3	$a+x+x^2+x^3$	GGA (G) → GCA (A)	63
	EC	4	$1+bx+bx^2+bx^5+bx^6+x^7$	1° codon	
632733	Nt	13 – 15	$b+x+x^2+x^3$	GGA (G) → GCA (A)	63
	EC	4	$1+ax+ax^2+ax^5+ax^6+x^7$	1° codon	

4 RESULTS AND DISCUSSION

In order to analyze the mismatching between an mRNA sequence and a codeword, we consider three other possibilities for nucleotides in each position in the mRNA sequence; i.e. we search for codewords that are one Hamming distance unit of a given mRNA sequence. This procedure makes sense, since the codes we are constructing can correct one error in any position.

Table 7: Protein identified by BCH codes over Z_{20} and F_4xZ_5
Abbreviations: IAAl-E3 heterodimer (IAAl-E3), NS2 (2-32) peptide on Hepatitis GB virus B (NS2 peptide), proto-oncogene tyrosine-protein kinase LCK (PROTO)

PDB number	Molecule	Labeling	Primitive polynomial (Z_5)	Mutation
		δ_5	Generator Polynomial (Z_5)	
	Length	Z_{20} or F_4xZ_5?	Primitive polynomial (Z_4 or F_4)	Position
		δ_4	Generator Polynomial (Z_4 or F_4)	
1U0I	IAAl-E3	Tay / Swa	$1+2x^2+2x^3+2x^4+x^6$	Reproduced (No mutation)
		6	$4+2x+4x^2+4x^3+2x^4+4x^5+4x^6+x^7+x^8+3x^9+x^{10}+x^{11}+3x^{12}+x^{13}$	
	21	Z_{20}	$1+x+3x^2+3x^4+2x^5+x^6$	Reproduced (No mutation)
		6	$3+2x+3x^2+3x^3+2x^4+3x^5+3x^6+x^7+x^8+2x^9+x^{10}+x^{11}+2x^{12}+x^{13}$	
1U0I	IAAl-E3	Tay / Swa	$1+2x^2+2x^3+2x^4+x^6$	Reproduced (No mutation)
		6	$4+2x+4x^2+4x^3+2x^4+4x^5+4x^6+x^7+x^8+3x^9+x^{10}+x^{11}+3x^{12}+x^{13}$	
	21	F_4xZ_5	$1+ax+x^3$	Reproduced (No mutation)
		6	$1+x^2+x^3+x^5+x^6+x^7+x^8+x^{10}+x^{11}+x^{13}$	
2LZP	NS2 peptide	Tay	$4+4x+4x^2+x^3$	A → R
		3	$1+3x+x^4$	
	31	Z_{20}	$3+3x+x^2+3x^3+2x^4+x^5$	15° amino acid
		4	$1+x+x^2+2x^3+2x^4+x^6$	
2LZP	NS2 peptide	Swa	$4+4x+2x^2+x^3$	A → G
		3	$1+2x^2+x^3+x^4$	
	31	Z_{20}	$3+3x+x^2+3x^3+2x^4+x^5$	15° amino acid
		4	$1+x+x^2+2x^3+2x^4+x^6$	
1H92	PROTO	Tay	$1+3x+3x^2+3x^3+3x^4+3x^5+x^6$	L → I
		3	$4+3x+2x^6+x^7$	
	63	F_4xZ_5	$a+bx+x^2+x^3$	23° amino acid
		4	$a+ax^2+ax^3+x^4$	
4A46	SSR2857 protein	Swa	$1+2x+3x^2+3x^4+2x^5+x^6$	A → Q
		3	$4+4x+4x^2+3x^3+2x^4+x^5+x^6+x^7$	
	63	Z_{20}	$a+bx+x^2+x^3$	15° amino acid
		4	$1+x+bx^2+x4$	

When studying mRNA sequences over Z_4, we did not identify more mRNA sequences than those identified by the algorithm introduced in [2, 4, 5]. However, the fact that we did not identify mRNA sequences by use of non-narrow sense BCH codes does not guarantee that they do not exist.

In the case of the alphabet F_4, we were able to identify more mRNA sequences than those identified by the algorithm introduced in [3]. Table 6 illustrates some mRNA sequences obtained from the NCBI database. These sequences were analyzed by the proposed procedure and were identified as codewords of non narrow-sense BCH codes over F_4. These results demonstrate that the proposed procedure generalizes the algorithm introduced in [3]. Note that every analyzed mRNA sequence differs by one nucleotide in one position when compared to the closest codeword in the obtained BCH code. Biologically, this difference is considered as an SNP (single nucleotide polymorphism).

Table 7 illustrates some proteins obtained from the RCSB Protein Data Bank. These proteins were analyzed by the proposed procedure and were identified by BCH codes over Z_{20} and $F_4 \text{x} Z_5$. Note that some of the analyzed proteins differ by one amino acid in only one position when compared to the closest codeword of the BCH code. This fact makes sense, since the designed codes are able to correct one error in any position. In Table 7, the two labelings, obtained from [11] and [12] and shown in Figure 1, are denoted by Tay and Swa, respectively.

12 CONCLUSIONS

A procedure for identifying mRNA sequences by use of BCH codes over Z_4 and F_4 has been proposed. This procedure generalizes the algorithm introduced in [2-5] and opens the possibility to identify a mathematical structure for an increased number of mRNA sequences.

A relation between coding theory concepts and protein properties has been introduced: 1) sequences over an alphabet of cardinality 20 with amino acid chains, 2) identified codewords with biologically functional proteins, 3) a code with a set of functional proteins and 4) correctable sequences for codeword $\boldsymbol{c}$ with proteins similar to a specific functional protein.

The methodology for identifying proteins by use of BCH code over Z_{20} and $F_4 \text{x} Z_5$ was realized by examples.

REFERENCES

[1] May, E.E., Vouk, M.A., Bitzer, D.L. and Rosnick, D.I., An error-correcting code framework for genetic sequence analysis. *Journal of the Franklin Institute* (2004). **341**(1–2): 89-109.

[2] Faria, L.C.B., Rocha, A.S.L. and Palazzo Jr, R., Transmission of intra-cellular genetic information: A system proposal. *Journal of Theoretical Biology* (2014). **358**: 208-231.

[3] Faria, L.C.B., Rocha, A.S.L., Kleinschmidt, J.H., Palazzo, R. and Silva-Filho, M.C., DNA sequences generated by BCH codes over GF(4). *Electronics Letters* (2010). **46**(3): 203-204.

[4] Faria, L.C.B., Rocha, A.S.L., Kleinschmidt, J.H., Silva-Filho, M.C., Bim, E., Herai, R.H., Yamagishi, M.E.B. and Palazzo, R., Jr., Is a Genome a Codeword of an Error-Correcting Code? *PLoS ONE* (2012). **7**(5): e36644.

[5] Rocha, A.S.L., Faria, L.C.B., Kleinschmidt, J.H., Palazzo, R., Jr. and Silva-Filho, M.C.

DNA sequences generated by Z4-linear codes. in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*. 2010.

[6] Kriegenburg, F., Ellgaard, L. and Hartmann-Petersen, R., Molecular chaperones in targeting misfolded proteins for ubiquitin-dependent degradation. *FEBS Journal* (2012). **279**(4): 532-542.

[7] Patterson, C. and Höhfeld, J., *Molecular Chaperones and the Ubiquitin–Proteasome System*, in *Protein Science Encyclopedia*. 2008, Wiley-VCH Verlag GmbH & Co. KGaA.

[8] Peterson, W.W. and Weldon, E.J., *Error-correcting Codes*. MIT Press. (1972).

[9] Shankar, P., On BCH codes over arbitrary integer tings (Corresp.). *Information Theory, IEEE Transactions on* (1979). **25**(4): 480-483.

[10] Blake, I.F., Codes over certain rings. *Information and Control* (1972). **20**(4): 396-404.

[11] Taylor, W.R., The classification of amino acid conservation. *Journal of Theoretical Biology* (1986). **119**(2): 205-218.

[12] Swanson, R., A unifying concept for the amino acid code. *Bulletin of Mathematical Biology* (1984). **46**(2): 187-203.