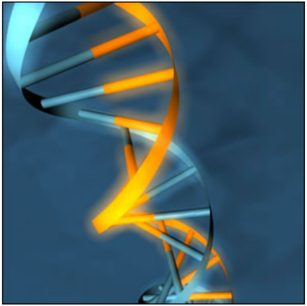
On the number of words of a given GC-content in some cyclic DNA-codes

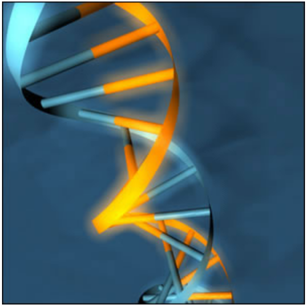
Luis Martínez, University of the Basque Country UPV/EHU
Joint work with Josu Sangroniz, University of the Basque Country UPV/EHU

Symetries of Graphs and Networks IV, Rogla 2014
July 1, 2014

Biological preliminaries

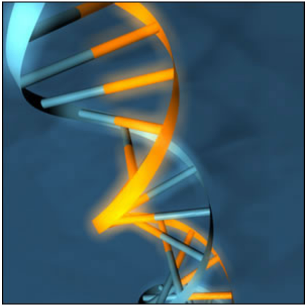


Biological preliminaries



A (Adenine), T (Thymine), G (Guanine), C (Cytosine)

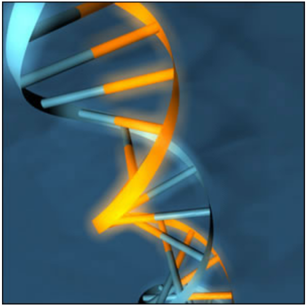
Biological preliminaries



A (Adenine), T (Thymine), G (Guanine), C (Cytosine)

DNA forms a double helix, and two single strands are coupled in a double strand

Biological preliminaries



A (Adenine), T (Thymine), G (Guanine), C (Cytosine)

DNA forms a double helix, and two single strands are coupled in a double strand

$A \longleftrightarrow T, C \longleftrightarrow G$

Biological preliminaries

Transcription: RNA A (Adenine), U (Uracil), G (Guanine), C (Cytosine)

Biological preliminaries

Transcription: RNA A (Adenine), U (Uracil), G (Guanine), C (Cytosine)

Translation: Each three nucleotides of RNA determine an aminoacid (or the beginning of a gene), according to the genetic code.

Biological preliminaries

Transcription: RNA A (Adenine), U (Uracil), G (Guanine), C (Cytosine)

Translation: Each three nucleotides of RNA determine an aminoacid (or the beginning of a gene), according to the genetic code.

For instance, UUA determines Leucine.

Biological preliminaries

Transcription: RNA A (Adenine), U (Uracil), G (Guanine), C (Cytosine)

Translation: Each three nucleotides of RNA determine an aminoacid (or the beginning of a gene), according to the genetic code.

For instance, UUA determines Leucine.

Thus, the complete sequence of a gene determines the sequence of aminoacids of a protein.

Mathematical formulation of DNA-codes

Recall that a code of length n over a finite alphabet Σ is a subset C of Σ^n.

Mathematical formulation of DNA-codes

Recall that a code of length n over a finite alphabet Σ is a subset C of Σ^n.

Intuitively speaking, a DNA-code is a code over the alphabet formed by the four nucleotides A,T,G,C.

Mathematical formulation of DNA-codes

Recall that a code of length n over a finite alphabet Σ is a subset C of Σ^n.

Intuitively speaking, a DNA-code is a code over the alphabet formed by the four nucleotides A,T,G,C.

In practice, usually $\mathbb{F}_4$ or $\mathbb{Z}/4\mathbb{Z}$ is considered for the set Σ.

Mathematical formulation of DNA-codes

Recall that a code of length n over a finite alphabet Σ is a subset C of Σ^n.

Intuitively speaking, a DNA-code is a code over the alphabet formed by the four nucleotides A,T,G,C.

In practice, usually $\mathbb{F}_4$ or $\mathbb{Z}/4\mathbb{Z}$ is considered for the set Σ.

DNA-codes have many applications in Biology and in Genetic Engineering, for instance

Mathematical formulation of DNA-codes

Recall that a code of length n over a finite alphabet Σ is a subset C of Σ^n.

Intuitively speaking, a DNA-code is a code over the alphabet formed by the four nucleotides A,T,G,C.

In practice, usually $\mathbb{F}_4$ or $\mathbb{Z}/4\mathbb{Z}$ is considered for the set Σ.

DNA-codes have many applications in Biology and in Genetic Engineering, for instance

1. In the design of bioarrays

Mathematical formulation of DNA-codes

Recall that a code of length n over a finite alphabet Σ is a subset C of Σ^n.

Intuitively speaking, a DNA-code is a code over the alphabet formed by the four nucleotides A,T,G,C.

In practice, usually $\mathbb{F}_4$ or $\mathbb{Z}/4\mathbb{Z}$ is considered for the set Σ.

DNA-codes have many applications in Biology and in Genetic Engineering, for instance

1. In the design of bioarrays
2. In biomolecular computing

Mathematical formulation of DNA-codes

Recall that a code of length n over a finite alphabet Σ is a subset C of Σ^n.

Intuitively speaking, a DNA-code is a code over the alphabet formed by the four nucleotides A,T,G,C.

In practice, usually $\mathbb{F}_4$ or $\mathbb{Z}/4\mathbb{Z}$ is considered for the set Σ.

DNA-codes have many applications in Biology and in Genetic Engineering, for instance

1. In the design of bioarrays
2. In biomolecular computing
3. As molecular barcodes

Mathematical formulation of DNA-codes

Several methods have been used to obtain DNA-codes, in particular

Mathematical formulation of DNA-codes

Several methods have been used to obtain DNA-codes, in particular

1. Additive codes

Mathematical formulation of DNA-codes

Several methods have been used to obtain DNA-codes, in particular

1. Additive codes
2. Linear codes

Mathematical formulation of DNA-codes

Several methods have been used to obtain DNA-codes, in particular

1. Additive codes
2. Linear codes
3. Cyclic codes

Mathematical formulation of DNA-codes

Several methods have been used to obtain DNA-codes, in particular

1. Additive codes
2. Linear codes
3. Cyclic codes
4. Cosets of linear codes

Mathematical formulation of DNA-codes

Definition

A linear DNA-code C is complementable if $u + (1, ..., 1) \in C$ for every $u \in C$.

Mathematical formulation of DNA-codes

Definition

A linear DNA-code C is complementable if $u + (1, \ldots, 1) \in C$ for every $u \in C$.

Theorem

A cyclic code C over $\mathbb{F}_4$ is complementable if and only if $X - 1$ does not divide the generator polynomial of the code C.

Mathematical formulation of DNA-codes

Definition

A linear DNA-code C is reversible if $(a_n, \ldots, a_1) \in C$ for every $(a_1, \ldots, a_n) \in C$.

Mathematical formulation of DNA-codes

Definition

A linear DNA-code C is reversible if $(a_n, \ldots, a_1) \in C$ for every $(a_1, \ldots, a_n) \in C$.

Definition

If q is a prime power and $g(X) = g_0 + \cdots + X^r \in \mathbb{F}_q[X]$ is a monic polynomial of degree r dividing $X^n - 1$, then the reciprocal polynomial of $g(X)$ is the polynomial $g_R(X) = g_0^{-1} X^r g(X^{-1})$. The polynomial $g(X)$ is called self-reciprocal if $g(X) = g_R(X)$.

Mathematical formulation of DNA-codes

Definition

A linear DNA-code C is reversible if $(a_n, \dots, a_1) \in C$ for every $(a_1, \dots, a_n) \in C$.

Definition

If q is a prime power and $g(X) = g_0 + \cdots + X^r \in \mathbb{F}_q[X]$ is a monic polynomial of degree r dividing $X^n - 1$, then the reciprocal polynomial of $g(X)$ is the polynomial $g_R(X) = g_0^{-1} X^r g(X^{-1})$. The polynomial $g(X)$ is called self-reciprocal if $g(X) = g_R(X)$.

Theorem (Massey)

A cyclic code C over $\mathbb{F}_4$ is reversible if and only if the generator polynomial of the code C is self-reciprocal.

Combinatorial restrictions on the words of a DNA-code

Some biological interesting combinatorial restrictions are usually imposed on the words of a DNA-code C

Combinatorial restrictions on the words of a DNA-code

Some biological interesting combinatorial restrictions are usually imposed on the words of a DNA-code C

1. Hamming constraint: $d(u, v) \geq d \; \forall u, v \in C$ with $u \neq v$.

Combinatorial restrictions on the words of a DNA-code

Some biological interesting combinatorial restrictions are usually imposed on the words of a DNA-code C

1. Hamming constraint: $d(u, v) \geq d \; \forall u, v \in C$ with $u \neq v$.
2. Complement constraint: $d(u, v^c) \geq d \; \forall u, v \in C$.

Combinatorial restrictions on the words of a DNA-code

Some biological interesting combinatorial restrictions are usually imposed on the words of a DNA-code C

1. Hamming constraint: $d(u, v) \geq d \ \forall u, v \in C$ with $u \neq v$.
2. Complement constraint: $d(u, v^c) \geq d \ \forall u, v \in C$.
3. Reverse complement constraint: $d(u, v^{rc}) \geq d \ \forall u, v \in C$.

Combinatorial restrictions on the words of a DNA-code

Some biological interesting combinatorial restrictions are usually imposed on the words of a DNA-code C

1. Hamming constraint: $d(u, v) \geq d \ \forall u, v \in C$ with $u \neq v$.
2. Complement constraint: $d(u, v^c) \geq d \ \forall u, v \in C$.
3. Reverse complement constraint: $d(u, v^{rc}) \geq d \ \forall u, v \in C$.
4. GC-content constraint: the number of positions in which a nucleotide G or C appears is the same for all the words of the code.

Combinatorial restrictions on the words of a DNA-code

As usual in the literature of DNA-codes, $\max_w A_4^{GC,RC}(n,d,w)$ will denote the maximum number number of words in a DNA-code of length n satisfying the Hamming constraint and the reverse-complement constraint with parameter d and the constant GC-content constraint.

Combinatorial restrictions on the words of a DNA-code

As usual in the literature of DNA-codes, $\max_w A_4^{GC,RC}(n, d, w)$ will denote the maximum number number of words in a DNA-code of length n satisfying the Hamming constraint and the reverse-complement constraint with parameter d and the constant GC-content constraint.

It is well known that, if C is a complementable reversible DNA-code with minimum distance d, we can put $C = C_0 \cup C_1 \cup C_2$, where C_0 is the set of words in C which coincide with their reverse complement and where $u^{rc} \in C$ if and only if $u \in C$, and $\max_w A_4^{GC,RC}(n, d, w) \geq |C_1|$.

Number of words with a given GC-content

Definition

Let $u \in \mathbb{F}_4^n$. We will call $\mathbb{F}_2$-weight of u, and we will denote it $wt_{\mathbb{F}_2}(u)$, to the number of coordinates of u which are in $\mathbb{F}_2$. If $C \subseteq \mathbb{F}_4^n$ is a code over $\mathbb{F}_4$, we define the $\mathbb{F}_2$-weight enumerator polynomial to be

$$W_{\mathbb{F}_2,C}(X) = \sum_{u \in C} X^{wt_{\mathbb{F}_2}(u)} = \sum_{w \geq 0} b_w X^w,$$

where $b_w = b_w(C)$ is the number of words of the code C with $\mathbb{F}_2$-weight equal to w.

Number of words with a given GC-content

Definition

Let $g \in \mathbb{F}_4[X]$ *be a divisor of* $X^n - 1$. *We will say that the cyclic code generated by* g *is Galois-supplemented if* $(g, g^\sigma) = 1$, *where* σ *is the Frobenius automorphism of* $\mathbb{F}_4$ *over* $\mathbb{F}_2$.

Number of words with a given GC-content

Definition

Let $g \in \mathbb{F}_4[X]$ be a divisor of $X^n - 1$. We will say that the cyclic code generated by g is Galois-supplemented if $(g, g^\sigma) = 1$, where σ is the Frobenius automorphism of $\mathbb{F}_4$ over $\mathbb{F}_2$.

Theorem

Let $C \subseteq \mathbb{F}_4^n$ be a Galois-supplemented cyclic code with generator polynomial g. Then,

$$W_{F_2,C}(X) = 2^{n-2\deg g}(X+1)^n.$$

Number of words with a given GC-content

Aboulion et al. gave a table of lower bounds for $\max_w A_4^{GC,RC}(n,d,w)$ for $n \leq 30$. In particular, for $\max_w A_4^{GC,RC}(29,11,w)$ they obtained the bound 38777664. By considering the Quadratic-residue code of length 29 over $\mathbb{F}_4$, which is complementable and reversible, and whose minimum distance is 11, and using the previous Theorem, we obtain that $\max_w A_4^{GC,RC}(29,11,w) \geq 77558760$, and so we have improved that bound for this set of parameters.

Future research

Future research

1. Study other combinatorial restrictions more precise that the constant GC-content constraint.

Future research

1. Study other combinatorial restrictions more precise that the constant GC-content constraint.
2. Take advantage of good symmetry groups on codes.

Future research

THANK YOU VERY MUCH FOR YOUR ATTENTION!