Journal of Engineering Technology and Applied Sciences, 2022
Received 21 October 2022, Accepted 5 December 2022
Published online 22 December 2022, Research Article

e-ISSN: 2548-0391
Vol. 7, No. 3, 219-230
doi: 10.30931/jetas.1192924

Citation: Oztas, E.S., Bulut Yilgor, M., "A Novel Method for Determining the Non-cds Region By Using Error-Correcting Codes". Journal of Engineering Technology and Applied Sciences 7 (3) 2022 : 219-230.

A NOVEL METHOD FOR DETERMINING THE NON-CDS REGION BY USING ERROR-CORRECTING CODES

Elif Segah Oztas[a*], Merve Bulut Yilgor[b],

[a]Department of Mathematics, Karamanoglu Mehmetbey University, Karaman, Turkiye
*(*corresponding author) esoztas@kmu.edu.tr*

[b]Deaperment of Basic Sciences, Faculty of Engineering and Architechture,
Altınbas University Istanbul, Turkiye
merve.yilgor@altinbas.edu.tr

Abstract

Our main motivation question is "Is there any relation between the non-coding region and useless error-correcting codes?". Then we focused CDS and non-CDS areas instead of exon and intron, because CDS involves in process of synthesis a protein and is involved by exons. We get the data of the genes from NCBI [21]. In this study, we introduce the method Fi-noncds that is used for determining the non-CDS region by using error-correcting codes. We obtained that the error-correction codes that can't correct any codes named zero error-correcting code, placed in non-CDS areas, densely. This result shows that non-CDS regions (non-coding areas in DNA) match zero error-correcting codes (useless error-correcting code). Frame lengths 7,8,9 and 10,11,12,13 and 14 were tested by the method. Optimal result for selected genes (TRAV1-1, TRAV1-2, TRAV2, TRAV7, WRKY33, HY5, GR-RBP2) is frame length 8, $n = 7$, $k = 2$, $dnaNo = 1$. Moreover, optimal results of the algorithm Fi-noncds matched the best sequence length 8 as in [1].

Keywords: Fi-noncds, non-CDS, error-correcting codes, zero error-correcting codes

1. Introduction

DNA codes was introduced by L. Adleman 1994 [6]. After this pioneer paper, studies of algebraic structures for DNA strings started. Some of these mathematical structures are dependent on one DNA base [5,14]. This means that one DNA base correspond one element of algebraic structure. For example, let algebraic structure be $Z_4 = \{0,1,2,3\}$ and correspondences for DNA bases are 0 → A, 1 → T, 2 → G, 3 → C. Studies that use more than one DNA base for one algebraic element was started by Oztas and Siap [4]. There are some studies for DNA codes by using DNA k-bases over finite field and rings [7,11,12,18]. However, these studies are not implemented in real DNA strings. These studies aim to solve DNA reversibility problem that introduces in [4], over algebraic structures and obtaining algebraic codes that can be corresponded to reversible DNA codes.

Otherwise, [8,10] studied focus on creation a error-correction code or codeword by using DNA strings. [10] a procedure was developed to determine whether such an error-correcting code is present in the base sequence. In [8] real DNA string of gene TRAV7 corresponded to a codeword over Z_4 with one error. This paper is an important example of relation between real DNA string an error-correction codes. This study was extended in [19].

In [13], Rosen presents a channel model for genome replication by inspiring from communication channel model. Further, an algorithm is given to investigate whether or not there is an underlying linear block code structure in DNA by partitioning DNA sequences with frame offsets.

In eukaryotic genes, a part of gene named CDS (coding DNA sequence) that involves in process of synthesis a protein. The remaining parts of the gene named as non-CDS and not used in DNA strings for the process of the synthesis a proteins. CDSs (coding DNA sequences) are located inside exon regions in genes. The region left over from the exons is called the intron.

Relation between Introns-exons and error-correcting codes are studied in [2, 3, 17]. Regions of introns and exons are defined as check-bits (redundancy) and information bits of error-correcting codes, respectively in [2,3]. In [17], there is a method for finding intron and exon areas by using error-correction codes, but it has unstable algorithm, limited calculations, and results. Then, we consider CDS and non-CDS area instead of exon and intron areas because of CDSs are involved in protein synthesis.

In this study, relation between CDS and non-CDS and error-correcting codes are studied to determine the region of non-CDS that different from [2,3]. we introduce the algorithm Fi-noncds that used for determining the non-CDS region by using error-correcting codes. We obtained that zero error-correcting code, placed in non-CDS areas, densely. This result shows that useless codes place in useless areas in genes. Moreover, the result shows that there is a new relation between genes and mathematical structures.

In the section that follows; in Section 2, the background is given about error-correcting codes that are used in the introduced algorithm. We introduce an algorithm to determine regions of non-CDS in Section [3]. In Seciton [4], results of application the method for selected genes TRAV1-1, TRAV1-2, TRAV2, TRAV7, WRKY33, HY5, GR-RBP2.

2. Background

In this study, error-correcting codes are used to determine the non-CDS area in genes. Then some background in information is given in this section.

A subspace of a vector space over a field F_q^n is a linear code of length n. Generator matrix is the base of the linear code and span the subspace. Hamming distance is number of different components between two vectors (in linear codes, vectors in a code are called codewords). Minimum Hamming distance of code is minimum hamming distance between any two of its codewords. The Hamming weight of a codeword is the number of components that are different from the zero. Minimum distance values exclude the distance zero. Finding the error correction capability of codes is defined that if d is odd then $d = 2t + 1$ (or if d is even, $d = 2t + 1$) where Minimum Hamming distance is d and error correction capability of code is t. For example, let d be 3 then $3 = 2t + 1$ and $t = 1$. It means that the code has minimum distance 3 and corrects 1 error in a codeword. If $d < 3$ then the code cannot correct any codeword. Then, we named "zero error-correction code" (zero-ecc) when $t = 0$ or $d < 3$. Zero-ecc are useless codes in coding theory. They can't correct any error in codes.

A linear code C of length n, dimension k over F_q^n (i.e., number of codewords is $C = q^k$) and minimum Hamming distance d is expressed, in terms of its parameters, with $[n, k, d]$-code. For example, let $C = \{0000, 1100, 0011, 1111\}$ is a subspace of the vector space F_2^4. Therefore C is a linear code of length 4 over F_2 because it has a generator matrix as

$$G = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \tag{1}$$

A generator matrix of a linear code \$C\$ is a matrix whose rows form a basis for C. The dimension of C is 2 $(k = 2)$ that is the number of rows of generator matrix. Hamming weight of all codewords are $w_H(0000) = 0, w_H(1100) = 2, w_H(0011) = 2, w_H(1111) = 4$. In a linear code, minimum Hamming weight (except zero) is also equal to the Minimum Hamming distance of the linear code $(d(C) = \min\{w_H(c) |\ c \in C\}$. Thus, $d(C) = 2$ and C is a [5,2,2]-code.

In linear code, we will use minimum weight of code because of minimum hamming weight is equal to minimum hamming distance of code and this gives the error-correction capability.

The error-correction capability of a code is closely related to the parameters $[n, k, d]$. To enhance the error-correcting capability, we need to improve the distance when the length (n) and dimension (k) are fixed.

In [9], for fixed length n and dimension k, the maximum possible Hamming distance d for a linear code is given. Linear codes with these parameters are called optimal codes, in other words they have the maximum error-correction capability for given n and k.

In this study, the algebraic structure F_4 is used. In $F_4 = \{0, 1, \alpha, \alpha^2\}$ is used and summation is defined under modulo 2 and multiplication is operated by the rule $\alpha^2 = \alpha + 1$.

Example 2.1. Let G be a generator matrix for linear code C_1 of length 4 .

$$G = \begin{bmatrix} 1 & \alpha & 0 & 0 \\ \alpha & 0 & 0 & \alpha \end{bmatrix} \tag{2}$$

Codewords are generated by G as follows.
$x(1, \alpha, 0, 0) + y(\alpha, 0, 0, \alpha)$ where $x, y \in F_4$. All codewords and weights are shown in Table 1. In here, the code C_1 is a $[4,2,2]$ $-$code.

Table 1. Codewords and weights of the code C_1.

x	y	**codeword**	**weight**
0	0	0000	0
1	0	$1\alpha 00$	2
α	0	$\alpha\alpha^2 00$	2
α^2	0	$\alpha^2 100$	2
0	1	$\alpha 00\alpha$	2
1	1	$\alpha^2\alpha 0\alpha$	3
α	1	$0\alpha^2 0\ \alpha$	3
α^2	1	$110\ \alpha$	3
0	α	$\alpha^2 00\alpha^2$	2
1	α	$\alpha\alpha 0\alpha^2$	3
α	α	$1\alpha^2 0\alpha^2$	3
α^2	α	$010\alpha^2$	2
0	α^2	1001	2
1	α^2	$0\alpha 01$	2
α	α^2	$\alpha^2\alpha^2 01$	3
α^2	α^2	$\alpha 101$	3

3. The method Fi-noncds

Exons include coding DNA sequence (CDS) and 3' and 5' untranslated regions. Exons combine in Process of protein creation. In this process, coding DNA sequence (CDS) areas that are located inside in exon, are translated to form proteins. Introns are non-coding sections of a gene and they is not read throughout the protein synthesis process. The intron in human DNA constitutes approximately %97 of the whole DNA [15]. We will consider non-CDS areas for zero-ecc. Non-CDS (non-coding areas) areas include introns and 3' and 5' untranslated regions. Identifying region of intron-exon and especially CDS in gene research are important problems. For this purpose, a CDS database was created at NCBI [16].

The absence of a non-CDS area in the structure of the region involved in protein or enzyme synthesis suggests a low error-correction rate in the non-CDS area. From this point of view, in this paper, the method named Fi-noncds is introduced by the authors.

The method satisfies to determine the non-CDS area of a gene by using error-correcting codes. Error-correction code (Ecc) was not used for determining non-CDS areas in the literature. However, there some studies about a relation between error correcting codes and and DNA strings [2, 19].

In [1], 8 DNA sequences for the determination of the binding site [20] for the separation of DNA chains were more common than sequences in the $6-13$ range on the Arabidopsis Thaliana plant. In this study, it has been revealed that, with the algebraic method and prediction algorithm used, similar to the [1] study, DNA fragments of 8 are also important in determining the non-coded regions. In 8 and [1, 19], genes Trav7 and Arabidopsis Thaliana were considered and then therefore we consider these genes.

To constitute a mathematical representation of DNA sequences, we associate the 4 nucleotides (A,T,G,C) with elements of the Galois Field of order 4 i.e., $F_4 = \{\ 0,1,\alpha,\alpha^2\}$. There are 24 possible correspondences can be made. We focus on the choice of 0, since it affects the minimum distance of the corresponding code. Hence, we have 4 non-isometric labels.

$$\overset{\textbf{Label 1}}{\begin{bmatrix} A & T & G & C \\ \alpha^2 & \alpha & 1 & 0 \end{bmatrix}} \overset{\textbf{Label 2}}{\begin{bmatrix} A & T & G & C \\ \alpha^2 & \alpha & 0 & 1 \end{bmatrix}} \overset{\textbf{Label 3}}{\begin{bmatrix} A & T & G & C \\ 0 & 1 & \alpha & \alpha^2 \end{bmatrix}} \overset{\textbf{Label 4}}{\begin{bmatrix} A & T & G & C \\ \alpha^2 & 0 & 1 & \alpha \end{bmatrix}} \tag{3}$$

In Algorithm 1, we consider the $k < n$. Then $2k < n_f + 1$ because of $n = n_f - k + 1$. Each frame has a starting point (SP) which is the order of the first nucleotide of this frame in the gene. End point (EP) of a frame is the order of the last nucleotide of this frame in the gene.

Algorithm 1 Fi-noncds method

```
procedure FI-NONCDS($GENE, n_f, k, dnaLabel$)
Ensure: $2k < n_f + 1$
    FGNE:=GENE($dnaLabel$) (The converted form of selected DNA to $F_4$ according the selected $dnaLabel$ )
    $N$ := Length(FGENE)
    $n \leftarrow n_f - k + 1$ Determined code length
    for $i \leftarrow 1, N - n_f + 1$ do                                   ▷ SP and EP of frame
        $fr \leftarrow FGENE[i : i + n_f - 1]$
        for $dim \leftarrow 1, k$ do                                     ▷ Generator matrix creation
            $G[dim] \leftarrow fr[dim : dim + n - 1]$
        end for
        $d \leftarrow$ Minimum hamming distance of the code generated by matrix $G$.
        $INDEXD[i] = d$
        $INDEX[i] = floor(\frac{d-1}{2})$
    end for
end procedure
```

The method Fi-noncds illustrated in Figure 1 works as follows:

- String of a gene is converted to the algebraic field F_4 according to a chosen label from the non-isometric label list (3).
- Chose the frame and generate the code.
- Hamming distance and error-correction capabilities are found and they are inserted to the lists INDEXD and INDEX, respectively.
- Result of the lists satisfies the estimation of non-CDS areas.

The number of zero-ecc is determined and the rate at which they are in the non-CDS area is determined. This ratio gives the estimation percentage for non-CDS area.

In Example 3.1, a generation matrix that is generated by using a chosen frame is presented. When we get the lists INDEXD and INDEX, we also classified these linear codes according to their properties, i.e., MDS, self-dual, self-orthogonal, cyclic, and optimal codes. These properties are also important in ECC. Relevant result by using MDS code end other properties is still an open problem.

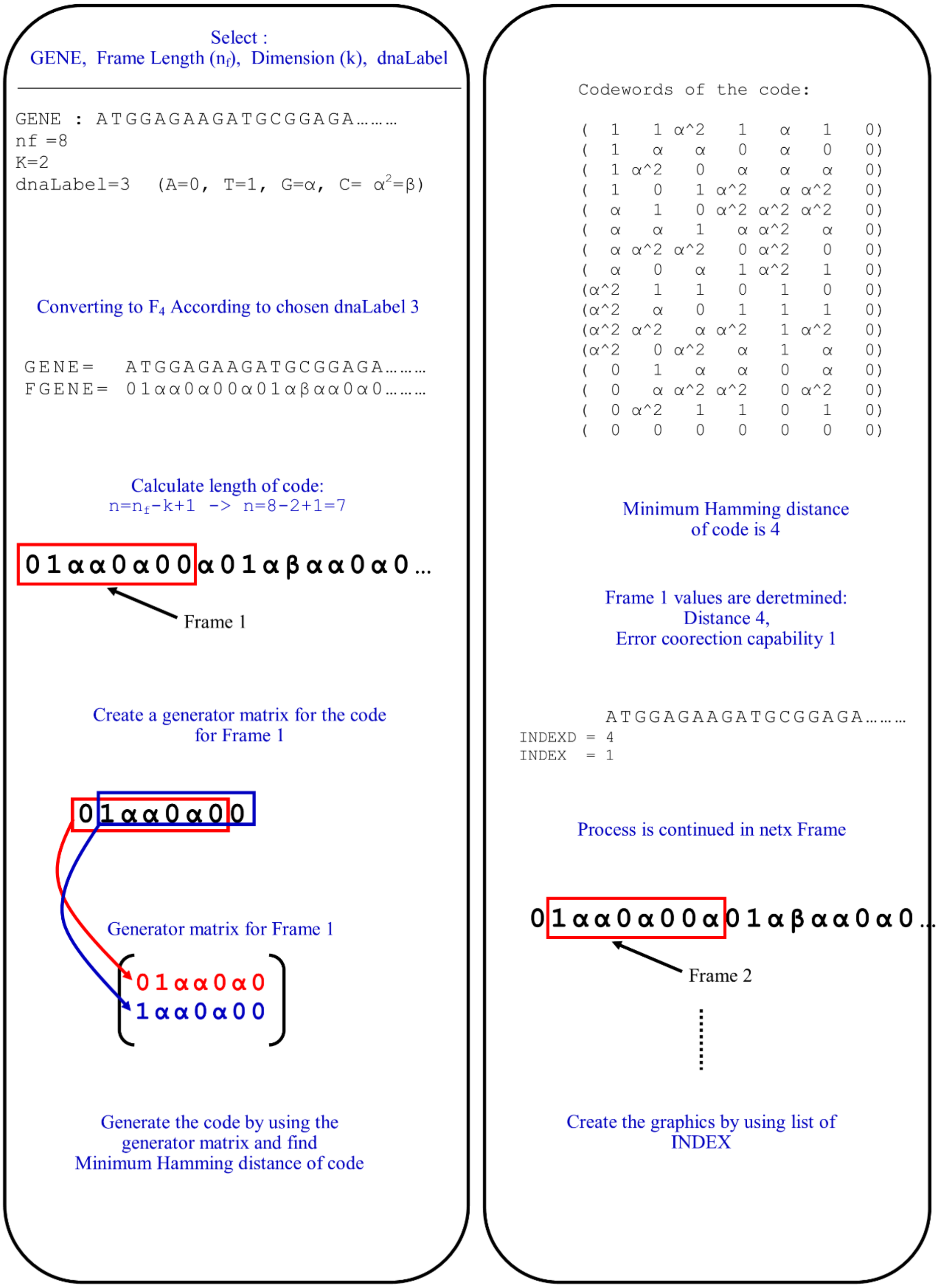
Example 3.1 The frame of length 7 of Trav7-3 (T cell receptor alpha variable 7-3 [Mus musculus (house mouse)]) gene with SP=28 and EP=34 is $GTCCTGT$. According to Label 1, let the dimension $k = 3$ and then find the corresponding vectors of length $n = 5$.

$$R1: \quad GTCCTGT \longrightarrow \quad GTCCT \longrightarrow \quad (1, \alpha, 0, 0, \alpha)$$
$$R2: \quad GTCCTGT \longrightarrow \quad TCCTG \longrightarrow \quad (\alpha, 0, 0, \alpha, 1)$$
$$R3: \quad GTCCTGT \longrightarrow \quad CCTGT \longrightarrow \quad (0, 0, \alpha, 1, \alpha)$$

Hence the generator matrix is obtained as follows.

$$G = \begin{bmatrix} 1 & \alpha & 0 & 0 & \alpha \\ \alpha & 0 & 0 & \alpha & 1 \\ 0 & 0 & \alpha & 1 & \alpha \end{bmatrix}. \tag{4}$$

The linear code C generated by G has parameters [5,3,3] which is an optimal code. Moreover, C is a cyclic MDS code.



Select :
GENE, Frame Length (n_f), Dimension (k), dnaLabel

```
GENE : ATGGAGAAGATGCGGAGA………
nf =8
K=2
dnaLabel=3  (A=0, T=1, G=α, C= α²=β)
```

Converting to F_4 According to chosen dnaLabel 3

```
GENE=   ATGGAGAAGATGCGGAGA………
FGENE=  01αα0α00α01αβαα0α0………
```

Calculate length of code:
n=n$_f$-k+1 -> n=8-2+1=7

Create a generator matrix for the code
for Frame 1

Generate the code by using the
generator matrix and find
Minimum Hamming distance of code

```
Codewords of the code:

(   1    1 α^2    1    α    1    0)
(   1    α    α    0    α    0    0)
(   1 α^2    0    α    α    α    0)
(   1    0    1 α^2    α α^2    0)
(   α    1    0 α^2 α^2 α^2    0)
(   α    α    1    α α^2    α    0)
(   α α^2 α^2    0 α^2    0    0)
(   α    0    α    1 α^2    1    0)
(α^2    1    1    0    1    0    0)
(α^2    α    0    1    1    1    0)
(α^2 α^2    α α^2    1 α^2    0)
(α^2    0 α^2    α    1    α    0)
(   0    1    α    α    0    α    0)
(   0    α α^2 α^2    0 α^2    0)
(   0 α^2    1    1    0    1    0)
(   0    0    0    0    0    0    0)
```

Minimum Hamming distance
of code is 4

Frame 1 values are deretmined:
Distance 4,
Error coorection capability 1

```
             ATGGAGAAGATGCGGAGA………
INDEXD = 4
INDEX  = 1
```

Process is continued in netx Frame

Create the graphics by using list of
INDEX

Figure 1. Illustration of the Fi-noncds

4. Test results

In this section, The method Fi-noncds are tested by using TRAV1-1 (NCBI Gene ID: 28693), TRAV1-2 (NCBI Gene ID: 28692), TRAV2 (NCBI Gene ID: 28691), TRAV7 (NCBI Gene ID: 28686) of Human genes; genes WRKY33 (NCBI Gene ID: 818429), HY5 (NCBI Gene ID: 830996), GR-RBP2 (NCBI Gene ID: 827019) of Arabidopsis thaliana.

In this study, the main window length is 7,8,9 and 10,11,12,13 and 14. Optimal results are obtained where $n = 7$, $k = 2$ and $dnaNo = 1$ (DNA Label) for the tested genes, area of CDS, non-CDS areas. Moreover in [1], sequence length is 8 that have important role in binding site. Also, in this study, best results are obtained where sliding window (main window or main sequence length) is 8 and $n = 7$, $k = 2$ as mentioned below.

The result is given as Table 2. The result shows that the error-correcting codes that cannot do any correction (zero-ecc) are densely placed in non-CDS area. These means, non-coding areas densely include algebraically non-coding useless codes. Figure 2,3,4,5,6,7,8 show the placement of non-CDS and zero-ecc results in the genes. (Higher green lines are CDS areas. The lower green line is the non-CDS area. Red points show where the code appeared in DNA string and the code has what error-correction rate.)

Table 2. The table shows that the zero-ecc rates and numbers in CDS and in non-CDS areas.

Gene name	n	k	dnano	Sum of zero-ecc	In non-CDS	Rate in Non-CDS	In CDS	Rate in CDS
TRAV2	7	2	1	24	22	91.67	2	8.33
HY5	7	2	1	100	86	86	14	14
GR-RBP2	7	2	1	116	97	83.62	19	16.38
WRKY33	7	2	1	118	83	70.34	35	29.66
TRAV1-1	7	2	1	36	25	69.44	11	30.55
TRAV7	7	2	1	19	12	63.16	7	36.84
TRAV1-2	7	2	1	15	15	57.69	11	42.31

sum of zero-ecc: Number of codes that can't do any correction and they have $d < 3$. *in non-CDS:* Number of the codes that have $d < 3$ and placed in non-CDS area. *rate in Non-CDS:* rate of to be in non-CDS area of zero-ecc. *in CDS:* Number of the codes that have $d < 3$ and placed in CDS area. *rate in CDS:* rate of to be in CDS area of zero-ecc.

5. Conclusion

In this study, the method Fi-noncds is introduced that can estimate the rate and regions that are non-CDS by using error-correction codes. The method for estimating the identified non-CDS, the accuracy for a certain percent non-CDS, and revealing data overlapping the results of [1]. According to the result, non-CDS areas have a relation between useless zero-ecc in mathematics.

In here, optimal result is obtained by using the parameter as sliding frame length =8, *n=7*, $k = 2$ and $dnaNo = 1$. As a result of calculations, some result was better than these parameters for only some of the genes in this paper. Some of the results were unacceptable according to a different parameter. Therefore, the improvement of the method Fi-noncds for more genes and parameters is an open problem.

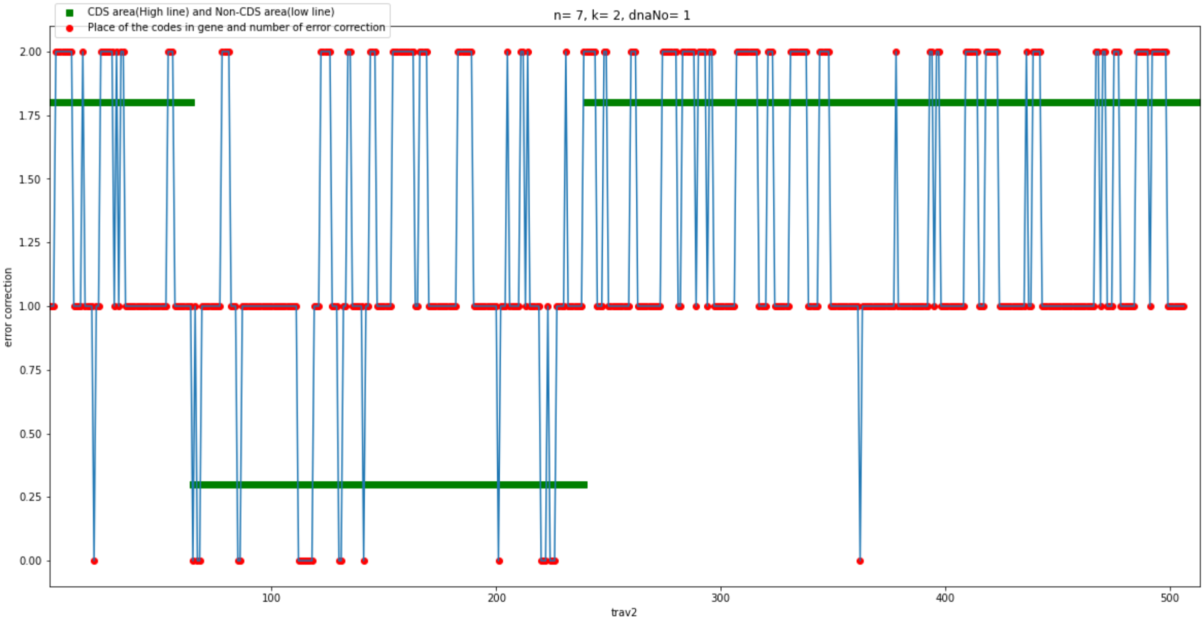


Figure 2. TRAV2

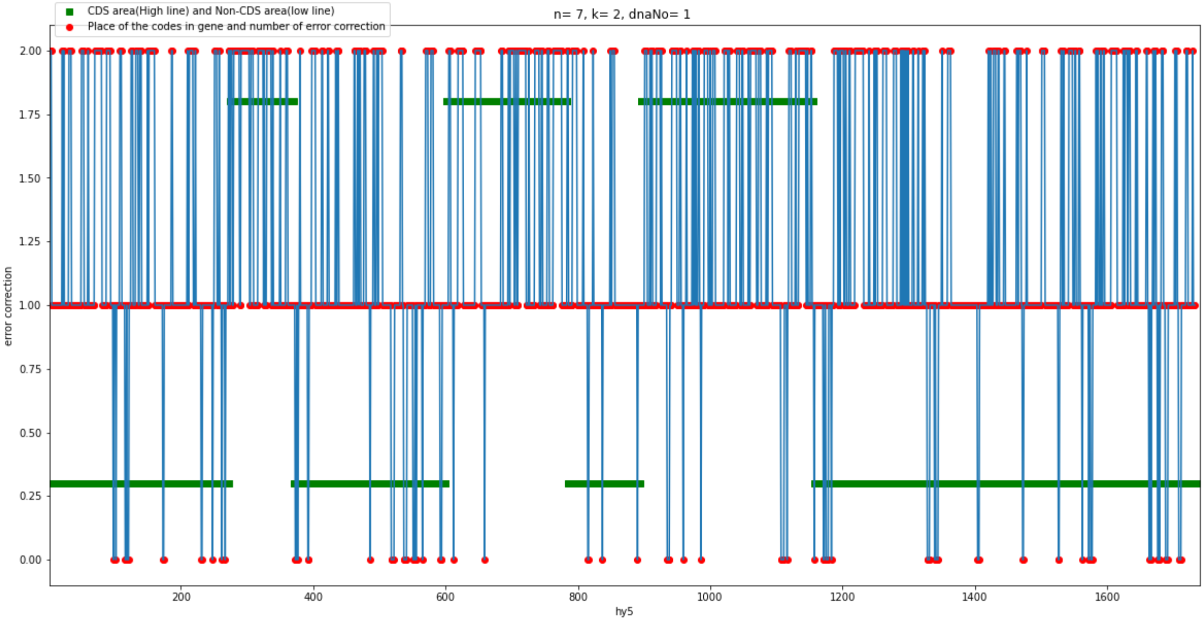


Figure 3. HY5

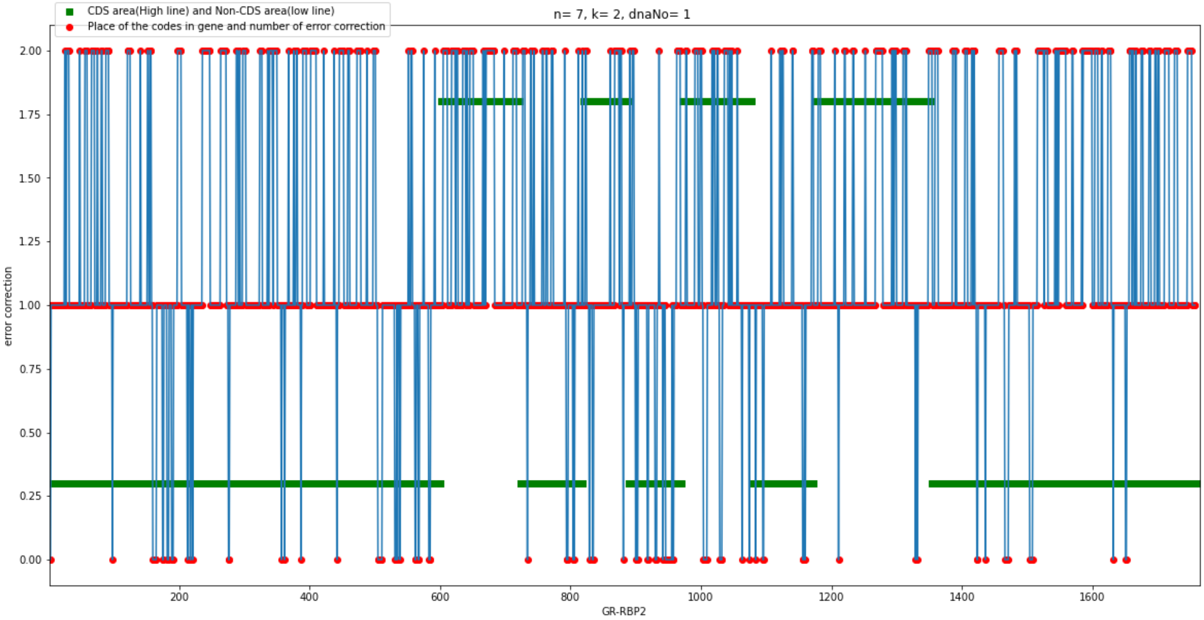


Figure 4. GR-RBP2

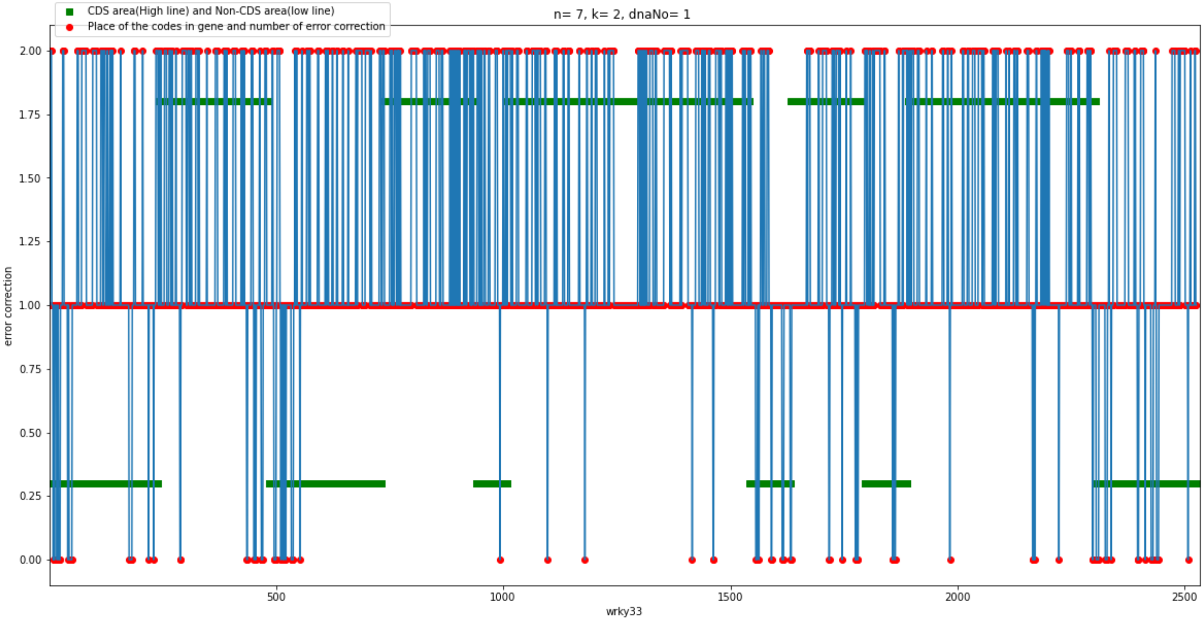


Figure 5. WRKY33

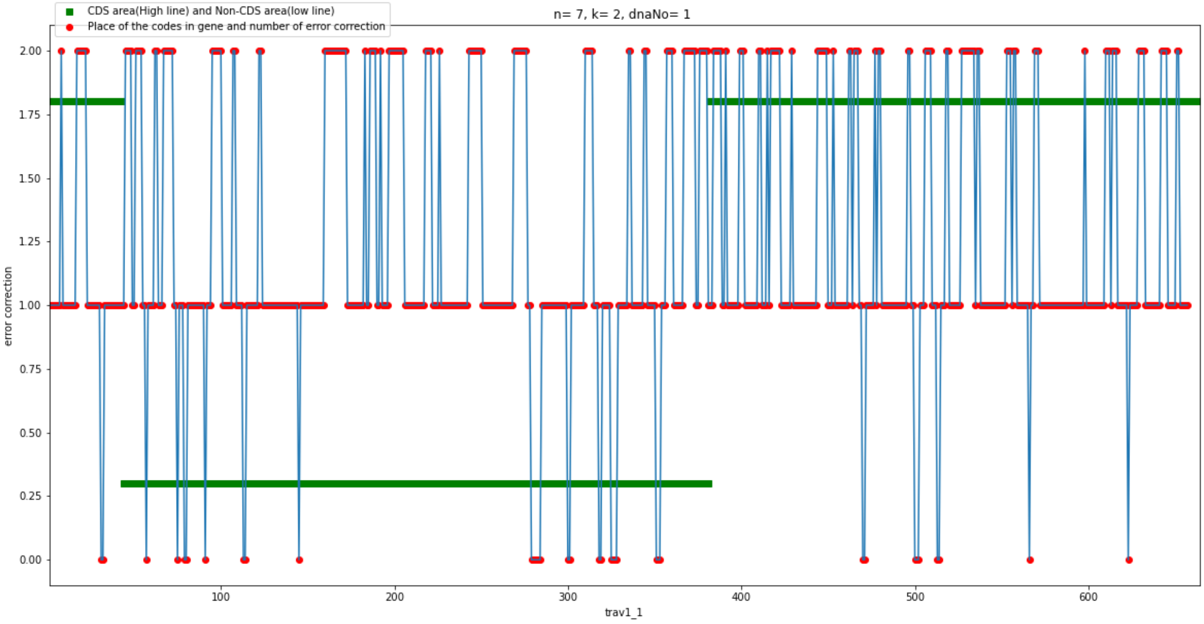


Figure 6. TRAV1-1

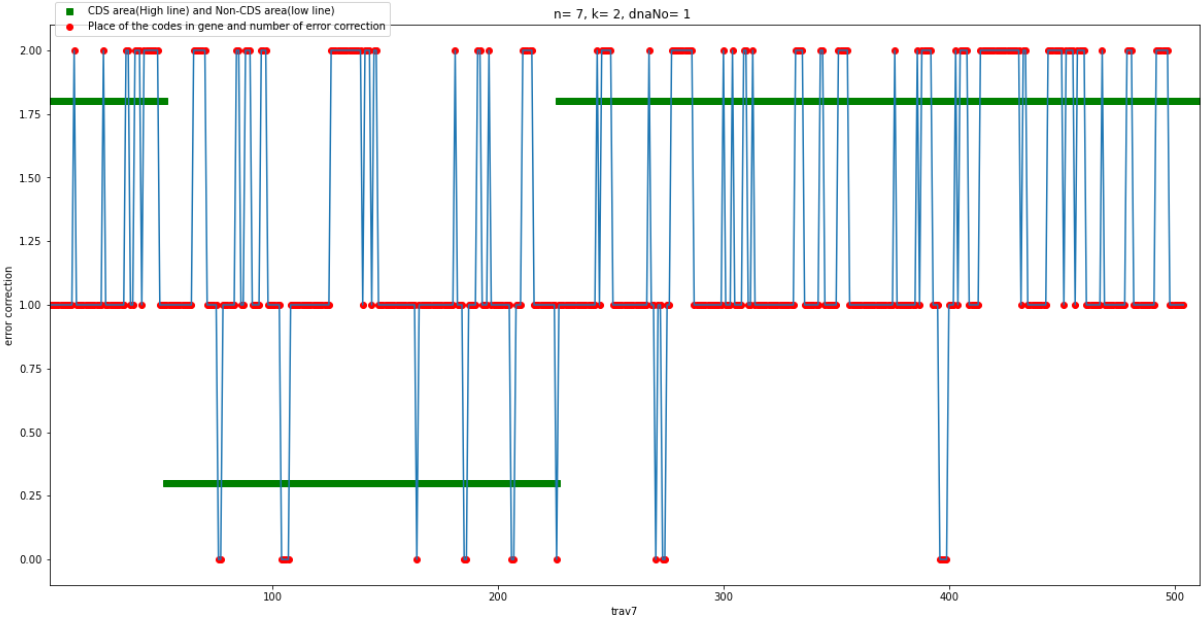


Figure 7. TRAV7

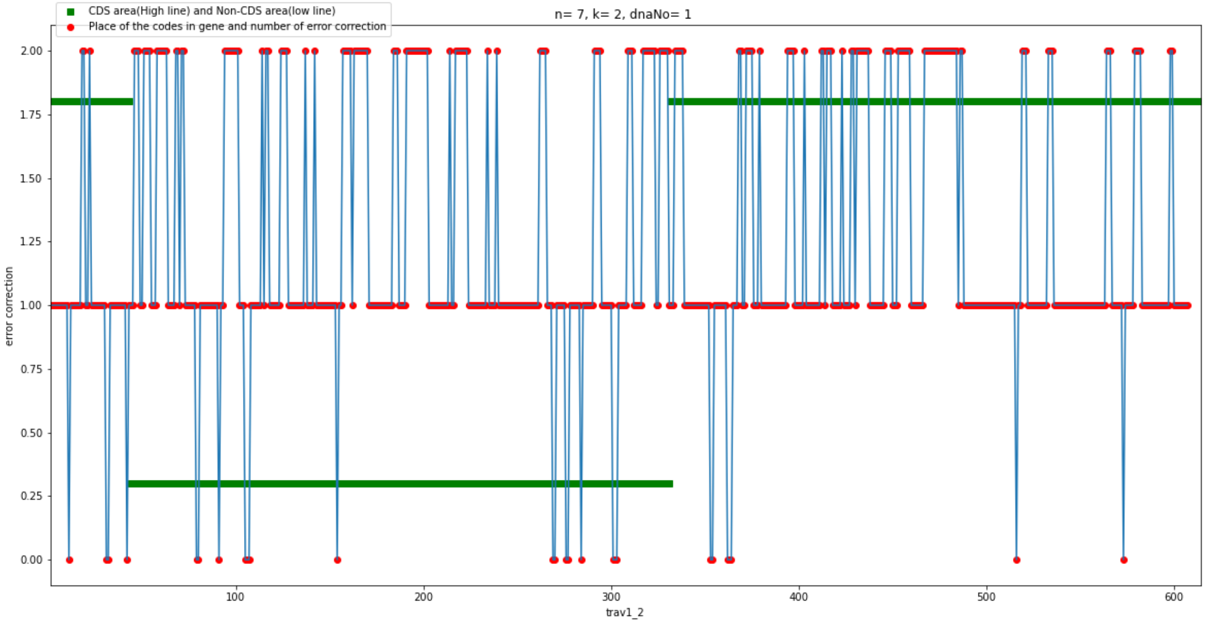


Figure 8. TRAV1-2

References

[1] Lichtenberg, J., Yilmaz, A., Welch, J. D., Kurz, K., Liang, X., Drews, F., Ecker, K., Lee, S. S., Geisler, M., Grotewold, E. and Welch, L. R., "The word landscape of the non-coding segments of the Arabidopsis thaliana genome", Bell Labs Tech. J 10(1).

[2] Forsdyke, D. R., "Are introns in-series error-detecting sequences?", Journal of Theoretical Biology, 93(4) (1981) : 861-866.

[3] Forsdyke, D. R., "Conservation of stem-loop potential in introns of snake venom phospholipase A2 genes: An application of FORS-D analysis", 12(6) (1995) : 1157 – 1165.

[4] Oztas, E. S., Siap, I., "Lifted polynomials over F-16 and their applications to DNA Codes", Filomat. 27 (2013) : 459-466.

[5] Abulraub, T., Ghrayeb A., Nian Zeng, X., "Construction of cyclic codes over $GF(4)$ for DNA computing", J. Franklin Inst. 343(4-5) (2006) : 448-457.

[6] Adleman, L. "Molecular computation of solutions to combinatorial problems", Science. 266 (5187) (1994) : 1021-1024.

[7] Bayram, A., Oztas, E.S., Siap, I., "Codes over $F\_4 + v\,F\_4$ and some DNA applications", Des. Codes Cryptogr. 80 (2) (2015): 379-393.

[8] Faria, L.C., Rocha, A.S., Kleinschmidt, J.H., Silva--Filho, M. C., Bim, E., Herai, R. H., Yamagishi, M. E., Palazzo, R. Jr., "Is a genome a codeword of an error--correcting code?", PloS one. 7 (5) e36644 (2012).

[9] Grassl, M., Bounds on the minimum distance of linear codes and quantum codes. http://www.codetables.de.

[10] Liebovitch, L.S., Tao, Y., Todorov, A.T., Levine, L., "Is there an error correcting code in the base sequence in DNA?", Biophys J. 71(3) (1996) : 1539-1544.

[11] Oztas, E.S., Siap, I., "On a generalization of lifted polynomials over finite fields and their applications to DNA codes" Int. J. Comput. Math. 92 (9) (2015) : 1976-1988.

[12] Oztas, E. S., Yildiz, B., Siap, I. "On DNA codes from a family of chain rings", J. Algebra Comb. Discrete Struct. Appl. 4 (1) (2017) : 93-102.

[13] Rosen, G. L.: "Examining Coding Structure and Reducdancy in DNA", IEEE Engineering in Medicine and Biology Magazine. 25 (2006) : 62-68.

[14] Siap, I., Abulraub, T., Ghrayeb, A., "Cyclic DNA codes over the ring $F_2[u]/(u^2-1)$ based on the deletion distance" J. Franklin Inst. 346 (8) (2009) : 731-740.

[15] Wong, G. K. and Passey, D. A., Huang, Y., Yang, Z. and Yu,J,, "Is "junk" DNA mostly intron DNA?", Genome research, 10 (11) (2000).

[16] NCBI, Database of The Consensus CDS (CCDS) project, https://www.ncbi.nlm.nih.gov/projects/CCDS.

[17] Bulut Yilgor, M. "Cyclic codes over the ring $F_2+uF_2+vF_2+v^2F_2$ and their applications to DNA codes", Phd thesis (2020).

[18] Bulut Yılgör, M., Gürsoy, F., Oztas, E.S. et al., "Cyclic codes over $F_2+uF_2+vF_2+v^2F_2$ with respect to the homogeneous weight and their applications to DNA codes" AAECC 32 (2021) : 621–636.

[19] Brandão, M. M. et al., "Ancient DNA sequence revealed by error-correcting codes" Sci. Rep. 5, 12051; doi: 10.1038/srep12051 (2015).

[20] Vajda, S. and Beglov, D., Wakefield, AE, Egbert, M. and Whitty, A., "Cryptic binding sites on proteins: definition, detection, and druggability", Curr Opin Chem Biol., 44 (2018).

[21] Gene [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; 2004 – [cited 2022 march 10]. Available from: https://www.ncbi.nlm.nih.gov/gene/.