

A novel method for determining the non-CDS region by using error-correcting codes

Elif Segah OZTAS¹ and Merve Bulut Yilgor²

¹Karamanoglu Mehmetbey University

²Altinbas Universitesi

March 30, 2022

Abstract

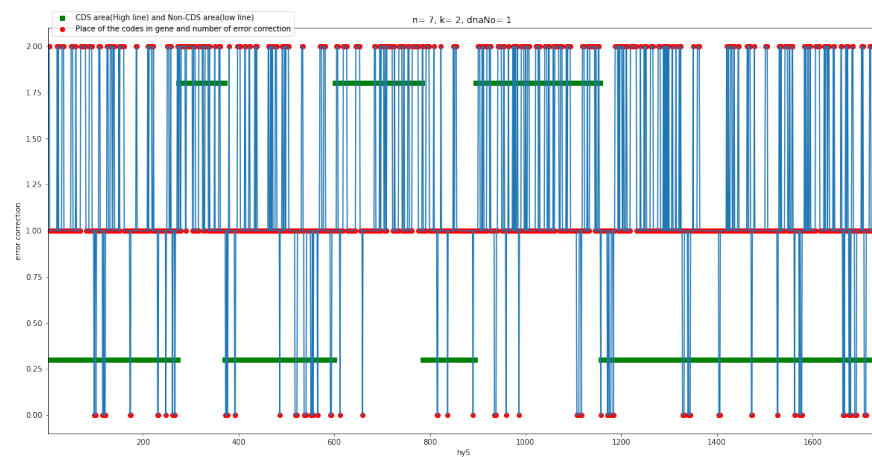
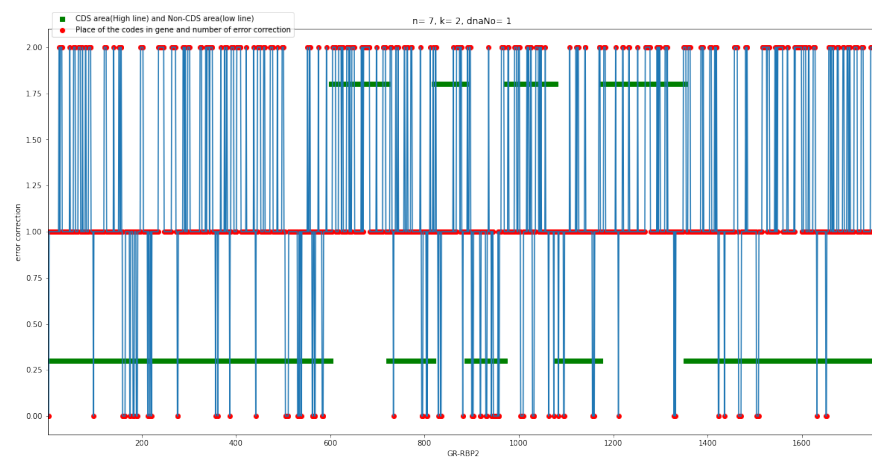
Our main motivation question is “Is there any relation between the non-coding region and useless error-correcting codes?”. Then we focused CDS and non-CDS areas instead of exon and intron, because CDS involves in process of synthesis a protein and is involved by exons. We get the data of the genes from NCBI (missing citation). 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 [Lichtenberg, Jens and Yilmaz, Alper and Welch, Joshua D and Kurz, Kyle and Liang, Xiaoyu and Drews, Frank and Ecker, Klaus and Lee, Stephen S. and Geisler, Matt and Grotewold, Erich ve Welch, Lonnie R.,The word landscape of the non-coding segments of the Arabidopsis thaliana genome,Bell Labs Tech. J, Volume 10, no 1].

Hosted file

KMUBAP4.pdf available at <https://authorea.com/users/467961/articles/561840-a-novel-method-for-determining-the-non-cds-region-by-using-error-correcting-codes>

Hosted file

KMUBAP4.tex available at <https://authorea.com/users/467961/articles/561840-a-novel-method-for-determining-the-non-cds-region-by-using-error-correcting-codes>



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

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

Converting to F₄ According to chosen dnaLabel 3

GENE = ATGGAGAAGATGCGGAGA.....
FGENE = 01αα0α00α01αβαα0α0.....

Calculate length of code:
n = n_f - k + 1 → n = 8 - 2 + 1 = 7

01αα0α00α01αβαα0α0...

Frame 1

Create a generator matrix for the code
for Frame 1

01αα0α00

Generator matrix for Frame 1

$\begin{bmatrix} 01\alpha\alpha0\alpha0 \\ 1\alpha\alpha0\alpha00 \end{bmatrix}$

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

Codewords of the code:

(1 1 α² 1 α 1 0)
(1 α α 0 α 0 0)
(1 α² 0 α α α 0)
(1 0 1 α² α α² 0)
(α 1 0 α² α² α² 0)
(α α 1 α α² α 0)
(α α² α² 0 α² 0 0)
(α 0 α 1 α² 1 0)
(α² 1 1 0 1 0 0)
(α² α 0 1 1 1 0)
(α² α² α α² 1 α² 0)
(α² 0 α² α 1 α 0)
(0 1 α α 0 α 0)
(0 α α² α² 0 α² 0)
(0 α² 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

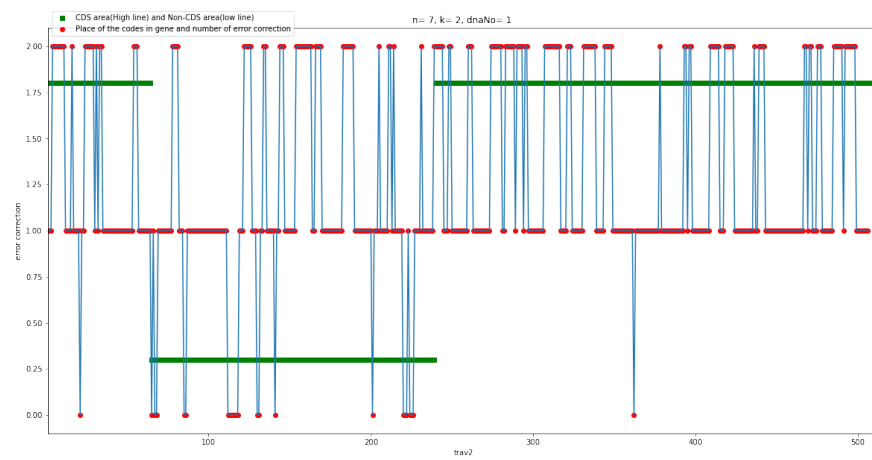
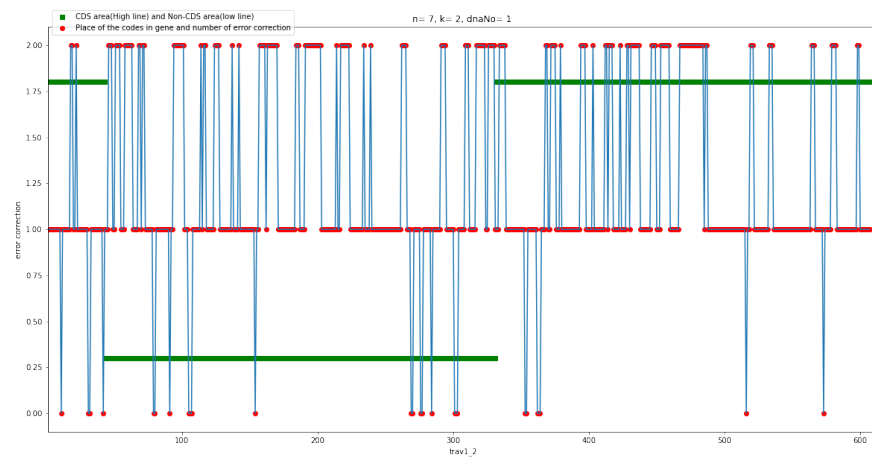
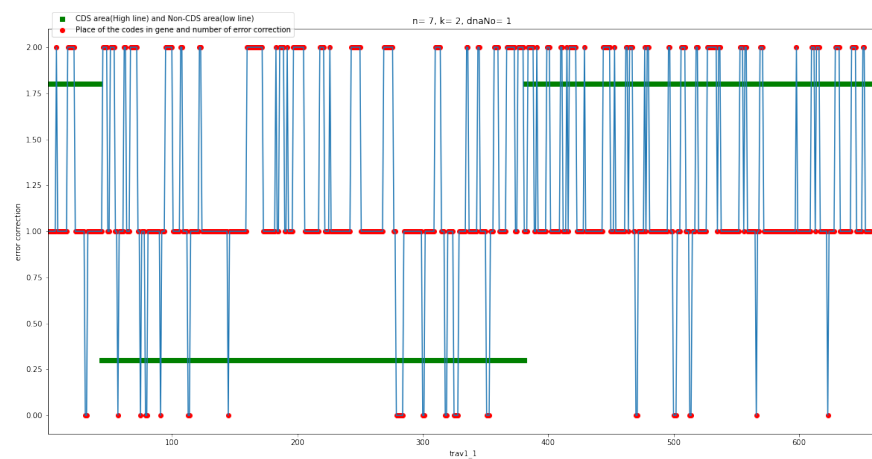
Process is continued in netx Frame

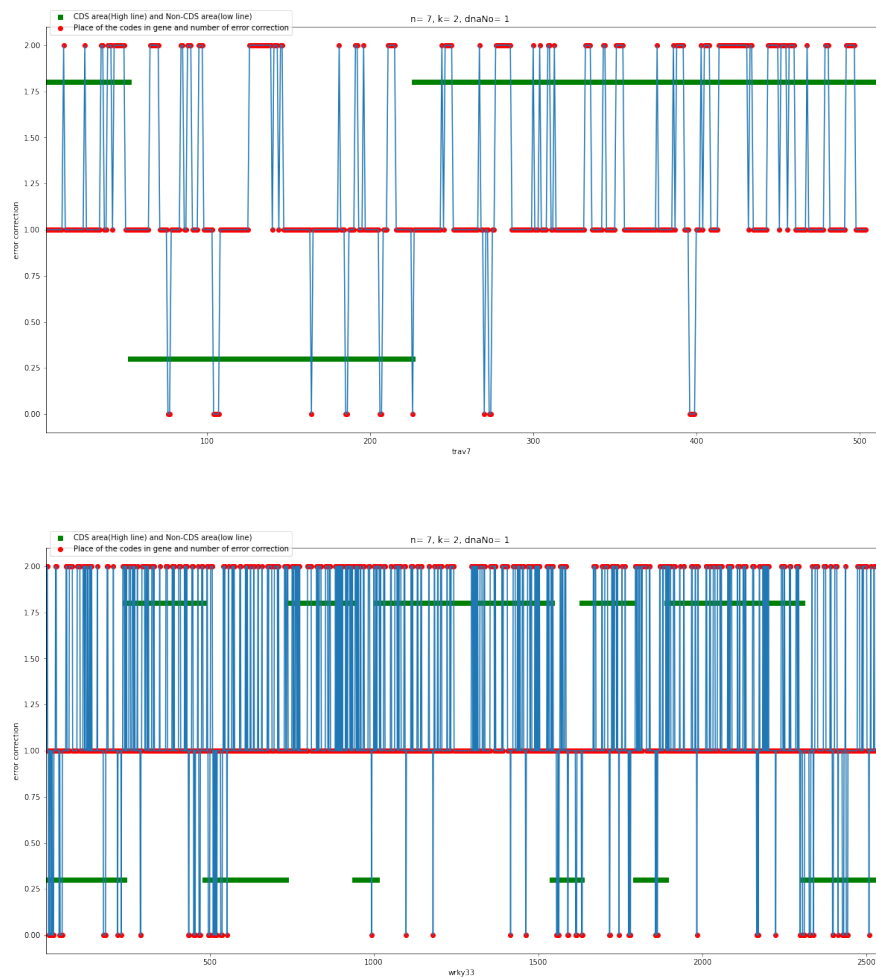
01αα0α00α01αβαα0α0...

Frame 2

⋮

Create the graphics by using list of
INDEX





References