AJMS
Asian Journal of Mathematical Sciences

RESEARCH ARTICLE

On the Covering Radius of DNA Code Over N

***P. Chella Pandian**

Department of Mathematics, Srimad Andavan Arts and Science College (A), Tiruchirappalli, Taminadu, India

Corresponding Email: chellapandianpc@gmail.com

Received: 12-04-2023; Revised: 28-04-2023; Accepted: 08-05-2023

ABSTRACT

In this paper, lower bound and upper bound on the covering radius of DNA codes over N with respect to lee distance are given. Also determine the covering radius of various Repetition DNA codes, Simplex DNA code Type α and Simplex DNA code Type β *and* bounds on the covering radius for MacDonald DNA codes of both types over N.

Keyword: DNA Code, Finite Ring, Covering Radius, Simplex Codes.

(2010) Mathematical Subject Classification: 94B25, 94B05, 11H31, 11H71.

INTRODUCTION

In, DNA is found naturally as a double stranded molecule, with a form similar to a twisted ladder. The backbone of the DNA helix is an alternating chain of sugars and phosphates, while the association between the two strands is variant combinations of the four nitrogenous bases adenine (A), thymine (T), guanine (G) and cytosine (C). The two ends of the strand are distinct and are conventionally denoted as 3' end and 5' end. Two strands of DNA can form (under suitable conditions) a double strand if the respective bases are Watson- Crick[17] complements of each other - A matches with T and C matches with G, also 3' end matches with 5' end.

The problem of designing DNA codes (sets of words of fixed length n over the alphabets $\{ A, C, G, T \}$ that satisfy certain combinatorial constraints has applications for reliably storing and retrieving information in synthetic DNA strands. These codes can be used in particular for DNA computing [1] or as molecular bar-codes.

There are many researchers doing research on code over finite rings. In particular, codes over Z_4 received much attention [2, 3, 4, 9, 11, 15, 16, 5]. The covering radiuses of binary linear codes were studied [4, 5]. Recently the covering radius of codes over Z_4 has been investigated with various distances [12]. In 1999, Sole et.al gave many upper and lower bounds on the covering radius of a code over Z_4 with different distances. In [14, 5], the covering radius of some particular codes over Z_4 have been investigated.

In this paper, investigate the covering radius of the Simplex DNA codes of both types and MacDonald DNA codes and repetition DNA codes over N. Also generalized some of the known bounds in [2]

Preliminary

Coding theory has several applications in Genetics and Bio engineering. The problem of designing DNA codes (sets of that words of fixed length *n* over the alphabet N = { *A, C, G, T}* that satisfy certain combinatorial constraints) has applications for reliably storing and retrieving information in synthetic DNA strands.

A DNA code of length *n* is a set of code words (x_1, $x_2, \ldots, x_n$) with $x_i \in \{A,C,G,T\}=N$ (representing the four nucleotides in DNA). Use a hat to denote the Watson-Crick complements of a nucleotide, so *A* matches with *T* and *C* matches with *G*.

The DNA codes are sets of words of fixed length *n* over the alphabet N *and* it follows the map $A \to 0, C \to 1, T \to 2$ and $G \to 3$. Therefore the problem of the DNA codes is corresponding to the problem of the Z_4-linear codes. These transpositions do not affect the GC-weight of the code word (the number of entries that are C or G). In my work, by using the above map in Z_4 with lee weight, so obtain the covering radius for repetition DNA codes .

Let $d = (d_1, d_2, \cdots, d_n) \in N^n$ and *n* be its length. Let *b* be an element of {*A, C, G, T* }.

For all $d = (d_1, d_2, \cdots, d_n) \in N^n$, define the weight of *d* at *b* tobe

$w_b(d) = |\{i | x_i = b\}|$.

A DNA linear code *C* of length *n* over N is an additive subgroup of N*n*. An element of *C* is called a DNA code word of *C* and a generator matrix of *C* is a matrix whose rows generate *C*. In [12], the Lee weight $w(x)$ of a vector *x* is 0 if $x_i = 0$; 1 if $x_i = 1$, 3 and 2 if $x_i = 2$. A linear Gray map φ from $N_4 \to Z_2^2$ is defined by $\varphi(x + 2y) = (y, x + y)$, for all $x + 2y$ *in* N. The image $\varphi(C)$, of a linear code *C* over N of length *n* by the Gray map is a binary code of length 2*n* with same cardinality [15].

Any DNA linear code *C* over N is equivalent to a code with Generator Matrix(*GM*) of the form

$$GM = \begin{bmatrix} I_{k_0} & A & B \\ 0 & 2I_{k_1} & 2D \end{bmatrix}$$

Where A, B and D are matrices over N. Then the DNA code C contains all DNA code words [v_0, v_1] GM, where v_0 is a vector of length k_1 over N and v_1 is a vector of length k_2 over Z_2. Thus C contains a total of $4^{k1}\ 2^{k2}$ code words. The parameters of C are given [n, $4^{k1}2^{k2}$, d], where d represents the minimum lee distance of C.

A DNA linear code C over N of length n, 2-dimension k, minimum lee distance d is called an [n, k, d_l] or simply an [n, k, d] code. In this paper, define the covering radius of dna codes over N with respect to lee distance and in particular study the covering radius of Simplex DNA codes of type α and type β namely, S_α and S_β and their MacDonald DNA codes and repetition DNA codes over N. Section 2 contains basic results for the covering radius of DNA codes over N. Section 3 determines the covering radius of different types of repetition DNA codes. Section 4 determines the covering radius of Simplex DNA codes and finally section 5 determines the bounds on the covering radius of MacDonald DNA codes.

Covering Radius of Repetition DNA Codes

$$r_d(C) = \max_{u \in \mathbb{N}^n} \left\{ \min_{c \in C} \{d(c,u)\} \right\}$$

Let d be a lee distance of a DNA code C over N. Thus, the covering radius of C:

The following result of Mattson [6] is useful for computing covering radius of codes over rings generalized easily from codes over finite fields.

Proposition 3.1 If C_0 and C_1 are codes over R generated by matrices GM_0 and GM_1 respectively and if C is the code generated by

$$GM = \left[\begin{array}{c|c} 0 & GM_1 \\ \hline GM_0 & A \end{array}\right]$$

then $r_d(C) \leq r_d(C_0) + r_d(C_1)$ and the covering radius of D (concatenation of C_0 and C_1) satisfy the following $r_d(D) \geq r_d(C_0) + r_d(C_1)$, for all distances d over N.

A q-ary repetition code C over a finite field $F_q = \{\alpha_0 = 0, \alpha_1 = 1, \alpha_2, \alpha_3, \ldots, \alpha_{q-1}\}$ is an [n, 1, n] code $C = \{\bar{\alpha} \mid \alpha \in (F_q)\}$ where $\bar{\alpha} = (\alpha, \alpha, \cdots, \alpha)$. The covering radius of C is $\lceil \frac{n(q-1)}{q} \rceil$ [11]. Using this, it can be seen easily that the covering radius of block of size n repetition code [n(q-1),1,n(q-1)] generated by

$$GM = [\overbrace{11\cdots1}^{n}\overbrace{\alpha_2\alpha_2\cdots\alpha_2}^{n}\overbrace{\alpha_3\alpha_3\cdots\alpha_3}^{n}\cdots\overbrace{\alpha_{q-1}\alpha_{q-1}\cdots\alpha_{q-1}}^{n}]$$

is $\lceil \frac{n(q-1)^2}{q} \rceil$, since it will be equivalent to a repetition code of length $(q-1)n$.

Consider the repetition dna code over N. There are two types of them of length n vi

1. Cytosine repetition code C_β : [n, 1, n] generated by $GM_\beta = \left[\overbrace{CC\cdots C}^{n}\right]$

2. Thymine repetition code C_α : (n, 2, 2n) generated by $GM\alpha = \left[\overbrace{TT\cdots T}^{n}\right]$

Theorem 3.2 Let C_β and C_α be the dna code of type β and α type in generator matrices GM_β and GM_α . Then, $\lfloor \frac{n}{2} \rfloor \leq r(C_\alpha) \leq n$ and $r(C_\beta) = n$.

Proof.

Let $x=\overbrace{AA\cdots A}^{\lfloor \frac{n}{2} \rfloor}\overbrace{TT\cdots T}^{\lceil \frac{n}{2} \rceil}$ and the code of C = {A A · · · A, T T · · · T } is generated by [T T …. T] is an [n, 1, 2n] code. Then, d(x, A A … A) = wt(x - A A … A) = $\lfloor \frac{n}{2} \rfloor$ and d(x, T T...T) = wt(x - T T … T) = $\lceil \frac{n}{2} \rceil$. Therefore, d(x, C_α) = min{ $\lfloor \frac{n}{2} \rfloor$, $\lceil \frac{n}{2} \rceil$}. Thus, by definition of covering radius

$$r(C\alpha) \geqslant \lceil \frac{n}{2} \rceil \qquad (3.1)$$

Let x be any word in N^n . Let us take x has ω_0 coordinates as 0's, ω_1 , coordinates as 1's, ω_2 coordinates as 2's, ω_3 coordinates as 3's, then $\omega_0 + \omega_1 + \omega_2 + \omega_3 = n$. Since C_α = {AA · · · A, T T · · · T} and lee weight of N : A is 0, C and G is 1 and T is 2.

Therefore, d(x, AA · · · A) = $n - \omega_0 + \omega_2$ and

d(x, T T · · · T) = $n - \omega_2 + \omega_0$.

Thus d(x, C_α) = min {$n - \omega_0 + \omega_2$, $n - \omega_2 + \omega_0$ } and hence, d (x, C_α) ≤ n = n. (3.2)

Hence, from the Equation (3.1) and (3.2), so $\lceil \frac{n}{2} \rceil \leq r(C_\alpha) \leq n$. Now, obtain the covering radius of C_β covering with respect to the lee weight.

Then d(x, AA · · · A) = $n - \omega_0 + \omega_2$,

d(x, CC · · · C) = $n - \omega_1 + \omega_3$,

d(x, T T · · · T) = $n - \omega_2 + \omega_0$ and

d(x, GG · · · G) = $n - \omega_3 + \omega_1$, for any $x \in N^n$.

This implies d(x, C_β) = min{$n - \omega_0 + \omega_2$, $n - \omega_1 + \omega_3$, $n - \omega_2 + \omega_0$, $n - \omega_3 + \omega_1$ } = n and hence r(C_β) ≤ n.

Let $x=\overbrace{AA\cdots A}^{t}\overbrace{CC\cdots C}^{t}\overbrace{TT\cdots T}^{t}\overbrace{GG\cdots G}^{n-3t}$, where $t=\lfloor \frac{n}{4} \rfloor$. Then d(x, AA · · · A) = n, d(x, CC · · · C) = 2n − 4t, d(x, T T · · · T) = n and d(x, GG · · · G) = 4t. Therefore r(C_β) ≥ min{2n, 2n − 4t, 4t} ≥ n.

Block Repetition Code

Let $GM=\left[\overbrace{CC\cdots C}^{n}\overbrace{TT\cdots T}^{n}\overbrace{GG\cdots G}^{n}\right]$ be a generator matrix of N in each block of repetition code length is n. Then, the parameters of Block Repetition Code(BRC) is [3n, 1, 4n]. The code of BRC = {c_0 = A · · · AA · · · AA · · · A, c_1 = C · · · C T · · · T G · · · G, c_2 = T · · · T A · · · AT · · · T, c_3 = G · · · GT · · · T C · · · C}, dimension of BRC is 1 and lee weight is 4n. Note that, the block repetition code has constant lee weight is 4n. Obtain, the following

Theorem 3.3 r(BRC^{3n}) = 3n.

Proof.

Let x = AA $\cdots$ A $\in$ N^{3n} . Then, d(x, BRC^{3n}) = 3n. Hence by definition,
r(BRC^{3n}) $\geq$ 3n.
Let x = (u|v|w) $\in$ N^{3n} with u, v and w have compositions (r_0 , r_1 , r_2 , r_3) ,(s_0 , s_1 , s_2 , s_3)
and (t_0 , t_1 , t_2 , t_3) respectively such that $\sum_{i=0}^{3} r_i = \sum_{i=0}^{3} s_i = n = \sum_{i=0}^{3} t_i$.
Then,

$d(x, c_0) = 3n - r_0 + r_2 - s_0 + s_2 - t_0 + t_2$,
$d(x, c_1) = 3n - r_1 + r_3 - s_2 + s_0 - t_3 + t_1$,
$d(x, c_2) = 3n - r_2 + r_0 - s_0 + s_2 - t_2 + t_0$ and
$d(x, c_3) = 3n - r_3 + r_1 - s_2 + s_0 - t_1 + t_3$.

Thus,

$d(x, BRC^{3n}) = \min\{3n - r_0 + r_2 - s_0 + s_2 - t_0 + t_2 , 3n - r_1 + r_3 - s_2 + s_0 - t_3 + t_1 ,$
$3n - r_2 + r_0 - s_0 + s_2 - t_2 + t_0 , 3n - r_3 + r_1 - s_2 + s_0 - t_1 + t_3.\}$
$d(x, BRC^{3n}) \leq 3n$ and hence, $r(BRC^{3n}) \leq 3n$.

Define a two block repetition dna code over N of each of length is n and the parameters of two block repetition cod BRC^{2n} : [2n, 1, 2n] is generated by
$GM = \left[\overbrace{CC\cdots C}^{n}\overbrace{TT\cdots T}^{n}\right]$. Use the above and obtain a following

Theorem 3.4 $r(BRC^{2n}) = 2n$.

Let $GM = \left[\overbrace{CC\cdots C}^{m}\overbrace{TT\cdots T}^{n}\right]$ be the generalized generator matrix for two different block repetition dna code of length are m and n respectively. In the parameters of two different block repetition code (BRC^{m+n}) are [m + n, 1, min{2m, m + n}] and Theorem 3.4 can be easily generalized for two different length using similar arguments to the following.

Theorem 3.5 $r(BRC^{m+n}) = m + n$.

Simplex DNA Code of Type α and Type β over N

In ref.[3] has been studied of Quaternary simplex codes of type α and type β. Type α Simplex code S_k^{α} is a linear dna code over N with parameters[4^k, k] and an inductive generator matrix given by

$$GM_k^{\alpha} = \left[\begin{array}{c|c|c|c} A\cdots A & C\cdots C & T\cdots T & G\cdots G \\ \hline GM_{k-1}^{\alpha} & GM_{k-1}^{\alpha} & GM_{k-1}^{\alpha} & GM_{k-1}^{\alpha} \end{array}\right] \quad (4.1)$$

with GM_1^{α} = [A(0) C(1) T (2) G(3)]. Type simplex code S_k^{β} is a punctured version of S_k^{α} with parameters [2^{k-1} , (2^k − 1), k] and an inductive generator matrix given by

$$GM_2^{\beta} = \left[\begin{array}{cccc|c|c} C & C & C & C & A & T \\ \hline A & C & T & G & C & C \end{array}\right] \quad (4.2)$$

$$GM_k^{\beta} = \left[\begin{array}{c|c|c} CC\cdots C & AA\cdots A & TT\cdots T \\ \hline GM_{k-1}^{\alpha} & GM_{k-1}^{\beta} & GM_{k-1}^{\beta} \end{array}\right] \quad (4.3)$$

and for k > 2, where GM_{k-1}^{α} is the generator matrix of S_{k-1}^{α} . For details the reader is ref. to [3]. Type α code with minimum lee weight is 4.

Theorem 4.1 $r(S_k^{\alpha}) \leq 2^{2k} + 1$.

Proof.

Let $x = CC \cdots C \in N^n$. By equation(4.1), the result of Mattson for finite rings and using Theorem 3.3, then $r(S_k^{\alpha}) \leq r(S_{k-1}^{\alpha}) + < \overbrace{CC\cdots C}^{2^{2(k-1)}}\overbrace{TT\cdots T}^{2^{2(k-1)}}\overbrace{GG\cdots G}^{2^{2(k-1)}} >$

$= r(S_{k-1}^{\alpha}) + 3.2^{2(k-1)}$

$= 3.2^{2(k-1)} + 3.2^{2(k-2)} + 3.2^{2(k-3)} + \ldots + 3.2^{2.1} + r(S_1^{\alpha})$

$r(S_k^{\alpha}) \leq 2^{2k} + 1$ (since $r(S_1^{\alpha}) = 5$) .

Theorem 4.2 $r(S_k^{\beta}) \leq 2k(2^k - 1) - 1$.

Proof.

By equation (4.3), Proposition 3.1 and Theorem 3.5, thus

$r(S_k^{\beta}) \leq r(S_{k-1}^{\beta}) + < \overbrace{CC\cdots C}^{4^{(k-1)}}\overbrace{TT\cdots T}^{2^{(2k-3)}-2^{(k-2)}} >$

$= r(S_{k-1}^{\beta}) + 2^{(2k-2)} + 2^{(2k-3)} - 2^{(k-2)}$

$\leq 2(2^{(2k-2)} + 2^{(2k-4)} + \ldots + 2^4) + 2(2^{(2k-3)} + 2^{(2k-5)} + \ldots + 2^3) -$

$2(2^{(k-2)} + 2^{(k-3)} + \ldots + 2) + r(S_2^{\beta})$

$r(S_k^{\beta}) \leq 2^{k-1}(2^k - 1) - 1$, (since $r(S_2^{\beta}) = 5$).

MacDonald DNA Codes of Type α and β Over N

The q-ary MacDonald code $M_{k,t}(q)$ over the finite field F_q is a unique $[(q^k - q^t)/(q-1), k, q^{k-1} - q^{t-1}]$ code in which every non-zero codeword has weight either q^{k-1} or $q^{k-1} - q^{t-1}$ [10]. In [13], he studied the covering radius of MacDonald codes over a finite field. In fact, he has given many exact values for smaller dimension. In [8], authors have defined the MacDonald codes over a ring using the generator matrices of simplex codes. For $2 \leq t \leq k-1$, let $GM_{k,t}^{\alpha}$ be the matrix obtained from GM_k^{α} by deleting columns corresponding to the columns of GM_t^{α} .

That is, $GM_{k,t}^{\alpha} = [GM_{k,}^{\alpha} \setminus 0 / (GM_{t,}^{\alpha})]$ (5.1)

and let $GM_{k,t}^{\beta}$ be the matrix obtained from GM_k^{β} by deleting columns corresponding to the columns of GM_t^{β} .

That is, $GM_{k,t}^{\beta} = [GM_k^{\beta} \setminus 0 / (GM_t^{\beta})]$ (5.2)

where [A\B] denotes the matrix obtained from the matrix A by deleting the columns of the matrix B and 0 is a $(k-t) \times 2^{2t}((k-t) \times 2^{t-1}(2t-1))$. The code generated by the matrix $GM_{k,t}^{\alpha}$ is called code of type α and the code generated by the matrix $GM_{k,t}^{\beta}$ is called Macdonald code of type β. The type α code is denoted by $M_{k,t}^{\alpha}$ and the type β code is denoted by $M_{k,t}^{\beta}$. The $M_{k,t}^{\alpha}$ code is $[4^k - 4^t, k]$ code over N and $M_{k,t}^{\beta}$ is a $[(2^{k-1} - 2^{t-1})(2^k + 2^t - 1), k]$ code over N. In fact, these codes

are punctured code of S_k^{α} and S_k^{β} respectively. Next Theorem gives a basic bound on the covering radius of above Macdonald codes.

Theorem 5.1 $r(M_{k,t}^{\alpha}) \leq 2^{2k} - 2^{2r} + r(M_{r,t}^{\alpha})$, for $t < r \leq k$.

Proof.

In equation(5.1), Proposition 3.1 and Theorem 3.3, thus

$$r(M_{k,t}^{\alpha}) \leq r(S_{k-1}^{\alpha}) + < \overbrace{CC\cdots C}^{2^{2(k-1)}}\ \overbrace{TT\cdots T}^{2^{2(k-1)}}\ \overbrace{GG\cdots G}^{2^{2(k-1)}} >$$

$$= 3.4^{k-1} + r(M_{k-1,t}^{\alpha}), \text{ for } k \geq r > t.$$

$$\leq 3.4^{k-1} + 3.4^{k-2} + \cdots + 3.4^{r} + r(M_{r,t}^{\alpha}), \text{ for } k \geq r > t.$$

$$r(M_{k,t}^{\alpha}) \leq 2^{2k} - 2^{2r} + r(M_{r,t}^{\alpha}), \text{ for } k \geq r > t.$$

Theorem 5.2 $r(M_{k,t}^{\beta}) \leq 2^{(k-1)}(2^k - 1) + 2^{(r-1)}(1 - 2^r) + r(M_{r,t}^{\beta})$, for $t < r \leq k$.

Proof.

Using Proposition 3.1, Theorem 3.5 and in equation(5.2), obtain

$$r(M_{k,t}^{\beta}) \leq r(M_{k-1,t}^{\beta}) + < \overbrace{CC\cdots C}^{4^{(k-1)}}\ \overbrace{TT\cdots T}^{2^{(2k-3)}-2^{(k-2)}} >$$

$$r(M_{k,t}^{\beta}) \leq 2^{(k-1)}(2^k - 1) + 2^{(r-1)}(1 - 2^r) + r(M_{r,t}^{\beta}), \text{ for } t < r \leq k.$$

REFERENCES

1. Adleman M. L., Molecular computation of solutions to combinatorial problems, Science 265, 1021-1024(1994).

2. Aoki T., Gaborit P., Harada M., Ozeki M. and Sol'e P., On the covering *radius of* Z_4 *Codes and their lattices,* IEEE Trans. Inform. Theory, vol. 45, no. 6, pp. 2162-2168(1999).

3. Bhandari M. C., Gupta M. K. and Lal A. K., On Z 4 Simplex codes and their gray images, Applied Algebra, Algebraic Algorithms and Error- Correcting Codes, AAECC- 13, Lecture Notes in Computer Science 1719, 170-180(1999).

4. Bonnecaze A., Sol'e, P., Bachoc, C. and Mourrain B., Type II codes over Z_4 , IEEE Trans. Inform. Theory, 43, 969-976(1997).

5. Chella Pandian P., On the Covering Radius of Some Codes Over R, International Journal of Research in Applied, Natural and Social Sciences, Vol. 2, No. 1, pp.61-70(2014).

6. Cohen G. D., Karpovsky, M. G., Mattson, H. F. and Schatz J. R., Covering radius- Survey and recent results, IEEE Trans. Inform. Theory, vol.31, no. 3, pp. 328-343(1985).

7. Cohen C., Lobstein, A. and Sloane N. J. A., Further Results on the Covering Radius of codes, IEEE Trans. Inform. Theory, vol. 32, no. 5, pp. 680-694(1986).

8. Colbourn C. J. and Gupta M. K., On quaternary MacDonald codes, Proc. Information Technology:Coding and Computing (ITCC), pp. 212- 215 April(2003).

9. Conway J. H. and Sloane N. J. A., 1993, Self-dual codes over the integers modulo 4, Journal of Combin. Theory Ser. A 62, 30-45(1993).

10. Dodunekov S. and Simonis, J., Codes and projective multisets, The Electronic Journal of Communications 5 R37(1998).

11. Dougherty S. T., Harada M. and Sol'e P., Shadow codes over Z_4 Finite Fields and Their Appl. (to appear).

12. Gupta M. K., David G. Glynn and Aaron Gulliver T., On Senary Simplex Codes. Lecture Notes in Computer Science.

13. Durairajan C., On Covering Codes and Covering Radius of Some Optimal Codes, Ph. D. Thesis, Department of Mathematics, IIT Kanpur (1996).

14. Gupta M. K and Durairajan C., On the Covering Radius of some Modular Codes. Journal of Advances in Mathematics of Computations 8(2), 9, 2014.

15. Hammons A. R., Kumar P. V., Calderbank A. R., Sloane N. J. A. and Sol'e P., The Z_4 - linearity of kerdock, preparata, goethals, and related codes, IEEE Trans. Inform. Theory, 40, 301-319 (1994).

16. Harada M., New extremal Type II codes over Z 4 . Des. Codes and Cryptogr. 13, 271-284 (1998).

17. Watson D. J. and Crick C. H. F., A structure for deoxyribose nucleic acid, Nature, vol. 25, pp. 737-738, 1953.