Seventh International Workshop on Optimal Codes and Related Topics
September 6-12, 2013, Albena, Bulgaria *pp. 170-175*

DNA codes based on additive self-dual codes over GF(4)[1]

Todor Todorov todorvt@abv.bg
Dept. of Information Technologies, University of Veliko Tarnovo
Zlatko Varbanov vtgold@yahoo.com
Dept. of Information Technologies, University of Veliko Tarnovo

Dedicated to the memory of Professor Stefan Dodunekov

Abstract. In this paper we translate in terms of coding theory constraints that are used in designing DNA codes for use in DNA computing. We focus in particular on linear and additive codes over GF(4), and we propose a new construction for DNA codes satisfying the Hamming distance constraint and the GC-content constraint.

1 Introduction

Coding theory has several applications in Genetics and Bioengineering. 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 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.

In this work we translate in terms of coding theory constraints that are used in designing DNA codes for use in DNA computing. We propose new construction for DNA codes satisfying a Hamming distance constraint and a GC-content constraint. Practically, the focus is on additive self-dual codes over $GF(4)$ and their graph representation.

The paper is organized as follows: Section 2 recalls basic notions for DNA codes and codes over a field with 4 elements. In Section 3 the constraints on DNA codes are translated into the terms of classical linear and additive codes. Section 4 lists some known constructions and presents our new construction. In the end of Section 4 we give a table with the obtained results for best known results for DNA codes satisfying a Hamming distance constraint and a GC-content constraint.

[1] This research is partially supported by Veliko Tarnovo University Science Fund, Grant RD672-08/2012

2 Linear and additive codes over $GF(4)$ and DNA codes

A DNA code of length n is a set of codewords $(x_1, \dots, x_n)$ with $x_i \in \{A, C, G, T\}$ (representing the four nucleotides in DNA). We use a hat to denote the Watson-Crick complement of a nucleotide, so $\hat{A} = T, \hat{T} = A, \hat{C} = G$, and $\hat{G} = C$.

The Hamming distance $H(x, y)$ between two codewords x and y is the number of coordinates in which x and y differ. The reverse of a codeword $x = (x_1, \dots, x_n)$ is denoted by $x^R = (x_n, \dots, x_1)$, and the reverse-complement of x is denoted by $x^{RC} = (\hat{x_n}, \dots, \hat{x_1})$.

In this paper we shall identify codes over $\{A, C, G, T\}$ with codes over other four-letter alphabet L, where L is $GF(4) = \{0, 1, \omega, \omega^2\}$, with $\omega^2 + \omega + 1 = 0$. The four symbols in $\{A, C, G, T\}$ are identified with the four symbols in L in the orders given above, so that $\hat{x} = x + 1$, for $x \in GF(4)$.

Let L^n be the n-dimensional vector space over the Galois field $GF(4)$. The *Hamming weight* of a vector $x \in L^n$, written $wt(x)$, is the number of nonzero entries of x, and Hamming distance $d(x, 0) = wt(x)$. A quaternary linear $[n, k]$-code C is a k-dimensional linear subspace of L^n. Any $k \times n$ matrix G (with entries in L) whose rows are a basis of the code C is a generator matrix of C. We call that the generator matrix G is given in systematic form if $G = (I|A)$ where I is $k \times k$ identity matrix and A is $k \times n - k$ matrix. A *minimum weight* (or *minimum distance*) of a linear code is the smallest weight among all nonzero codewords. A quaternary linear $[n, k, d]$-code C is an $[n, k]$-code with minimum distance d. A weight enumerator of a code C is the polynomial $W_C(z) = \sum_{i=0}^{n} A_i z^i$, where A_i is the number of codewords of weight i.

A quaternary additive $(n, 2^k)$ code of length n is an additive subgroup of L^n with 2^k codewords. The definitions for Hamming weight, generator matrix, minimum distance, and weight enumerator are the same as the definitions about linear codes. By $(n, 2^k, d)$ we denote an additive code of length n with 2^k codewords that has minimum distance d. About additive codes over $GF(4)$, there is an inner product arising from the trace map. The trace map $Tr : GF(4) \to GF(2)$ is given by $Tr(x) = x + x^2$. The conjugate of $x \in GF(4)$, denoted $\bar{x}$, is the following image: $\bar{0} = 0, \bar{1} = 1$, and $\bar{\omega} = \omega^2$. Now we can define the trace inner product of two vectors $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ in L^n as:

$$x * y = \sum_{i=1}^{n} Tr(x_i . \bar{y}_i) \qquad (1)$$

If C is an additive code, its dual code with respect to (1) is the code $C^{\perp} = \{x \in GF(4)^n | x * c = 0 \text{ for all } c \in C\}$. If C is an $(n, 2^k)$ code, then $C^{\perp}$ is an $(n, 2^{2n-k})$ code. The code C is self-orthogonal (with respect to (1)) if C is a subset of $C^{\perp}$, and self-dual if $C = C^{\perp}$. In particular, if C is self-dual, then $C^{\perp}$ is an $(n, 2^n)$ code.

In our work we will the following map: $0 \to A, 1 \to T, 2 \to C$, and $3 \to G$. In this case the Watson-Crick complement (the transpositions $A \leftrightarrow T$ and $C \leftrightarrow G$) is presented as $\hat{x} = x + 1$, for $x \in GF(4)$. These transpositions do not affect the GC-weight of the codeword (the number of entries that are C or G).

3 Constraints on DNA codes

- *Hamming distance constraint*: the Hamming distance constraint for a DNA code C is that $H(x, y) \geq d$ for all $x, y \in C$ with $x \neq y$, for some prescribed minimum distance d. This constraint will be enforced in all of the codes we consider, in addition to some combination of the constraints described below.

- *Reverse constraint*: the reverse constraint is that $H(x^R, y) \geq d$ for all $x, y \in C$, including $x = y$. It is useful as an intermediate step in constructing codes with the reverse-complement constraint. A natural idea is to start with a code that is fixed by the reverse permutation R, which exchanges column i and column $n + 1 - i$, for $1 \leq i \leq n$.

- *Reverse-complement constraint*: this constraint is that $H(x^{RC}, y) \geq d$ for all $x, y \in C$, including $x = y$. To construct codes satisfying the reverse-complement constraint, it can be useful to begin with codes over L that contain a special codeword we denote by 1^L, which is the all-one word for $L = GF(4)$. Note that $x^{RC} = x^R + 1^L$, so an additive code containing 1^L that is fixed by the permutation $x \to x^R$ is also fixed by the map $x \to x^{RC}$.

- *GC-content constraint*: this constraint is that each codeword $x \in C$ has the same GC-weight. Starting from a linear code, the question is how to compute the GC-weight enumerator. It can of course be obtained by specializing the complete weight enumerator, but this turns out to be quickly time consuming, since finding the complete weight enumerator may in itself take a long time. We propose in the following a simple way to compute the number of codewords with fixed GC-weight of a special class of additive self-dual codes over $GF(4)$.

In this work we consider DNA codes with fixed GC-content that satisfy given Hamming distance constraint. By $A_4^{GC}(n, d, u)$ we denote the maximum size of a DNA code of length n with fixed GC-content u that satisfies the Hamming distance constraint for a given d.

4 Constructions on DNA codes

Tables with lower bounds on DNA codes can be found in [1] and [2]. Also, in [1] and [2] can be found some constructions on DNA codes as:

- Binary construction
- Lexicographic construction
- Linear reverse construction
- Cyclic (and extended cyclic) code construction
- Shortening and puncturing

Here in our work we present new construction based on additive self-dual codes with circulant generator matrix in graph form (below are the definitions).

A graph code is an additive self-dual code over $GF(4)$ with generator matrix $G = \Gamma + \omega I$ where I is the identity matrix and Γ is the adjacency matrix of a simple undirected graph, which must be symmetric with 0's along the diagonal.

Example:

$$\Gamma = \begin{pmatrix} 0\,1\,1 \\ 1\,0\,0 \\ 1\,0\,0 \end{pmatrix}, \quad C = \Gamma + \omega I = \begin{pmatrix} \omega\,1\,1 \\ 1\,\omega\,0 \\ 1\,0\,\omega \end{pmatrix}$$

Schlingemann [3] first proved (in terms of quantum stabilizer states) that for any self-dual quantum code, there is an equivalent graph code. This means that there is a one-to-one correspondence between the set of simple undirected graphs and the set of additive self-dual codes over $GF(4)$. We have seen that every graph represents an additive self-dual code over $GF(4)$, and that every additive self-dual code over $GF(4)$ can be represented by a graph.

A matrix B of the form

$$B = \begin{pmatrix} b_0 & b_1 & \dots & b_{n-2} & b_{n-1} \\ b_{n-1} & b_0 & b_1 & \dots & b_{n-2} \\ \dots & \dots & \dots & \dots & \dots \\ b_2 & \dots & b_{n-1} & b_0 & b_1 \\ b_1 & b_2 & \dots & b_{n-1} & b_0 \end{pmatrix}$$

is called a circulant matrix. The vector $(b_0, b_1, \dots, b_{n-1})$ is called generator vector for the matrix B. An additive code with circulant generator matrix is called circulant code.

An additive circulant graph (ACG) code is a code corresponding to graph with circulant adjacency matrix [4]. Circulant graphs must be regular, i.e., all vertices must have the same number of neighbors. It is easy to see that such matrix has the following property: $b_i = b_{n-i}$, for all $i = 1, \dots, n-1$, and $b_0 = \omega$. Then, the entries in the generator matrix of ACG code depend on the coordinates $(b_1, b_2, \dots, b_{\lfloor n/2 \rfloor})$ only. Therefore, we can restrict our search space to the $2^{\lfloor n/2 \rfloor}$ codes over $GF(4)$ of length n corresponding to graphs with circulant adjacency matrices.

Besides a smaller search space, the special form of the generator matrix of a graph code makes it easier to determine the minimum distance, since any codeword obtained as a linear combination of i rows of the generator matrix must have weight at least i. If, for example, we want to check whether a code has minimum distance at least d, we only need to consider combinations of $d-1$ or fewer rows of its generator matrix.

Graph codes are proper to construct DNA codes with fixed GC-content u that satisfy Hamming distance constraint for given d. If we know already that the minimum distance of the code is at least d, then $H(x, y) \geq d$ (for any two codewords x and y), and the Hamming distance constraint is satisfied. Other good property is that the generator matrix G of the code has just one position in any row (and column) that is neither 0 nor 1. Then any row of G has GC-weight 1, any codeword that is a sum of two rows has GC-weight 2, any codeword that is a sum of three rows has GC-weight 3, etc. It is easy to see that any codeword that is a sum of u rows has GC-weight u. Then, the corresponding DNA code with $H(x, y) \geq d$ and fixed GC-content u consists of all codewords that are linear combinations of u rows of the generator matrix G. In this way, we obtain the following:

Theorem 1. *Any ACG code of length n with minimum distance d consists of DNA codes of length n with $H(x, y) \geq d$, fixed GC-content u $(1 \leq u \leq n)$, and $A_4^{GC}(n, d, u) = \binom{n}{u}$.*

Using this result, we can improve some lower bounds on $A_4^{GC}(n, d, u)$. Practically, there are no published results for DNA codes of length greater than 30 [2]. By the proposed method, we can construct codes of length $n > 30$, given Hamming distance d, and fixed GC-content u faster than the other known methods. Also, we can search for already known lower bounds on $A_4^{GC}(n, d, u)$ for $n \leq 30$, that are less than the corresponding value $\binom{n}{u}$.

Table 1 consists of the obtained new results. By the described method we construct ACG codes with parameters $(29, 2^{29}, d \geq 10)$ (generator vector 01100001011101) and $(30, 2^{30}, d \geq 11)$ (generator vector 011000011011111). By these codes we obtain that $A_4^{GC}(29, 10, u) = \binom{29}{14} = 77558760$ and $A_4^{GC}(30, 11, u) = \binom{30}{15} = 155117520$. The corresponding DNA codes of length $n > 30$ with minimum distance $d = 10$ were constructed in [4].

n/d	10	11	old bound
29	77558760	–	4859904 [2]
30	–	155117520	1417920 [2]
31	300540195	–	–
32	601080390	–	–
33	1166803110	–	–
34	2333606220	–	–

Table 1: New lower bounds on $A_4^{GC}(n, d, u)$, $29 \leq n \leq 34, d = 10$ or 11

5 Conclusion

In this work we have presented some connections between DNA codes and linear and additive codes over a field with 4 elements. We use a special class of additive self-dual codes and we propose a new construction on DNA codes based on the form of generator matrices of the codes in this special class. By this construction we improve some lower bounds on DNA codes that satisfy the Hamming distance constraint and the GC-content constraint for given parameters. Also, this construction can be used to construct new DNA codes that satisfy the reverse constraint and the reverse-complement constraint for given parameters.

References

[1] P. Gaborit, O. D. King, Linear constructions for DNA codes, *Theoretical Computer Science* 334, 2005, 99-113.

[2] A. A. Niema, The construction of DNA codes using a computer algebra system, PhD Thesis, University of Glamorgan, 2011.

[3] D. Schlingemann, Stabilizer codes can be realized as graph codes, *Quantum Inf. Comput.* 2, 2002, 307-323, arXiv:quant-ph/0111080.

[4] Z. Varbanov, Additive self-dual graph codes over GF(4), Proc. 6th Int. Workshop on Optimal Codes and Related Topics, Varna, Bulgaria, June 2009, 189-195.