DNA Codes based on additive self-dual codes over $\mathbb{F}_4$

Zlatko Varbanov, Todor Todorov

University of Veliko Tarnovo, Bulgaria

7^{th} Int.Workshop on Optimal Codes and Related Topics, Albena, Bulgaria

September 10, 2013

Overview

DNA codes

DNA codes

- DNA codes - sets of words of fixed length n over the alphabet $\{A, C, G, T\}$
- We use a hat to denote the Watson-Crick complement of a nucleotide, so $\hat{A} = T$, $\hat{T} = A$, $\hat{C} = G$, and $\hat{G} = C$

DNA codes problems

- Looking for code that satisfy certain combinatorial constraints reliably storing and retrieving information in synthetic DNA strands
- Good codes can be used in particular for DNA computing or as molecular bar-codes

Linear codes

- $\mathbb{F}_q$ – a field with q elements.
- **Linear** $[n, k]$ **code** C **of length n** – k-dimensional linear subspace of $\mathbb{F}_q^n$.
- **Weight** of a codeword $c \in C$ $(wt(c))$ – the number of nonzero components of c.
- **Hamming distance** $H(x, y)$ between two codewords x and y – the number of coordinates in which x and y differ.
- **Minimum weight (distance)**:
- $d = d(C) = min\{wt(c)|c \in C, c \neq 0\} \to [n, k, d]$ code.
- **Generator matrix of** C – $k \times n$ matrix with entries in $\mathbb{F}_q$ whose rows are a basis of C.
- **Weight enumerator** of C: $C(y) = \sum_{i=0}^{n} A_i y^i$

Linear codes

Self-orthogonal and self-dual codes

- **Inner product** – $x.y = \sum_{i=1}^{n} x_i y_i, \quad x, y \in \mathbb{F}_q^n$
- **Dual code** – $C^{\perp} = \{x \in \mathbb{F}_q^n \mid x.c = 0, \ \forall c \in C\}$
- C – **self-orthogonal** code if $C \subseteq C^{\perp}$
- C – **self-dual** code if $C = C^{\perp}$ $(k = n/2)$

Additive codes over $\mathbb{F}_4$

Additive code

Let $\mathbb{F}_4 = \{0, 1, \omega, \omega^2\}$. An additive code C over $\mathbb{F}_4$ of length n is an additive subgroup of $\mathbb{F}_4^n$. We call C an $(n, 2^k)$ code.

Trace map

Trace map $Tr : \mathbb{F}_4 \to \mathbb{F}_2$ is given by $Tr(x) = x + x^2$.
In particular $Tr(0) = Tr(1) = 0$ and $Tr(\omega) = Tr(\omega^2) = 1$.
The *conjugate* of $x \in \mathbb{F}_4$ (denoted $\bar{x}$) is the following image of x:
$\bar{0} = 0, \bar{1} = 1$, and $\bar{\omega} = \omega^2$.

Additive self-dual codes

Trace inner product

The *trace inner product* of two vectors
$x = (x_1, x_2, \ldots, x_n)$, $y = (y_1, y_2, \ldots, y_n)$ in $\mathbb{F}_4^n$ is

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

Self-dual codes

- **Dual code** – $C^\perp = \{x \in \mathbb{F}_4^n \mid x \star c = 0, \ \forall c \in C\}$
- C – **self-orthogonal** code if $C \subseteq C^\perp$
- C – **self-dual** code if $C = C^\perp$ (it is $(n, 2^n)$ code)

DNA codes

DNA codes

- The reverse of a codeword $x = (x_1, \ldots, x_n)$ is denoted by $x^R = (x_n, \ldots, x_1)$
- The reverse-complement of $x = (x_1, \ldots, x_n)$ is denoted by $x^{RC} = (\hat{x}_n, \ldots, \hat{x}_1)$

Mapping

- In our work we will the following map: $0 \to A$, $1 \to T$, $\omega \to C$, and $\omega^2 \to G$
- In this case the Watson-Crick complement (the transpositions $A \leftrightarrow T$ and $C \leftrightarrow G$) is presented as $\hat{x} = x + 1$, for $x \in \mathbb{F}_4$
- These transpositions do not affect the GC-weight of the codeword

Constraints on DNA codes

Hamming distance constraint

$H(x, y) \geq d$ for all $x, y \in C$ with $x \neq y$, for some prescribed minimum distance d.

Reverse constraint

$H(x^R, y) \geq d$ for all $x, y \in C$, including $x = y$.

Reverse-complement constraint

$H(x^{RC}, y) \geq d$ for all $x, y \in C$, including $x = y$.

GC-content constraint

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.

Constructions

Some constructions on DNA codes as:

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

Gaborit and King, 2005; Niema, 2011

Graph codes

Graph code

A graph code is an additive self-dual code over $\mathbb{F}_4$ 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.

$$\Gamma = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}, \qquad G = \begin{pmatrix} \omega & 1 & 1 \\ 1 & \omega & 1 \\ 1 & 1 & \omega \end{pmatrix}$$

There is a one-to-one correspondence between the set of simple undirected graphs and the set of additive self-dual codes over $\mathbb{F}_4$ [Schlingemann, 2002].

Additive Circulant codes

A $n \times n$ 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 self-dual code with circulant generator matrix is called additive circulant code.

Additive Circulant Graph codes

- An additive circulant graph (ACG) code is a code corresponding to graph with circulant adjacency matrix.
- The generator vector of such a matrix has the following property: $b_i = b_{n-i}$, for all $i = 1, ..., n-1$, and $b_0 = \omega$.
- Then, the entries in the generator matrix of ACG code depend on the coordinates $(b_1, b_2, ..., b_{\lfloor n/2 \rfloor})$

New DNA codes construction

Construction

- 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 constant GC-content u that satisfies the Hamming distance constraint for a given d.
- We present new construction based on additive self-dual codes with circulant generator matrix in graph form

New DNA codes construction

Construction

The graph codes are proper to construct DNA codes with fixed GC-content u that satisfy Hamming distance constraint for given d:

- $H(x, y) \geq d$
- G – just one position in any row (and column) that is neither 0 nor 1
- Any codeword that is a sum of u rows has GC-weight u

Main result

Theorem

Any graph 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 $A_4^{GC}(n,d,u) = \binom{n}{u}$

New results on DNA codes

Example (new lower bound)

- We construct $(29, 2^{29}, d \geq 10)$ ACG code. For this code the largest value for $A_4^{GC}(n, d, u) = \binom{n}{u}$ code is when $u = 14$ or 15. This value is $\binom{29}{14} = 77558760$ (old bound 4859904 [Niema, 2011]).

Example (new lower bound)

- There exists $(31, 2^{31}, d \geq 10)$ ACG code [Varbanov, 2009]. For this code the largest value for $A_4^{GC}(n, d, u) = \binom{n}{u}$ code is when $u = 15$ or 16. This value is $\binom{31}{15} = 300540195$
- All of the entries for $n > 30$ are new bests

New results on DNA codes

n/d	10	11	old bound
29	77 558 760	–	4 859 904 [Niema, 2011]
30	–	155 117 520	1 417 920 [Niema, 2011]
31	300 540 195	–	–
32	601 080 390	–	–
33	1 166 803 110	–	–
34	2 333 606 220	–	–

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

References

Gaborit, P., King, O.D. (2005)
Linear constructions for DNA codes
Theoretical Computer Science 334, 99– 113.

Niema, A.A.(2011)
The construction of DNA codes using a computer algebra system
PhD Thesis, University of Glamorgan.

Varbanov, Z.(2009)
Additive circulant graph codes over $GF(4)$,
6th Int. Workshop on Optimal Codes and Related Topics, Varna, Bulgaria, 189–195.

Tnanks for your attention!