

# Binary templates for comma-free DNA codes

Oliver D. King<sup>a,\*</sup>, Philippe Gaborit<sup>b</sup>

<sup>a</sup>Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, USA

<sup>b</sup>LACO, Université de Limoges, 123 av. A. Thomas, 87060 Limoges, France

Received 29 June 2004; received in revised form 1 July 2005; accepted 5 July 2005

Available online 17 October 2006

## Abstract

Arita and Kobayashi proposed a method for constructing comma-free DNA codes using binary templates, and showed that the separation  $d$  of any such binary template of length  $n$  satisfies  $d \leq n/2$ . Kobayashi, Kondo and Arita later produced an infinite family of binary templates with  $d \geq 11n/30$ . Here we demonstrate the existence of an infinite family of binary templates with  $d > n/2 - (18n \log_e n)^{1/2}$ . We also give an explicit construction for an infinite family of binary templates with  $d > n/2 - 19n^{1/2} \log_e n$ .  
© 2006 Elsevier B.V. All rights reserved.

**Keywords:** Comma-free codes; DNA barcodes; Legendre sequences

## 1. Introduction

Suppose  $\mathbf{x} = (x_1, \dots, x_n) \in \mathcal{F}^n$  and  $\mathbf{y} = (y_1, \dots, y_m) \in \mathcal{F}^m$  are words of length  $n$  and  $m$ , respectively, over an alphabet  $\mathcal{F}$ . If  $n = m$ , we let  $D(\mathbf{x}, \mathbf{y})$  denote the Hamming distance between  $\mathbf{x}$  and  $\mathbf{y}$ , i.e., the number of positions in which they differ. If  $n < m$  we define

$$D(\mathbf{x}, \mathbf{y}) = \min_{0 \leq i \leq m-n} D(\mathbf{x}, \mathbf{y}^{[i:n]}),$$

where  $\mathbf{y}^{[i:n]} = (y_{i+1}, \dots, y_{i+n})$  is the substring of  $\mathbf{y}$  beginning in the  $(i+1)$ th position and having length  $n$ . We denote the concatenation of  $\mathbf{x}$  and  $\mathbf{y}$  by  $\mathbf{xy} = (x_1, \dots, x_n, y_1, \dots, y_m)$ , and we define  $\langle \mathbf{x} \rangle = (x_2, \dots, x_{n-1})$  to be  $\mathbf{x}$  with its first and last positions removed. We denote the reverse of  $\mathbf{x} = (x_1, \dots, x_n)$  by  $\mathbf{x}^R = (x_n, \dots, x_1)$ . In the case of DNA codes—codes over the alphabet  $\{A, C, G, T\}$  of nucleotides—we denote the reverse complement of  $\mathbf{x} = (x_1, \dots, x_n)$  by  $\mathbf{x}^{RC} = (\bar{x}_n, \dots, \bar{x}_1)$ , where  $\bar{A} = T$ ,  $\bar{T} = A$ ,  $\bar{C} = G$ , and  $\bar{G} = C$ .

A *comma-free* code [5,6] is a subset  $\mathcal{C}$  of  $\mathcal{F}^n$  with the property that  $D(\mathbf{x}, \langle \mathbf{yz} \rangle) > 0$  for all (not-necessarily distinct)  $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{C}$ . A code  $\mathcal{C}$  is *comma-free of index  $d$*  [9] if  $D(\mathbf{x}, \langle \mathbf{yz} \rangle) \geq d$  for all (not-necessarily distinct)  $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{C}$ . Comma-free codes were introduced by Crick et al. [5] as a model for a non-overlapping, self-synchronizing genetic code. Comma-free codes of index greater than 1 can provide error-correction in addition to self-synchronization—see [11] for a recent overview.

\* Corresponding author. Present address: Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA. Tel.: +1 617 258 6918; fax: +1 617 258 6691.

E-mail addresses: [oking@wi.mit.edu](mailto:oking@wi.mit.edu) (O.D. King), [gaborit@unilim.fr](mailto:gaborit@unilim.fr) (P. Gaborit).

<sup>1</sup> Supported in part by a fellowship from the NIH/NHGRI.

Although the genetic code turned out not to be comma-free (see [7] for a historical account), comma-free binary codes have found other applications in communications (see e.g. [17]) and in bioengineering. DNA codes, with distance constraints imposed to reduce the probability of unwanted hybridizations, have been synthesized for use as molecular barcodes in yeast deletion libraries [15], and for encoding problem-instances in DNA computing [1]. For some applications in which codewords may concatenate, it is desirable to have a large set  $\mathcal{C}$  of words in  $\{A, C, G, T\}^n$  for which

$$\min_{\substack{\mathbf{x}, \mathbf{y}, \mathbf{z}, \mathbf{w} \in \mathcal{C} \\ \mathbf{x} \neq \mathbf{y}}} \min \left\{ \begin{array}{l} D(\mathbf{x}, \mathbf{y}), D(\mathbf{x}, \mathbf{z}^{RC}), D(\mathbf{x}, \langle \mathbf{wz} \rangle), D(\mathbf{x}, \langle \mathbf{w}^{RC} \mathbf{z}^{RC} \rangle), \\ D(\mathbf{x}, \langle \mathbf{w}^{RC} \mathbf{z} \rangle), D(\mathbf{x}, \langle \mathbf{wz}^{RC} \rangle) \end{array} \right\}$$

is large [3]. We will call this minimum, say  $d$ , the *separation* of  $\mathcal{C}$ . A code  $\mathcal{C}$  with separation  $d$  is comma-free of index  $d$  in the sense described above, with the additional properties that  $\mathcal{C} \cap \mathcal{C}^{RC} = \emptyset$ , that the minimum Hamming distance between distinct codewords in  $\mathcal{C} \cup \mathcal{C}^{RC}$  is at least  $d$ , and that  $\mathcal{C} \cup \mathcal{C}^{RC}$  is also comma-free of index  $d$ . (Here  $\mathcal{C}^{RC} = \{\mathbf{x}^{RC} : \mathbf{x} \in \mathcal{C}\}$ .)

Arita and Kobayashi [3] proposed a “template-code” strategy for constructing comma-free DNA codes with fixed GC-content (a rough indicator of melting temperature) and large separation. They combine a binary template word  $\mathbf{x}$  with a binary error-correcting code  $\mathcal{C}$  to form  $\mathbf{x} \otimes \mathcal{C} := \{\mathbf{x} \otimes \mathbf{y} : \mathbf{y} \in \mathcal{C}\}$ , where  $(x_1, \dots, x_n) \otimes (y_1, \dots, y_n) = (w_1, \dots, w_n)$  with  $w_i = A$  if  $x_i = 1$  and  $y_i = 1$ ;  $w_i = C$  if  $x_i = 0$  and  $y_i = 0$ ;  $w_i = G$  if  $x_i = 0$  and  $y_i = 1$ ; and  $w_i = T$  if  $x_i = 1$  and  $y_i = 0$ . They define

$$\|\mathbf{x}\| = \min\{D(\mathbf{x}, \mathbf{x}^R), D(\mathbf{x}, \langle \mathbf{xx} \rangle), D(\mathbf{x}, \langle \mathbf{x}^R \mathbf{x}^R \rangle), D(\mathbf{x}, \langle \mathbf{x}^R \mathbf{x} \rangle), D(\mathbf{x}, \langle \mathbf{xx}^R \rangle)\},$$

and define  $d_n$  to be the maximum value of  $\|\mathbf{x}\|$  taken over all  $\mathbf{x} \in \{0, 1\}^n$ . The separation of the DNA code  $\mathbf{x} \otimes \mathcal{C}$  is the smaller of  $\|\mathbf{x}\|$  and the minimum Hamming distance of  $\mathcal{C}$  (i.e.,  $\min\{D(\mathbf{x}, \mathbf{y}) : \mathbf{x}, \mathbf{y} \in \mathcal{C}, \mathbf{x} \neq \mathbf{y}\}$ ). As there are binary codes  $\mathcal{C}$  of length  $n$  having any minimum Hamming distance up to  $n$ ,  $d_n$  is the largest attainable separation for a DNA code of the form  $\mathbf{x} \otimes \mathcal{C}$ . Arita and Kobayashi [3] showed that  $d_n \leq n/2$  for all  $n$ . They also computed  $d_n$  exactly for  $n \leq 32$  by exhaustive search, and together with Kondo [10] showed that  $d_n \geq 11n/30$  for infinitely many  $n$ . Here we show that  $d_n > n/2 - (18n \log_e n)^{1/2}$  for infinitely many  $n$  (specifically, for  $n$  prime) using a non-constructive counting argument. We also give explicit constructions with  $d > n/2 - 19n^{1/2} \log_e n$  when  $n$  is a prime of the form  $4k + 3$ , using Legendre sequences.

### 2. Random constructions

Fix  $n$ , and for  $i = 0, \dots, n$  and  $k = 0, \dots, n$  define

$$\begin{aligned} S_i^n(k) &= \{\mathbf{x} \in \{0, 1\}^n : D(\mathbf{x}, \langle \mathbf{xx} \rangle^{[i:n]}) = k\}, \\ R_i^n(k) &= \{\mathbf{x} \in \{0, 1\}^n : D(\mathbf{x}, \langle \mathbf{x}^R \mathbf{x}^R \rangle^{[i:n]}) = k\}, \\ T_i^n(k) &= \{\mathbf{x} \in \{0, 1\}^n : D(\mathbf{x}, \langle \mathbf{xx}^R \rangle^{[i:n]}) = k\}, \\ U_i^n(k) &= \{\mathbf{x} \in \{0, 1\}^n : D(\mathbf{x}, \langle \mathbf{x}^R \mathbf{x} \rangle^{[i:n]}) = k\}. \end{aligned}$$

Note that

$$\begin{aligned} \bigcup_{i=1}^{n-1} \bigcup_{k=0}^d S_i^n(k) &= \{\mathbf{x} \in \{0, 1\}^n : D(\mathbf{x}, \langle \mathbf{xx} \rangle) \leq d\}, \\ \bigcup_{i=1}^{n-1} \bigcup_{k=0}^d R_i^n(k) &= \{\mathbf{x} \in \{0, 1\}^n : D(\mathbf{x}, \langle \mathbf{x}^R \mathbf{x}^R \rangle) \leq d\}, \\ \bigcup_{i=1}^{n-1} \bigcup_{k=0}^d T_i^n(k) &= \{\mathbf{x} \in \{0, 1\}^n : D(\mathbf{x}, \langle \mathbf{xx}^R \rangle) \leq d\}, \end{aligned}$$

$$\bigcup_{i=1}^{n-1} \bigcup_{k=0}^d U_i^n(k) = \{\mathbf{x} \in \{0, 1\}^n : D(\mathbf{x}, \langle \mathbf{x}^R \mathbf{x} \rangle) \leq d\},$$

$$\bigcup_{k=0}^d R_0^n(k) = \{\mathbf{x} \in \{0, 1\}^n : D(\mathbf{x}, \mathbf{x}^R) \leq d\},$$

so that  $d_n$  is the largest integer  $d$  for which the cardinality of the set

$$\bigcup_{i=1}^{n-1} \bigcup_{k=0}^d S_i^n(k) \cup \bigcup_{i=0}^{n-1} \bigcup_{k=0}^d R_i^n(k) \cup \bigcup_{i=1}^{n-1} \bigcup_{k=0}^d T_i^n(k) \cup \bigcup_{i=1}^{n-1} \bigcup_{k=0}^d U_i^n(k)$$

is strictly less than  $2^n$ . (For  $R_i^n(k)$  the index  $i$  starts at 0 to account for  $D(\mathbf{x}, \mathbf{x}^R)$ .) By the union bound, this cardinality is at most

$$\sum_{i=1}^{n-1} \sum_{k=0}^d \#S_i^n(k) + \sum_{i=0}^{n-1} \sum_{k=0}^d \#R_i^n(k) + \sum_{i=1}^{n-1} \sum_{k=0}^d \#T_i^n(k) + \sum_{i=1}^{n-1} \sum_{k=0}^d \#U_i^n(k),$$

so if this sum (or any upper bound for this sum) is less than  $2^n$  then  $d_n \geq d$ . Below we give upper bounds for each term in this sum, which may be used to compute a lower bound for  $d_n$ . To simplify the arguments, we consider only the case of prime  $n > 2$ .

**Proposition 1.** *Suppose  $n = 2m + 1$  is prime. Then for all  $k = 0, \dots, n$  and all  $i = 1, \dots, n - 1$ , and also for  $i = 0$  in the case of  $\#R_i^n(k)$ ,*

- (a)  $\#S_i^n(k) = \begin{cases} 2 \binom{n}{k} & \text{if } k \text{ is even,} \\ 0 & \text{if } k \text{ is odd,} \end{cases}$
- (b)  $\#R_i^n(k) = \begin{cases} 2^{m+1} \binom{m}{k/2} & \text{if } k \text{ is even,} \\ 0 & \text{if } k \text{ is odd,} \end{cases}$
- (c)  $\#T_i^n(k) \leq 2^{\lfloor i/2 \rfloor} \sum_{j=0}^{\lfloor k/2 \rfloor} \binom{\lfloor i/2 \rfloor}{j} \binom{n-i}{k-2j},$
- (d)  $\#U_i^n(k) \leq 2^{\lceil (n-i)/2 \rceil} \sum_{j=0}^{\lfloor k/2 \rfloor} \binom{\lfloor (n-i)/2 \rfloor}{j} \binom{i}{k-2j}.$

**Proof.** Let  $\phi : \{1, \dots, n\} \rightarrow \{1, \dots, n\}$  (not necessarily one-to-one), and consider a directed graph  $G$  with nodes  $v_1, \dots, v_n$  and with an edge from vertex  $v_r$  to  $v_{\phi(r)}$  for  $r = 1, \dots, n$ . Given a bitstring  $\mathbf{x} = (x_1, \dots, x_n)$ , we will color the edge from  $v_r$  to  $v_{\phi(r)}$  black if  $x_r = x_{\phi(r)}$  and gray otherwise. Note that the number of gray edges in  $G$  is exactly the Hamming distance between the bitstrings  $(x_1, \dots, x_n)$  and  $(x_{\phi(1)}, \dots, x_{\phi(n)})$ . The bitstring  $\mathbf{x}$  is completely determined once we specify which edges are gray and also specify the value  $x_r$  for a single node  $v_r$  from each weakly connected component of  $G$ . Note also that since the  $x_r$  are binary valued, the number of gray edges in any cycle in  $G$  must be even.

For the remainder of this proof we suppose  $n = 2m + 1$  is prime, and we interpret  $n \bmod n$  as  $n$  rather than 0.

(a) To compute  $\#S_i^n(k)$  for  $0 < i < n$ , we define  $\phi(r) = (i + r) \bmod n$ , so that  $(x_{\phi(1)} \dots x_{\phi(n)}) = (x_1, \dots, x_n, x_1, \dots, x_n)^{\lfloor i:n \rfloor}$ . Since  $n$  is prime,  $\gcd(i, n) = 1$ , and so  $i$  has order  $n$  in the additive group  $\mathbb{Z}/n\mathbb{Z}$ . Thus the resulting graph  $G$  consists of a single weakly connected component, which is a cycle of length  $n$  (and is in fact strongly connected). The total number of gray edges must therefore be even, but the gray edges can be specified arbitrarily otherwise. The number of ways to specify the value for one element in each connected component is 2, and the number of ways to specify which  $k$  of the  $n$  edges are gray is  $\binom{n}{k}$  for  $k$  even and 0 otherwise.

(b) To compute  $\#R_i^n(k)$  for  $0 \leq i < n$ , we define  $\phi(r) = (n + 1 - i - r) \bmod n$ , so that  $(x_{\phi(1)} \dots x_{\phi(n)}) = (x_n, \dots, x_1, x_n, \dots, x_1)^{\lfloor i:n \rfloor}$ . Note that  $\phi$  is in this case an involution with a single fixed-point (since  $n$  is odd), so the resulting graph

$G$  consists of  $m + 1$  weakly connected components (which are again strongly connected). One of them is a node with an edge to itself, and the remaining  $m$  are disjoint cycles of length 2. The edge from the node to itself cannot be gray, and for each cycle of length two either both edges are gray or neither are. Hence the number of ways to specify the value for one element in each connected component is  $2^{m+1}$ , and the number of ways to specify which  $k$  of the  $n$  edges are gray is  $\binom{m}{k/2}$  for  $k$  even and 0 otherwise.

(c) To bound  $\#T_i^n(k)$  for  $0 < i < n$ , we define  $\phi(r) = (i + r) \bmod n$  for  $1 \leq r \leq n - i$  and  $\phi(r) = (n + 1 - i - r) \bmod n$  for  $n - i < r \leq n$ , so that  $(x_{\phi(1)} \dots x_{\phi(n)}) = (x_1, \dots, x_n, x_n, \dots, x_1)^{[i:n]}$ . Note that the restriction of  $G$  to just the  $i$  vertices  $v_s$  with  $n - i < s \leq n$  consists of  $\lceil i/2 \rceil$  weakly connected components including  $\lceil i/2 \rceil$  cycles of length 2. Also, since  $n$  is prime, each vertex  $v_r$  with  $1 \leq r \leq n - i$  is in the same weakly connected component as some vertex  $v_s$  with  $n - i < s \leq n$  (otherwise the order of  $i$  in the additive group  $\mathbb{Z}/n\mathbb{Z}$  would be less than  $n$ ). Hence the entire graph  $G$  consists of at most  $\lceil i/2 \rceil$  weakly connected components and has at least  $\lfloor i/2 \rfloor$  cycles of length 2. If we let  $j$  denote the number of these  $\lfloor i/2 \rfloor$  cycles of length 2 for which both edges are gray, the number of ways to specify which  $k$  of the  $n$  edges are gray is at most  $\sum_{j=0}^{\lfloor k/2 \rfloor} \binom{\lfloor i/2 \rfloor}{j} \binom{n-i}{k-2j}$ .

(d) The argument for  $U_i^n(k)$  is similar to that for  $T_i^n(k)$ . Alternatively, note that  $D(\mathbf{x}, (\mathbf{x}\mathbf{x}^R)^{[i:n]}) = D(\mathbf{x}^R, ((\mathbf{x}\mathbf{x}^R)^R)^{[n-i:n]}) = D(\mathbf{x}^R, ((\mathbf{x}^R)^R \mathbf{x}^R)^{[n-i:n]})$ . Hence for any  $k$ ,  $U_{n-i}^n(k) = \{\mathbf{x}^R : \mathbf{x} \in T_i^n(k)\}$ , so  $\#U_{n-i}^n(k) = \#T_i^n(k)$ .  $\square$

Next we will give a simpler lower bound on  $d_n$ , using the following bound for large deviations of the binomial distribution:

**Lemma 2** (Chernoff [4]; see also e.g. [2, Theorem A.1.1] for a proof). *If  $c > 0$ , then*

$$\sum_{k=0}^{n/2 - cn^{1/2}} \binom{n}{k} < 2^n e^{-2c^2}.$$

We will also use the weaker result that  $\binom{n}{k} < 2^n e^{-2c^2}$  for  $k \leq n/2 - cn^{1/2}$ , and the trivial observation that  $\binom{n}{k} \leq 2^n$  for any  $k$ .

**Theorem 3.** *Suppose  $n = 2m + 1$  is prime. Then*

$$d_n > n/2 - (18n \log_e n)^{1/2}.$$

**Proof.** Fix  $c > 0$  and suppose for the remainder of this proof that  $d \leq n/2 - cn^{1/2}$ . Then by Proposition 1(a) and Lemma 1,

$$\sum_{k=0}^d \#S_i^n(k) \leq \sum_{k=0}^d 2 \binom{n}{k} < 2^{n+1} e^{-2c^2}$$

so we have

$$\sum_{i=1}^{n-1} \sum_{k=0}^d \#S_i^n(k) < (n - 1) 2^{n+1} e^{-2c^2}$$

Similarly,  $d \leq n/2 - cn^{1/2}$  implies that

$$\frac{d}{2} \leq \frac{2m + 1}{4} - \frac{c}{2}(2m + 1)^{1/2} < \frac{m + 1}{2} - \frac{c}{2}(m + 1)^{1/2},$$

so by Proposition 1(b) and Lemma 1

$$\sum_{k=0}^d \#R_i^n(k) = \sum_{k=0}^{\lfloor d/2 \rfloor} 2^{m+1} \binom{m}{k} \leq \sum_{k=0}^{\lfloor d/2 \rfloor} 2^{m+1} \binom{m + 1}{k} < 2^{m+1} 2^{m+1} e^{-c^2/2},$$

and so

$$\sum_{i=0}^{n-1} \sum_{k=0}^d \#R_i^n(k) < n2^{n+1} e^{-c^2/2}.$$

Next, note that for any  $0 < i < n$ , any  $k \leq d$  and any  $j \leq k/2$ , either

$$j \leq \frac{i}{4} - \frac{c}{3}n^{1/2}$$

or

$$k - 2j \leq \frac{n-i}{2} - \frac{c}{3}n^{1/2},$$

since otherwise

$$d \geq k = 2j + (k - 2j) > 2 \left( \frac{i}{4} - \frac{c}{3}n^{1/2} \right) + \left( \frac{n-i}{2} - \frac{c}{3}n^{1/2} \right) = \frac{n}{2} - cn^{1/2}.$$

In the former case we have  $j \leq \lceil i/2 \rceil / 2 - (c/3)(\lceil i/2 \rceil)^{1/2}$ , so by Lemma 1

$$\binom{\lfloor i/2 \rfloor}{j} \leq \binom{\lceil i/2 \rceil}{j} < 2^{\lceil i/2 \rceil} e^{-2c^2/9}$$

and trivially

$$\binom{n-i}{k-2j} \leq 2^{n-i},$$

in the latter case we have  $k - 2j \leq (n-i)/2 - (c/3)(n-i)^{1/2}$ , so by Lemma 1

$$\binom{n-i}{k-2j} < 2^{n-i} e^{-2c^2/9}$$

and trivially

$$\binom{\lfloor i/2 \rfloor}{j} \leq 2^{\lfloor i/2 \rfloor}.$$

Hence in either case we have

$$\binom{\lfloor i/2 \rfloor}{j} \binom{n-i}{k-2j} 2^{\lceil i/2 \rceil} < 2^{\lceil i/2 \rceil} 2^{n-i} 2^{\lceil i/2 \rceil} e^{-2c^2/9} \leq 2^{n+1} e^{-2c^2/9},$$

so by Proposition 1(c)

$$\#T_i^n(k) \leq 2^{\lceil i/2 \rceil} \sum_{j=0}^{\lfloor k/2 \rfloor} \binom{\lfloor i/2 \rfloor}{j} \binom{n-i}{k-2j} < (\lfloor k/2 \rfloor + 1) 2^{n+1} e^{-2c^2/9},$$

and so

$$\sum_{i=1}^{n-1} \sum_{k=0}^d \#T_i^n(k) < (n-1)2^n e^{-2c^2/9} \sum_{k=0}^d (\lfloor k/2 \rfloor + 1) \leq \frac{n^3}{4} 2^{n+1} e^{-2c^2/9}.$$

(Here we used that

$$(n-1) \sum_{k=0}^d (\lfloor k/2 \rfloor + 1) = \begin{cases} (n-1)(d+1 + \lfloor d/2 \rfloor^2 + \lfloor d/2 \rfloor) & \text{for } d \text{ odd,} \\ (n-1)(d+1 + (d/2)^2) & \text{for } d \text{ even,} \end{cases}$$

which one can check by induction is at most  $n^3/4$  when  $n \geq 3$  and  $d \leq n/2$ .)

Since  $\#T_i^n(k) = \#U_{n-i}^n(k)$ , we also have

$$\sum_{i=1}^{n-1} \sum_{k=0}^d \#U_i^n(k) < \frac{n^3}{4} 2^{n+1} e^{-2c^2/9}.$$

Now note that if  $n \geq 2$  then  $n \leq n^3/4$ , so if  $d \leq n/2 - cn^{1/2}$  we have

$$\begin{aligned} & \sum_{i=1}^{n-1} \sum_{k=0}^d \#S_i^n(k) + \sum_{i=0}^{n-1} \sum_{k=0}^d \#R_i^n(k) + \sum_{i=1}^{n-1} \sum_{k=0}^d \#T_i^n(k) + \sum_{i=1}^{n-1} \sum_{k=0}^d \#U_i^n(k) \\ & \leq (n-1) 2^{n+1} e^{-2c^2} + n 2^{n+1} e^{-c^2/2} + \frac{n^3}{4} 2^{n+1} e^{-2c^2/9} + \frac{n^3}{4} 2^{n+1} e^{-2c^2/9} \\ & \leq n^3 2^{n+1} e^{-2c^2/9}. \end{aligned}$$

Thus if we take  $c \geq ((9/2)\log_e(2n^3))^{1/2}$  then  $n^3 2^{n+1} e^{-2c^2/9} \leq 2^n$  so  $d_n > n/2 - cn^{1/2}$ . Since  $n \geq 2$ , it suffices to take  $c = (18 \log_e n)^{1/2}$ .  $\square$

**Remark 1.** In the above proof we have not tried too hard to optimize the constant term in  $c$ .

**Remark 2.** If we take  $c \geq ((9/2)\log_e(2^k n^3))^{1/2}$  for some  $k$  in the above proof, we get that  $\|\mathbf{x}\| > n/2 - cn^{1/2}$  for at least  $2^n - 2^{n-k+1} + 1$  of the  $2^n$  bitstrings  $\mathbf{x}$  of length  $n$ . This can be useful for constructing larger DNA codes by using multiple templates as in [10].

**3. Pseudo-random constructions**

In this section, we give a concrete example of an infinite family of templates for which the separation can be shown to be not-too-much below the non-constructive lower bound of Theorem 1. Here it will be convenient to consider words  $\mathbf{x}$  over the alphabet  $\{+1, -1\}$  rather than  $\{0, 1\}$ —this has no effect on  $\|\mathbf{x}\|$  or  $d_n$ . For prime  $p > 2$  and for any integer  $i$ , the Legendre symbol  $(i/p)$  is defined by

$$\left(\frac{i}{p}\right) = \begin{cases} +1 & \text{if } i^{(p-1)/2} \equiv +1 \pmod{p}, \\ -1 & \text{if } i^{(p-1)/2} \equiv -1 \pmod{p}, \\ 0 & \text{if } i \equiv 0 \pmod{p}. \end{cases}$$

(Equivalently,  $(i/p) = +1$  if  $i$  is a quadratic residue mod  $p$ , and  $(i/p) = -1$  if  $i$  is a quadratic non-residue mod  $p$ ). The following properties of the Legendre symbol will be useful for our purposes:

- (a)  $\left(\frac{i+p}{p}\right) = \left(\frac{i}{p}\right)$ ,
- (b)  $\left(\frac{ij}{p}\right) = \left(\frac{i}{p}\right) \left(\frac{j}{p}\right)$ ,
- (c)  $\left(\frac{-1}{p}\right) = \begin{cases} +1 & \text{if } p \equiv 1 \pmod{4}, \\ -1 & \text{if } p \equiv 3 \pmod{4}, \end{cases}$
- (d)  $\sum_{i=0}^{p-1} \left(\frac{i}{p}\right) \left(\frac{i+j}{p}\right) = \begin{cases} -1 & \text{for } j \not\equiv 0 \pmod{p}, \\ p-1 & \text{for } j \equiv 0 \pmod{p}. \end{cases}$
- (e)  $\left| \sum_{i=0}^k \left(\frac{i}{p}\right) \left(\frac{i+j}{p}\right) \right| \leq 18p^{1/2} \log_e p$  for  $k < p$  and  $j \not\equiv 0 \pmod{p}$ .

Properties (a)–(d) are standard (see e.g. [14,13]), and the bound on the incomplete character sum in property (e) is a specialization to polynomials of degree 2 of Corollary 1 in [12], which follows from a theorem of Weil’s [16].

Define a modified two-valued Legendre symbol  $(i/p)'$  by  $(i/p)' = 1$  if  $i \equiv 0 \pmod{p}$  and  $(i/p)' = (i/p)$  otherwise. We will call the sequence  $\mathbf{x} = (x_1, x_2, \dots, x_p)$  with  $x_i = ((i-1)/p)$  the Legendre sequence of length  $p$ , and the sequence

with  $x_i = ((i - 1)/p)'$  the modified Legendre sequence of length  $p$ —they differ only in the first position. Modified Legendre sequences pass various tests of randomness, and are among a family of sequences sometimes referred to as pseudo-random or pseudonoise sequences (see e.g. [12,8]). Exactly half of the numbers  $\{1, \dots, p - 1\}$  are quadratic residues mod  $p$ . This is useful here since for any sequence  $\mathbf{y} = (y_1, \dots, y_n)$  with  $y_i = +1$  in  $w$  positions and  $y_i = -1$  in  $n - w$  positions,  $\|\mathbf{y}\| \leq D(\mathbf{y}, \mathbf{y}^R \mathbf{y}^R) \leq 2w(n - w)/n$ ; as  $2w(n - w)/n$  is maximized when  $w = n - w = n/2$ , this gives the upper bound  $\|\mathbf{y}\| \leq n/2$  (see [3, Lemma 2.5]). The sharply-peaked periodic autocorrelation in property (d) can be used to bound  $D(\mathbf{x}, \langle \mathbf{x}\mathbf{x} \rangle)$  for a modified Legendre sequence  $\mathbf{x}$ . By properties (a) and (b) we have  $((p - j)/p) = (-j/p) = (-1/p)(j/p)$ , so a Legendre sequence with its first position truncated is palindromic when  $(-1/p) = +1$  and anti-palindromic otherwise—this is one feature of Legendre sequences that is *not* typical of random sequences, but will be useful here for bounding  $D(\mathbf{x}, \mathbf{x}^R \mathbf{x}^R)$ . Property (e) will be useful for bounding  $D(\mathbf{x}, \langle \mathbf{x}^R \mathbf{x} \rangle)$  and  $D(\mathbf{x}, \langle \mathbf{x}\mathbf{x}^R \rangle)$ .

**Theorem 4.** *Let  $p$  be an odd prime of the form  $p = 4k + 3$ , and let  $\mathbf{x} = (x_1, x_2, \dots, x_p)$  with  $x_i = ((i - 1)/p)'$ . Then  $\|\mathbf{x}\| > p/2 - 19 p^{1/2} \log_e p$ .*

**Proof.** First note that if  $x_i$  and  $y_i$  take values in  $\{+1, -1\}$  for any finite index set  $I$ , then

$$\#\{i \in I : x_i = y_i\} + \#\{i \in I : x_i \neq y_i\} = \#I$$

and

$$\#\{i \in I : x_i = y_i\} - \#\{i \in I : x_i \neq y_i\} = \sum_{i \in I} x_i y_i,$$

so

$$\#\{i \in I : x_i \neq y_i\} = \frac{1}{2} \left( \#I - \sum_{i \in I} x_i y_i \right).$$

For  $p$  a prime of the form  $4k + 3$ , by properties (a)–(c) we have  $((p - j)/p) = (-j/p) = (-1/p)(j/p) = -(j/p)$ . Then by property (d), for  $j = 1, \dots, p - 1$  we have

$$\begin{aligned} D(\mathbf{x}, \langle \mathbf{x}\mathbf{x} \rangle^{[j:p]}) &= \frac{1}{2} \left( p - \sum_{i=0}^{p-1} \left( \frac{i}{p} \right)' \left( \frac{i+j}{p} \right)' \right) \\ &= \frac{1}{2} \left( p - \sum_{i=0}^{p-1} \left( \frac{i}{p} \right) \left( \frac{i+j}{p} \right) - \left( \frac{j}{p} \right) - \left( \frac{p-j}{p} \right) \right) \\ &= \frac{1}{2} (p - (-1)), \end{aligned}$$

so  $D(\mathbf{x}, \langle \mathbf{x}\mathbf{x} \rangle) = (p + 1)/2$ .

For  $j = p - 1$  we have

$$\begin{aligned} D(\mathbf{x}, \langle \mathbf{x}^R \mathbf{x}^R \rangle^{[j:p]}) &= \frac{1}{2} \left( p - \sum_{i=0}^{p-1} \left( \frac{i}{p} \right)' \left( \frac{p-1-j-i}{p} \right)' \right) \\ &= \frac{1}{2} \left( p - \sum_{i=0}^{p-1} \left( \frac{i}{p} \right)' \left( \frac{-i}{p} \right)' \right) \\ &= \frac{1}{2} \left( p - \sum_{i=0}^{p-1} \left( \frac{i}{p} \right) \left( \frac{-i}{p} \right) - 1 \right) \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2} \left( p + \sum_{i=0}^{p-1} \binom{i}{p} \binom{i}{p} - 1 \right) \\
&= \frac{1}{2} (p + (p-1) - 1) = p-1
\end{aligned}$$

and for  $j = 0, 1, \dots, p-2$  we have

$$\begin{aligned}
D(\mathbf{x}, (\mathbf{x}^R \mathbf{x}^R)^{[j:p]}) &= \frac{1}{2} \left( p - \sum_{i=0}^{p-1} \binom{i}{p}' \left( \frac{p-1-j-i}{p} \right)' \right) \\
&= \frac{1}{2} \left( p - \sum_{i=0}^{p-1} \binom{i}{p} \left( \frac{p-1-j-i}{p} \right) - 2 \left( \frac{p-1-j}{p} \right) \right) \\
&= \frac{1}{2} \left( p + \sum_{i=0}^{p-1} \binom{i}{p} \left( \frac{i+j+1}{p} \right) + 2 \left( \frac{j+1}{p} \right) \right) \\
&= \frac{1}{2} \left( p + (-1) + 2 \left( \frac{j+1}{p} \right) \right) = (p-1)/2 \pm 1.
\end{aligned}$$

Thus  $D(\mathbf{x}, \mathbf{x}^R) = D(\mathbf{x}, (\mathbf{x}^R \mathbf{x}^R)^{[0:p]}) = (p-1)/2 \pm 1$ , and since  $((j+1)/p) = -1$  for some  $0 < j < p-2$ ,  $D(\mathbf{x}, (\mathbf{x}^R \mathbf{x}^R)) = (p-1)/2 - 1$ .

For  $j = p-1$  we have

$$\begin{aligned}
D(\mathbf{x}, (\mathbf{x}^R)^{[j:p]}) &= \frac{1}{2} \left( p - \sum_{i=0}^{p-1-j} \binom{i}{p}' \left( \frac{i+j}{p} \right)' - \sum_{i=p-j}^{p-1} \binom{i}{p}' \left( \frac{p-1-j-i}{p} \right)' \right) \\
&= \frac{1}{2} \left( p - \binom{0}{p}' \left( \frac{p-1}{p} \right)' - \sum_{i=1}^{p-1} \binom{i}{p}' \left( \frac{-i}{p} \right)' \right) \\
&= \frac{1}{2} \left( p - (-1) + \sum_{i=1}^{p-1} \binom{i}{p} \left( \frac{i}{p} \right) \right) \\
&= \frac{1}{2} (p+1 + (p-1)) = p,
\end{aligned}$$

and for  $j = 1, \dots, p-2$ , by two applications of property (e) we have

$$\begin{aligned}
D(\mathbf{x}, (\mathbf{x}^R)^{[j:p]}) &= \frac{1}{2} \left( p - \sum_{i=0}^{p-1-j} \binom{i}{p}' \left( \frac{i+j}{p} \right)' - \sum_{i=p-j}^{p-1} \binom{i}{p}' \left( \frac{p-1-j-i}{p} \right)' \right) \\
&= \frac{1}{2} \left( p - \sum_{i=0}^{p-1-j} \binom{i}{p} \left( \frac{i+j}{p} \right) - \binom{j}{p} - \sum_{i=p-j}^{p-1} \binom{i}{p} \left( \frac{p-1-j-i}{p} \right) - \left( \frac{p-1-j}{p} \right) \right) \\
&= \frac{1}{2} \left( p - \sum_{i=0}^{p-1-j} \binom{i}{p} \left( \frac{i+j}{p} \right) - \binom{j}{p} + \sum_{i=p-j}^{p-1} \binom{i}{p} \left( \frac{i+j+1}{p} \right) + \left( \frac{j+1}{p} \right) \right) \\
&\geq \frac{1}{2} (p - 18p^{1/2} \log_e p - 1 - 18p^{1/2} \log_e p - 1) \\
&= \frac{1}{2} (p - 36p^{1/2} \log_e p - 2) > p/2 - 19p^{1/2} \log_e p.
\end{aligned}$$

This shows that  $D(\mathbf{x}, \langle \mathbf{x}\mathbf{x}^R \rangle) > p/2 - 19p^{1/2} \log_e p$ . A similar argument shows that  $D(\mathbf{x}, \langle \mathbf{x}^R \mathbf{x} \rangle) > p/2 - 19p^{1/2} \log_e p$ , which completes the proof.  $\square$

**Remark 3.** Arita and Kobayashi [3] also consider the problem of finding the  $\mathbf{x} \in \{0, 1\}^n$  that maximizes  $\|\mathbf{x}\|' := \min\{D(\mathbf{x}, \mathbf{x}^R), D(\mathbf{x}, \langle \mathbf{x}\mathbf{x} \rangle), D(\mathbf{x}, \langle \mathbf{x}^R \mathbf{x}^R \rangle)\}$ . (This differs from  $\|\mathbf{x}\|$  in that it does not consider  $D(\mathbf{x}, \langle \mathbf{x}\mathbf{x}^R \rangle)$  or  $D(\mathbf{x}, \langle \mathbf{x}^R \mathbf{x} \rangle)$ .) Arita and Kobayashi showed that  $\|\mathbf{x}\|' \leq n/2$  for all  $\mathbf{x}$ . Note that the proof above shows that  $\|\mathbf{x}\|' = (n-1)/2 - 1$  when  $\mathbf{x}$  is the modified Legendre sequence of length  $n$  for prime  $n = 4k + 3$ .

**Remark 4.** All primes other than 2 can be written uniquely as either  $4k + 1$  or  $4k + 3$  for some non-negative integer  $k$ . There are infinitely many primes of each type by Dirichlet's theorem. When  $p$  is a prime of the form  $4k + 1$  and  $\mathbf{x}$  is the modified Legendre sequence of length  $p$ ,  $\|\mathbf{x}\| = 0$  since  $\mathbf{x}^R = (\mathbf{x}\mathbf{x})^{[1:p]}$ . However, these sequences are useful for the modified definition of  $\|\mathbf{x}\|$  in which all occurrences of  $\mathbf{x}^R$  are replaced by the bitwise complement of  $\mathbf{x}^R$ , as considered in [10].

## Acknowledgments

The authors thank B. Poonen and K. Paterson for advice on incomplete character sums, and H. Niederreiter for directing them to Ref. [12].

## References

- [1] L.M. Adleman, Molecular computation of solutions to combinatorial problems, *Science* 266 (5187) (1994) 1021–1024.
- [2] N. Alon, J.H. Spencer, *The Probabilistic Method*, Wiley, New York, 2000.
- [3] M. Arita, S. Kobayashi, DNA sequence design using templates, *New Gener. Comput.* 20 (3) (2002) 263–278.
- [4] H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Ann. Math. Statist.* 23 (4) (1952) 493–509.
- [5] F.H.C. Crick, J.S. Griffith, L.E. Orgel, Codes without commas, *Proc. Nat. Acad. Sci. U.S.A.* 43 (1957) 416–421.
- [6] S.W. Golomb, B. Gordon, L.R. Welch, Comma-free codes, *Canad. J. Math.* 10 (1958) 202–209.
- [7] B. Hayes, The invention of the genetic code, *Amer. Sci.* 86 (1) (1998) 8–14.
- [8] T. Hellese, P.V. Kumar, Sequences with low correlation, in: V. Pless, G. Huffman (Eds.), *Handbook of Coding Theory*, Elsevier, Amsterdam, 1998, pp. 1765–1852.
- [9] B.H. Jiggs, Recent results in comma-free codes, *Canad. J. Math.* 15 (1963) 178–187.
- [10] S. Kobayashi, T. Kondo, M. Arita, On template method for DNA sequence design, *DNA8*, *Lecture Notes in Computer Science*, vol. 2568, Springer, Berlin, 2002, pp. 205–214.
- [11] V.I. Levenshtein, Combinatorial problems motivated by comma-free codes, *J. Combin. Designs* 12 (3) (2004) 184–196.
- [12] C. Mauduit, A. Sárközy, On finite pseudorandom binary sequences. I. Measure of pseudorandomness, the Legendre symbol, *Acta Arith.* 82 (4) (1997) 365–377.
- [13] M.R. Schroeder, *Number Theory in Science and Communication*, Springer, Berlin, 1997.
- [14] J.-P. Serre, *A Course in Arithmetic*, Springer, New York, 1973.
- [15] D. Shoemaker, D.A. Lashkari, D. Morris, M. Mittmann, R.W. Davis, Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy, *Nat. Genet.* 14 (4) (1996) 450–456.
- [16] A. Weil, *Sur les Courbes Algébriques et les Variétés Qui s'en Déduisent*, Hermann, Paris, 1948.
- [17] S.B. Wicker, Deep space applications, in: V. Pless, G. Huffman (Eds.), *Handbook of Coding Theory*, Elsevier, Amsterdam, 1998, pp. 2119–2169.