The Constructions of DNA Codes from Linear Self-Dual Codes over Z$_4$

B. Feng, S.S. Bai, B.Y. Chen, X.N. Zhou
College of science
Anhui University of Science and Technology
Huainan Anhui, China

Abstract--**It is possible to construct DNA codes from classical linear codes. However, it is difficult to determine the GC weight enumerator. In this paper, we use the Z_4 -linear self-dual codes K_{4m} to construct DNA codes. K_{4m} have some good properties, so DNA codes can be easily constructed. Moreover, GC weight enumerator of the DNA codes can be easily determined.**

Keywords-self-dual codes; Z_4 codes; DNA codes

I INTRODUCTION

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[1] 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 [2] or as molecular bar-codes.

In this paper,we construct DNA codes from K_{4m} which are linear self-dual codes over Z_4 . K_{4m} have some particular properties, so the constructed DNA codes from them also have these properties, i.e.the GC-weight enumerator of the constructed DNA codes can be easily determined.

The paper is organized as follows. In Section 2, we recall basic notions for linear self-dual codes over Z_4 and DNA codes; In Section 3, we translate the linear self-dual codes K_{4m} into DNA codes and take Z_4 and Z_8 for example; Also in Section 3, we detail the process of our constructions of DNA codes and finally give their advantages and tables of the constructed DNA codes.

II PRELIMINARIES

A. Linear Self-Dual Codes Over Z_4

Let [3] $Z_4(=\{0,1,2,3\})$ denote the ring of integers modulo four. A Z_4 -code C of length n is a Z_4 -submodule of Z_4^n .We define an inner product on Z_4^n by $x \cdot y$ = $x_1y_1+\cdots+x_ny_n$ (mod4) where $x=(x_1,\cdots,x_n)$ and $y=(y_1\cdots,y_n)$ .The[4] dual code $C^{\perp}$ of C is defined as $\{x \in Z_4^n \mid x \cdot y = 0$ for all $y \in$ C$\}$ under the standard inner product $x \cdot y$. A code C is a self-dual code if $C^{\perp}$=C.

Theorem 2.1 Any code over Z_4 is permutation-equivalent to a code C with a generator matrix of the form

$$\begin{pmatrix} I_{k_1} & A & B_1+2B_2 \\ O & 2I_{k_2} & 2D \end{pmatrix} \tag{1}$$

Where A, B_1, B_2 , and D are (1, 0)-matrices. We say that a code with the generator matrix has type $4^{k_1}2^{k_2}$.

Definition 2.1[5] Let C be a Z_4 -code and n be its length. Let a be an element of Z_4 ,i.e. a =0,1,2 or 3.For all $x=(x_1,\cdots,x_n)\in Z_4^n$,define the weight of x at a to be $w_a(x)=\left|\{i \mid x_i=a\}\right|$.

Definition 2.2[6] the complete weight enumerator (or c.w.e.) of C is $\mathrm{cwe}_C(\mathrm{W,X,Y,Z})=\sum_{a\in C}\mathrm{W}^{n_0(a)}\mathrm{X}^{n_1(a)}\mathrm{Y}^{n_2(a)}\mathrm{Z}^{n_3(a)}$ where $n_i(a)$ is the number of components of a that are congruent to i modulo 4.

Klemm[7] has studied the conditions satisfied by the complete weight enumerator of a self-dual code over Z_4 .The MacWilliams transform determines a group of substitutions, each of which fixes the complete weight enumerator.

Example 2.1 Let K_4 denote the Z_4 -linear code with generator matrix

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 2 & 0 & 2 \\ 0 & 0 & 2 & 2 \end{pmatrix} \tag{2}$$

K_4 is of type $4^1 2^2$ .Therefore $|K_4| = 16$. $K_4^{\perp}$ is also of type $4^1 2^2$ with respect to theorem 2.1.Therefore $|K_4^{\perp}| = 16$. It is obvious that any two rows of (2), distinct or not, are orthogonal. Therefore $K_4 \subseteq K_4^{\perp}$ .Hence $K_4 = K_4^{\perp}$ and K_4 is a self-dual code.

Let K_4 be the Z_4 -linear codes. Its 16 codewords and the numbers $w_a(c)$ of them, where $a \in Z_4$ and $c \in K_4$, are shown in the following table.

TABLE I

K_4	w_0	w_1	w_2	w_3
(0000)	4	0	0	0
(1111)	0	4	0	0
(2222)	0	0	4	0
(3333)	0	0	0	4
(0202)	2	0	2	0
(1313)	0	2	0	2
(2020)	2	0	2	0
(3131)	0	2	0	2
(0022)	2	0	2	0
(1133)	0	2	0	2
(2200)	2	0	2	0
(3311)	0	2	0	2
(0220)	2	0	2	0
(1331)	0	2	0	2
(2002)	2	0	2	0
(3113)	0	2	0	2

Then the complete weight enumerator of C is defined to be the homogeneous polynomial of degree n in four indeterminates X_0, X_1, X_2 and X_3

$$\mathrm{cwe}_{K_4}(X_0,X_1,X_2,X_3)=\frac{1}{2}((X_0+X_2)^4+(X_1+X_3)^4+(X_0-X_1)^4+(X_1-X_3)^4) \quad (3)$$

B. *DNA Codes*

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. Every DNA molecule consists of two complementary strands which are sequences of four different nucleotide bases. These are called adenine (A), cytosine (C), guanine (G) and thymine (T). The two ends of the strand are distinct and are conventionally denoted as 3' end and 5' end. A matches with T and C matches with G, also 3' end matches with 5' end.

Coding theory has several applications in Genetics and Bioengineering. The problem of designing DNA codes (sets of words of fixed length n over the alphabet $\{A, C, G, T\}$ 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 codewords $(x_1, \cdots, x_n)$ with $x_i \in \{A, C, G, \mathrm{T}\}$ (representing the four nucleotides in DNA). We use a hat to denote the Watson-Crick complement of a nucleotide, so A matches with T and C matches with G.

III CONSTRUCTIONS FOR DNA CODES

Because the DNA codes are sets of words of fixed length n over the alphabet $\{A, C, G, T\}$,we can follow the map A→0, C→1,T→2, and G→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 codeword (the number of entries that are C or G).In our work, by using the above map, we construct DNA codes from the self-dual Z_4 -codes K_{4m} .

Definition 3.1[8] Let K_{4m} (m $\geq$ 1) denote the Z_4 -linear code with generator matrix

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 & 1 \\ 0 & 2 & 0 & \cdots & 0 & 2 \\ 0 & 0 & 2 & \cdots & 0 & 2 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 2 & 2 \end{pmatrix} \quad (4)$$

Which is a $(4m-1) \times 4m$ matrix. K_{4m} (m $\geq$ 1) was introduced by Klemn(1989).

Definition 3.2 Let $d=(d_1,\cdots,d_n) \in \{\mathrm{A,C,T,G}\}^n$ and n be its length.Let b be an element of $\{\mathrm{A,G,T,C}\}$ .For all $d=(d_1,\cdots,d_n) \in \{\mathrm{A,C,T,G}\}^n$, define the weight of d at b to be $w_b(d) = |\{i | x_i = b\}|$.

Theorem 3.1 Let C be the $K_{4m}(m \geq 1)$ codes. Then there exists DNA codes of length $4m$, with 2^{4m} codewords, and GC weight enumerator is

$$W_{GC} = X_C{}^{4m} + X_G{}^{4m} + C_{4m}^{2m} X_C{}^{2m} X_G{}^{2m} \quad (5)$$

Next we present illustrative examples to show how the proposed construction works.

Example 3.1 Let C be the self-dual Z_4 -codes K_4 .From Theorem 3.1, we obtain DNA codes of length 4, with 16 codewords, and GC weight enumerator is

$$W_{GC} = X_C{}^4 + X_G{}^4 + 6 X_C{}^2 X_G{}^2 . \quad (6)$$

Obviously K_4 are just corresponding to 16 DNA codes. So we have the following table:

TABLE II

K_4	D_4	$w_A(d)$	$w_C(d)$	$w_T(d)$	$w_G(d)$
(0000)	(AAAA)	4	0	0	0
(1111)	(CCCC)	0	4	0	0
(2222)	(TTTT)	0	0	4	0
(3333)	(GGGG)	0	0	0	4
(0202)	(ATAT)	2	0	2	0
(1313)	(CGCG)	0	2	0	2
(2020)	(TATA)	2	0	2	0
(3131)	(GCGC)	0	2	0	2
(0022)	(AATT)	2	0	2	0
(1133)	(CCGG)	0	2	0	2
(2200)	(TTAA)	2	0	2	0
(3311)	(GGCC)	0	2	0	2
(0220)	(ATTA)	2	0	2	0
(1331)	(CGGC)	0	2	0	2
(2002)	(TAAT)	2	0	2	0
(3113)	(GCCG)	0	2	0	2

Then the complete weight enumerator of D_4 is defined to be the homogeneous polynomial of degree n in four indeterminates X_A, X_C, X_T and X_G:

$$\mathrm{cwe}_{D_4}(X_A,X_C,X_T,X_G)=\frac{1}{2}((X_A+X_T)^4+(X_C+X_G)^4+(X_A-X_C)^4+(X_C-X_G)^4) \quad (7)$$

The GC-weight enumerator of D_4:

$$W_{GC}=X_C{}^4+X_G{}^4+6X_C{}^2X_G{}^2 \quad (8)$$

Example 3.2 Let K_8 be the Z_4 -linear codes with generator matrix

$$\begin{pmatrix} 1&1&1&1&1&1&1&1\\ 0&2&0&0&0&0&0&2\\ 0&0&2&0&0&0&0&2\\ 0&0&0&2&0&0&0&2\\ 0&0&0&0&2&0&0&2\\ 0&0&0&0&0&2&0&2\\ 0&0&0&0&0&0&2&2 \end{pmatrix} \quad (9)$$

Clearly, any two rows of the generator matrix, distinct or not, are orthogonal. Therefore $K_8 = K_8{}^{\perp}$ and K_8 is self-dual. From Theorem 2.1.1 and Theorem 2.1, K_8 is of type $4^1 2^6$ and then $K_8{}^{\perp}$ is also of type $4^1 2^6$ .Let C be the self-dual Z_4 -codes K_8 .From Theorem 3.1, we obtain DNA codes of length 8, with 256 codewords. Therefore

$$\mathrm{cwe}_{D_8}(X_A,X_C,X_T,X_G)=\frac{1}{2}((X_A+X_T)^8+(X_C+X_G)^8+(X_A-X_C)^8+(X_C-X_G)^8) \quad (10)$$

The GC-weight enumerator of D_8:

$$W_{GC}=X_C{}^8+X_G{}^8+70X_C{}^4X_G{}^4 \quad (11)$$

Obviously, one of the advantages of DNA codes construction is that it is less of complexity than pre-existing DNA codes construction. Moreover, GC weight enumerator can be easily determined.

IV CONCLUSIONS

In this paper, we have introduced a new way of constructing good DNA codes. Due to K_{4m} be the self-dual Z_4 -codes, the constructed DNA codes have some particular properties. It is less of complexity than pre-existing DNA codes construction. Moreover, GC weight enumerator can be easily determined. But the constructed DNA codes have not applied to Genetics and Bioengineering. We must stick at the study of DNA codes.

ACKNOWLEDGMENTS

This work was supported by the the Anhui Provincial Natural Science Foundation (1408085MA05).

REFERENCES

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

[2] L.M.Adleman, Molecular computation of solutions to combinatorial problems, Science 266, 1021-1024, 1994.

[3] Masaaki Harada, Optimal self-dual Z4-codes and a unimodular lattice in dimension 41, Finite Fields and Their Applications 18, 529–536, 2012.

[4] A. Bonnecaze, P. Solé, C. Bachoc, B. Mourrain, Type II codes over Z4, IEEE Trans. Inform. Theory 43, 969–976, 1997.

[5] Z.X. Wan, Quaternary Codes, World Scientific, Singapore, 1997.

[6] Alexis Bonnecaze, A. R. Calderbank, Quaternary Quadratic Residue Codes and Unimodular Lattices, IEEE Transactions on Inffirmation Theory, Vol. 41, No.2, March 1995.

[7] M. Klemm, Se\bstduale Codes tiber dem Ring der ganzen Zahlen modulo 4, A rch. Math,vol. 53, pp. 201-207, 1989.

[8] J.H. Conway, N.J.A. Sloane, Self-dual codes over the integers modulo 4, J. Combin. Theory Ser. A 62, 30–45, 1993.