Finite Fields and Their Applications 37 (2016) 328–343

Contents lists available at ScienceDirect

Finite Fields and Their Applications

www.elsevier.com/locate/ffa

Construction of DNA codes by using algebraic number theory

Haibo Hong [a,*], Licheng Wang [b], Haseeb Ahmad [b], Jing Li [b], Yixian Yang [b], Changzhong Wu [c]

[a] *School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China*
[b] *State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[c] *Anhui Post and Telecommunication College, Hefei, 230031, China*

ARTICLE INFO

Article history:
Received 24 November 2014
Received in revised form 11 August 2015
Accepted 30 October 2015
Available online 18 November 2015
Communicated by Michael Tsfasman

MSC:
94B05
94B15
20G40
12E20
05E15

Keywords:
DNA codes
Irreducible cyclic codes
Gauss sum
Gauss period

ABSTRACT

The canonical structure of DNA has four bases – Thymine (T), Adenine (A), Cytosine (C), and Guanine (G) – and DNA codes are regarded as words over the alphabet set $\Sigma = \{A, C, G, T\}$, satisfying certain combinatorial conditions. Good DNA codes are desirable for DNA computation, DNA microarray technologies and molecular barcodes, etc. One of the main tasks in DNA code designing is to build more codewords and better GC-content for given fixed word length n. Existing heuristic methods work well for small n. In this paper, we present a systematic method for constructing good DNA codes for large n by using irreducible cyclic codes. Being different from traditional DNA constructions, our method is based on algebraic number theory rather than classical heuristic algorithms and the conventional coding theory. Furthermore, comparing with the traditional DNA codes, our codes have larger number of codewords and better GC-content. As far as we know, it is the very first time to

* Corresponding author.
E-mail address: honghaibo1985@163.com (H. Hong).

http://dx.doi.org/10.1016/j.ffa.2015.10.008

utilize irreducible cyclic codes for constructing a type of DNA codes.

1. Introduction

DNA codes are the sets of words of fixed length over the genetic alphabet $\{A, C, G, T\}$. Meanwhile, hybridizations of DNA follow the principle that short single strands of DNA should be hybridized with their Watson–Crick complements, i.e., $\overline{A} = T$, $\overline{T} = A$, $\overline{C} = G$, and $\overline{G} = C$.

DNA codes have wide applications for storage and retrieval of information in synthetic DNA strands. DNA codes are also used for DNA computation [1,3,7–10,12,20,22], as probes in DNA microarray technologies [24] and as molecular barcodes for chemical libraries [5]. In these applications, the primary property required is that short single strands of DNA should be hybridized with their Watson–Crick complements, but other undesirable hybridizations should be unlikely [3]. This is achieved by specifying a minimum Hamming distance between words and as well as between the reverse complements of words.

Another desirable property concerns with the melting temperature. DNA melting is the process by which double-stranded DNA unwinds and separates into single-stranded strands through the breakage of hydrogen bonding between the bases. Similar melting temperatures could be achieved approximately by ensuring that each word contains the same number of positions, which are either G or C.

By employing the notation from code theory, we use $\mathcal{D} = (n, M, d, w)_4$ for denoting a DNA code defined over the genetic alphabet $\{A, C, G, T\}$, the meanings of corresponding parameters are specified as follows: n is the length of the code; M is the number of the codewords; d is the minimum Hamming distance[1]; w is GC-content, which is defined as the number of letters of each codeword either as G or C.

The DNA codes designing problem is about to find the largest possible set of DNA codewords, each of length n, satisfying certain combinatorial conditions [23,4,2]. Let $\mathcal{D} = (n, M, d, w)_4$ and $c = (c_0, c_1, c_2, \cdots, c_{n-1}) \in \mathcal{D}$, then the conditions are presented as follows:

(1) (HD) $H(c, c') \geq d$, $\forall c, c' \in \mathcal{D}, c \neq c'$
(2) (R) $H(c, {c'}^{-1}) \geq d$, $\forall c, c' \in \mathcal{D}$

[1] Note that here, Hamming distance between two codewords x and y, denoted by $H(x, y)$, is defined as the number of positions i at which the ith letter in x differs from the ith letter in y. For instance, $H(ACTG, ATGG) = 2$.