Bounds for DNA codes with constant GC-content

Oliver D. King*

Department of Biological Chemistry and Molecular Pharmacology
Harvard Medical School, Boston, MA, USA
oliver_king@hms.harvard.edu

Submitted: Jun 10, 2003; Accepted: Aug 30, 2003; Published: Sep 8, 2003
MR Subject Classification: 05B40

Abstract

We derive theoretical upper and lower bounds on the maximum size of DNA codes of length n with constant GC-content w and minimum Hamming distance d, both with and without the additional constraint that the minimum Hamming distance between any codeword and the reverse-complement of any codeword be at least d. We also explicitly construct codes that are larger than the best previously-published codes for many choices of the parameters n, d and w.

Introduction

Libraries of DNA words satisfying certain combinatorial constraints have applications to DNA barcoding and DNA computing (see e.g. [17] and the references therein). The goal is to design libraries that are as large as possible given the constraints.

We first review some terminology and notation – see [16, 17] for more context. Let Z_q denote the q-character alphabet $\{0, \ldots, q-1\}$. By a *q-ary word* of length n we mean an element $\mathbf{x}$ of Z_q^n, which we write as $\mathbf{x} = x_1 \cdots x_n$. A *$q$-ary code* of length n is just a subset of Z_q^n, and the elements of the code are called *codewords*. The *Hamming distance* $H(\mathbf{x}, \mathbf{y})$ between two q-ary words $\mathbf{x}$ and $\mathbf{y}$ of length n is defined to be the number of coordinates in which they differ, and the *Hamming weight* of $\mathbf{x}$ is the number of coordinates in which it is nonzero. The maximum cardinality of a q-ary code of length n for which the minimum Hamming distance between two distinct codewords is at least d is denoted $A_q(n, d)$. If we also require each codeword to have Hamming weight w (i.e., that the code be a *constant-weight code*), the maximum cardinality is denoted $A_q(n, d, w)$.

A *DNA code* is a q-ary code with $q = 4$; we identify the elements 0, 1, 2, 3 $\in Z_4$ with the nucleotides A, C, G, T (in that order). The *reverse complement* of a DNA word $\mathbf{x} = x_1 \cdots x_n$ is denoted by $\mathbf{x}^{RC}$, and is defined to be the word $\overline{x_n} \cdots \overline{x_1}$ where $\overline{x_i}$

*Supported in part by a fellowship from NIH/NHGRI

is the Watson-Crick complement of x_i (i.e., $\overline{A} = T$, $\overline{T} = A$, $\overline{C} = G$, and $\overline{G} = C$). By requiring the minimum Hamming distance between two DNA codewords to be sufficiently large, one can make it unlikely that a codeword hybridizes to the reverse-complement of any other codeword. By requiring the minimum Hamming distance between a DNA codeword and the reverse-complement of a DNA codeword to be sufficiently large, one can make it unlikely that a codeword hybridizes to any other codeword or to itself [9]. We denote by $A_4^{RC}(n,d)$ the maximum size of a DNA code of length n in which $H(\mathbf{x},\mathbf{y}) \geq d$ for all distinct codewords $\mathbf{x}$ and $\mathbf{y}$ and $H(\mathbf{x},\mathbf{y}^{RC}) \geq d$ for all (not-necessarily distinct) codewords $\mathbf{x}$ and $\mathbf{y}$. If we also require each codeword to have Hamming weight w the maximum cardinality is denoted $A_4^{RC}(n,d,w)$.

The *GC-content* of a DNA word is defined to be the number of positions in which the word has coordinate C or G. It may be desirable that all codewords in a DNA code have roughly the same GC-content, so that they have similar melting temperatures (see e.g. [9]); $A_4^{GC}(n,d,w)$ and $A_4^{GC,RC}(n,d,w)$ are defined analogously to $A_4(n,d,w)$ and $A_4^{RC}(n,d,w)$, except that in the former two cases it is the GC-content (rather than the Hamming weight) of each codeword that is required to be w.

Theoretical upper and lower bounds on $A_4^{RC}(n,d,w)$, with no restriction on GC-content, are given in [17]. Explicit constructions using stochastic local search [23, 24] and a "template-map" strategy [14] provide lower bounds on $A_4^{GC}(n,d,w)$ and $A_4^{GC,RC}(n,d,w)$ for a limited range of parameters n, d and w. In this paper we derive theoretical upper and lower bounds on $A_4^{GC}(n,d,w)$ and $A_4^{GC,RC}(n,d,w)$ for all parameters, and we use lexicographic constructions to find explicit codes that improve on many of the lower bounds in [14, 23, 24].

Upper bounds

Before giving upper bounds on the sizes of DNA codes with constant GC-content, we note some simple special cases:

Proposition 1 *For $n > 0$, with $0 \leq d \leq n$ and $0 \leq w \leq n$,*

$$A_4^{GC}(n,d,0) = A_2(n,d) \tag{1}$$

$$A_4^{GC}(n,d,w) = A_4^{GC}(n,d,n-w) \tag{2}$$

$$A_4^{GC}(n,n,w) = \begin{cases} 4 \text{ if } w = n/2 \\ 3 \text{ if } n/3 \leq w < n/2 \text{ or } n/2 < w \leq 2n/3 \\ 2 \text{ if } w < n/3 \text{ or } w > 2n/3 \end{cases} \tag{3}$$

$$A_4^{GC,RC}(n,n,w) = \begin{cases} 2 \text{ if } w = n/2 \\ 1 \text{ if } w \neq n/2 \end{cases} \tag{4}$$

$$A_4^{GC}(n,1,w) = \binom{n}{w} 2^n \tag{5}$$

$$A_4^{GC,RC}(n,1,w) = \begin{cases} \frac{1}{2}\left(\binom{n}{w}2^n - \binom{n/2}{w/2}2^{n/2}\right) \text{ if } n \text{ is even and } w \text{ is even,} \\ \frac{1}{2}\binom{n}{w}2^n \text{ if } n \text{ is odd or } w \text{ is odd.} \end{cases} \tag{6}$$

Proof. (1): Changing all 0's in a binary code to A's and all 1's to T's gives a Hamming-distance-preserving bijection between the set of all binary codes of length n and the set of all DNA codes of length n with constant GC-content 0.

(2): Interchange A's with C's, and T's with G's.

(3): By (2) we may assume $w \leq n/2$. If no two codewords agree in any position, then there can be at most four codewords by the pigeonhole principle. Hence $A(n,n,w) \leq 4$ for all w. If there are four codewords none of which agree in any position, then each of the four nucleotides must occur exactly once in each of the n positions, so the average GC-content of the four words is exactly $n/2$. This implies that $A(n,n,w) \leq 3$ for $w < n/2$, since in a code with constant GC-content w, the average GC-content is w. If three words each have GC-content $w < n/3$, then there is some position j in which none of the words has a C or G, and at least two of the three words must agree in this position (both A or both T). Hence $A(n,n,w) \leq 2$ if $w < n/3$. The following constructions demonstrate the reverse inequalities: For $w = n/2$, the four words A^wC^w, C^wA^w, T^wG^w and G^wT^w have pairwise distance n; for $n/3 \leq w < n/2$ the three words C^wA^{n-w}, $T^{n-w}C^w$ and $A^{\lfloor (n-w)/2 \rfloor}G^wT^{\lceil (n-w)/2 \rceil}$ have pairwise distance n; for $w < n/3$ the two words C^wA^{n-w} and G^wT^{n-w} are distance n apart.

(4): For $w = n/2$, the two words A^wC^w and C^wA^w satisfy the distance and reverse-complement constraints. For $w \neq n/2$, the word C^wA^{n-w} satisfies the constraints. These are the largest sets possible, by (3) together with Theorem 7.

(5): This is the total number of DNA words of length n and GC-content w.

(6): When n and w are even, there are $\binom{n/2}{w/2}2^{n/2}$ words with GC-content w that are their own reverse complements, otherwise there are none.

Johnson-type bounds

A code of length n can be *shortened* to a (usually smaller) code of length $n-1$ without increasing the minimum Hamming distance, by choosing any character $b \in Z_q$ and any position $i \in \{1, \ldots, n\}$, keeping just those codewords that have b in their i-th position, and then deleting the i-th position from these codewords [16]. This procedure is used in proving the following bounds.

Theorem 2 *For $0 \leq d \leq n$ and $0 < w < n$,*

$$A_4^{GC}(n,d,w) \leq \lfloor \frac{2n}{w} A_4^{GC}(n-1,d,w-1) \rfloor \quad (7)$$

$$A_4^{GC}(n,d,w) \leq \lfloor \frac{2n}{n-w} A_4^{GC}(n-1,d,w) \rfloor. \quad (8)$$

Proof. (7): In any set of M words with length n, minimum Hamming distance at least d and constant GC-content w, there is some position i in which at least $\lceil wM/2n \rceil$ codewords have nucleotide C, or some position i in which at least $\lceil wM/2n \rceil$ codewords have nucleotide G — otherwise, the average GC-content would be less than w. Keeping just these codewords, and deleting position i, gives a code with length $n-1$, GC-content $w-1$, and minimum Hamming distance at least d. Inequality (8) is analogous, based on the observation that there is some position with at least $\lceil (n-w)M/2n \rceil$ A's or $\lceil (n-w)M/2n \rceil$ T's.

Remark 3 Upper bounds on $A_4^{GC}(n, d, w)$ are obtained by repeatedly applying inequalities (7) and (8), in any order, until $n = d$, $n = w$ or $w = 0$, at which point (1)–(3) may be used. (Different orders of applying (7) and (8) may result in different bounds.) One may continue using (8) even after $w = 0$ (or (7) even after $n = w$), until $n = d$, but this amounts to upper-bounding $A_4^{GC}(n, d, 0) = A_2(n, d)$ with the Singleton bound, 2^{n-d+1} (see e.g. [3]). Tighter upper bounds for $A_2(n, d)$ are known for many n and d — see for example [15].

Theorem 4 *Suppose there is a set of M words of length n, constant GC-content w, and minimum Hamming distance at least d. Write $wM = nk + r$ with $0 \le r < n$. Then*

$$\begin{aligned} M(M-1)\,d \;\; &\le \;\; (n-r)\,(M^2 - \lfloor \tfrac{k}{2} \rfloor^2 - \lceil \tfrac{k}{2} \rceil^2 - \lfloor \tfrac{M-k}{2} \rfloor^2 - \lceil \tfrac{M-k}{2} \rceil^2) \\ &+ \;\; r\,(M^2 - \lfloor \tfrac{k+1}{2} \rfloor^2 - \lceil \tfrac{k+1}{2} \rceil^2 - \lfloor \tfrac{M-k-1}{2} \rfloor^2 - \lceil \tfrac{M-k-1}{2} \rceil^2). \end{aligned} \tag{9}$$

Proof. Let a_i, c_i, g_i and t_i denote the number of occurrences of A, C, G and T (respectively) in the i-th position of the M codewords. Note that $\sum_{i=1}^n (c_i + g_i) = wM$. The sum of the Hamming distances over all M^2 ordered pairs of codewords is $D = \sum_{i=1}^n (M^2 - a_i^2 - c_i^2 - g_i^2 - t_i^2)$. Subject only to the constraints that $a_i + c_i + g_i + t_i = M$ for each i and that $\sum_{i=1}^n (c_i + g_i) = wM$, the expression D is maximized when $c_i + g_i$ is as close as possible to wM/n for each i, when a_i is as close as possible to t_i for each i, and when c_i is as close as possible to g_i for each i. This is also true when a_i, c_i, g_i and t_i are constrained to be integers, as can be proved using the same type of argument as in [19], for example. Hence the right-hand-side of (9) is an upper bound for the sum of the M^2 pairwise Hamming distances. For the left-hand-side, note that since the Hamming distance between distinct codewords is at least d, the sum of the Hamming distances taken over all M^2 ordered pairs of codewords is at least $M(M-1)\,d$.

If we relax the constraint that the counts a_i, c_i, g_i and t_i be integers, Theorem 4 simplifies to the following:

Theorem 5 *If $2dn > w^2 + 4w(n-w) + (n-w)^2$, then*

$$A_4^{GC}(n, d, w) \le \frac{2dn}{2dn - (w^2 + 4w(n-w) + (n-w)^2)}. \tag{10}$$

Remark 6 Versions of the bounds in Theorems 2, 4 and 5 for binary constant-weight codes [11, 12] are called Johnson bounds. Johnson bounds have been generalized to q-ary constant-weight codes [25, 7] and to q-ary *constant-composition codes* (where the number of occurrences of each character in each codeword is prescribed) [22]. They can also be generalized to a setting in which the q characters $\{0, \ldots, q-1\}$ are partitioned into any number of subsets, with the total number of occurrences from each subset specified.

Constant-weight codes correspond to the partition $\{0,\ldots,q-1\}=\{0\}\cup\{1,\ldots,q-1\}$, and constant-composition codes to the partition $\{0,\ldots,q-1\}=\{0\}\cup\cdots\cup\{q-1\}$. Our bounds for DNA codes with constant GC-content correspond to the partition $\{0,1,2,3\}=\{0,3\}\cup\{1,2\}$.

Halving bound

Any upper bound for $A_4^{GC}(n,d,w)$ yields an upper bound for $A_4^{GC,RC}(n,d,w)$ by the following result, an analogue of the halving bound for DNA codes with unrestricted GC-content in [17]. The same proof works here, since the reverse-complement of a DNA word has the same GC-content as the word itself.

Theorem 7 *For $0<d\le n$ and $0\le w\le n$,*

$$A_4^{GC,RC}(n,d,w)\le \frac{1}{2}A_4^{GC}(n,d,w). \tag{11}$$

Proof. If $\{\mathbf{x}_i\}_{i=1}^M$ is a set of M codewords with constant GC-content w, minimum Hamming distance at least d, and with $H(\mathbf{x}_i,\mathbf{x}_j^{RC})\ge d$ for all $1\le i,j\le M$, then $\{\mathbf{x}_i\}_{i=1}^M\cup\{\mathbf{x}_i^{RC}\}_{i=1}^M$ is a set of words with constant GC-content w and minimum Hamming distance at least d. This set has cardinality $2M$ provided that $\{\mathbf{x}_i\}_{i=1}^M\cap\{\mathbf{x}_i^{RC}\}_{i=1}^M=\emptyset$, which holds for $d>0$.

Lower bounds

Gilbert-type bounds

If $\mathcal{C}$ is set of words in Z_q^n with the property that the Hamming distance between any pair of words in $\mathcal{C}$ is at least d, and if $\mathcal{C}$ is maximal in the sense that no more points from Z_q^n can be added to $\mathcal{C}$ without violating this distance constraint, then the balls of Hamming radius $d-1$ around the points in $\mathcal{C}$ cover all of Z_q^n. This is the idea behind the Gilbert bound for q-ary codes (see e.g. [20]), and a similar argument applies to constant-weight codes (see e.g. [4]). Here we give an analogue for DNA codes with constant GC-content:

Theorem 8 *For $0\le d\le n$ and $0\le w\le n$,*

$$A_4^{GC}(n,d,w)\ge \frac{\binom{n}{w}2^n}{\sum_{r=0}^{d-1}\sum_{i=0}^{\min\{\lfloor r/2\rfloor,w,n-w\}}\binom{w}{i}\binom{n-w}{i}\binom{n-2i}{r-2i}2^{2i}}. \tag{12}$$

Proof. The numerator gives the total number of words with GC-content w. The denominator gives the number of these words that have distance at most $d-1$ from any fixed codeword $\mathbf{x}$. (In the denominator, $\binom{w}{i}\binom{n-w}{i}\binom{n-2i}{r-2i}2^{2i}$ is the number of words $\mathbf{y}$ with GC-content w for which $H(\mathbf{x},\mathbf{y})$ is exactly r, and for which there are exactly $w-i$ positions j with x_j and y_j both in $\{C,G\}$.)

Remark 9 Replacing $d-1$ with $\lfloor(d-1)/2\rfloor$ as the upper index of the outer summation in the denominator of (12) gives an upper-bound for $A_4^{GC}(n,d,w)$, since the balls of Hamming radius $\lfloor(d-1)/2\rfloor$ centered around codewords must be disjoint. This is an analogue of the sphere-packing bound for q-ary codes — see e.g. [20].

Now define $V(n,w,d) = \#\{\mathbf{x} \in Z_4^n : \mathbf{x} \text{ has GC-content } w \text{ and } H(\mathbf{x},\mathbf{x}^{RC}) = d\}$. Note that since no nucleotide is its own complement, $V(n,w,d)=0$ unless n and d have the same parity (i.e., are both even or are both odd).

Lemma 10 *For $n=2m$ and $d=2e$ even,*

$$V(2m,w,2e) = \sum_{i=\max\{0,w-m,\lceil(w-e)/2\rceil\}}^{\lfloor w/2\rfloor} \binom{m}{i}\binom{m-i}{w-2i}\binom{m-w+2i}{e-w+2i}2^{m+2w-4i}; \qquad (13)$$

For $n=2m+1$ and $r=2e+1$ odd,

$$V(2m+1,w,2e+1) = V(2m,w,2e) + V(2m,w-1,2e). \qquad (14)$$

Proof. In (13), the index i ranges over the number of positions $j \le m$ for which both x_j and x_{2m-j+1} belong to $\{C,G\}$. There are $\binom{m}{i}$ ways to select these positions, and $\binom{m-i}{w-2i}2^{w-2i}$ ways to select the positions for the remaining $w-2i$ occurrences of C's or G's. There are then $m-w+i$ positions $j\le m$ for which both x_j and x_{2m-j+1} belong to $\{A,T\}$. Note that the j-th coordinate of $\mathbf{x}$ necessarily differs from the j-th coordinate of $\mathbf{x}^{RC}$ in the $w-2i$ positions $j\le m$ for which one of x_j and x_{2m-j+1} is in $\{A,T\}$ and the other is in $\{C,G\}$, so there are $\binom{m-w+2i}{e-w+2i}$ ways to choose the remaining $e-w+2i$ positions $j\le m$ in which x_j differs from the complement of x_{2m+1-j}. After all these choices have been made, there are two choices for the nucleotide in each position $j\le m$; for the $m-w+2i$ positions $j\le m$ for which x_j and x_{2m-j+1} both belong to $\{C,G\}$ or both belong to $\{A,T\}$, the nucleotide at x_{2m-j+1} is forced by the choice of x_j; for the other $w-2i$ positions $j\le m$, there are two choices for the nucleotide x_{2m-j+1}.

In (14), the first summand gives the number of words with $x_{m+1}\in\{A,T\}$ and the second summand gives the number of words with $x_{m+1}\in\{C,G\}$.

Theorem 11 *For $0\le d\le n$ and $0\le w\le n$,*

$$A_4^{GC,RC}(n,d,w) \ge \frac{\sum_{r=d}^{n} V(n,d,r)}{2\sum_{r=0}^{d-1}\sum_{i=0}^{\min\{\lfloor r/2\rfloor,w,n-w\}}\binom{w}{i}\binom{n-w}{i}\binom{n-2i}{r-2i}2^{2i}}. \qquad (15)$$

Proof. The numerator gives the total number of words with GC-content w that have distance at least d from their reverse-complements, and the denominator gives an upper-bound on the number of these words that have distance at most $d-1$ from any fixed codeword. (The denominator is an upper-bound rather than an exact count, because the balls of radius $d-1$ around a word and its reverse-complement might overlap, and because when counting the number of words in these balls we may be including some words $\mathbf{y}$ that do not satisfy the condition $H(\mathbf{y},\mathbf{y}^{RC})\ge d$.)

Lexicographic codes

See [6] for an introduction to lexicographic codes. The idea is that all words in Z_q^n are listed in lexicographic order, i.e., with $\mathbf{x} = x_1 \cdots x_n$ listed before $\mathbf{y} = y_1 \cdots y_n$ if $x_i < y_i$, where i is the first position in which $\mathbf{x}$ and $\mathbf{y}$ differ. Then, starting with the empty code, one proceeds down this list and adds to the code any word whose addition does not violate any of the combinatorial constraints. (Ordinarily these would be a Hamming distance and possibly a Hamming weight constraint, but GC-content and reverse-complement Hamming distance constraints can be enforced as well.) Since the resulting lexicographic codes can accommodate no more codewords without a constraint being violated, they meet or exceed Gilbert-type lower bounds; they often do much better [6]. There are many variants of the standard lexicographic construction, for example the words may be ordered as a Gray code, or one may start with an arbitrary codeword as a seed rather than with the empty code [4]. We used three variants, singly and in combination, to construct DNA codes with the desired constraints:

(i) We used different orderings of the characters A, C, G and T when putting the 4^n DNA words of length n in lexicographic order. There are $4! = 24$ orderings of the four characters, but because of the symmetry between A and T and between C and G, only 6 of these 24 orderings need to be considered.

(ii) We used offsets, as in [19]: one starts at an arbitrary place in the list of words rather than at the beginning, and loops back around to the beginning of the list when the end is reached.

(iii) We used a "factored" ordering of the DNA words. The 2^n binary words of length n were listed in lexicographic order, $\mathbf{u}_1 = 0 \cdots 0, \ldots, \mathbf{u}_{2^n} = 1 \cdots 1$. As in [17], we define a mapping $\odot$ from pairs of binary words of length n to DNA words of length n, given by $\mathbf{x} \odot \mathbf{y} = \mathbf{z}$ where $z_i = A$ if $x_i = 0$ and $y_i = 1$; $z_i = C$ if $x_i = 1$ and $y_i = 0$; $z_i = G$ if $x_i = y_i = 1$; and $z_i = T$ if $x_i = y_i = 0$. Note that $\odot$ is a bijection, and that the Hamming weight of $\mathbf{x}$ is equal to the GC-content of $\mathbf{z}$. We ordered the 4^n DNA words so that $\mathbf{u}_i \odot \mathbf{u}_j$ comes before $\mathbf{u}_k \odot \mathbf{u}_m$ if $i < k$ or if $i = k$ and $j < m$.

When combining variants (ii) and (iii) above, two offsets can be used: one for the binary words in the first slot of $\mathbf{x} \odot \mathbf{y}$, and another for those in the second slot.

We used the above three approaches to construct DNA codes with constant GC-content, both with and without the reverse-complement constraint, for a variety of parameters n, d and w. Using offsets of zero and an average of about ten random offsets, we found codes that are larger than the codes given in [14, 14, 24] for many choices of parameters. The sizes of the lexicographic codes are given in Tables 1 and 2, and the offsets used to generate these codes are given in Tables 3 and 4.

Product bounds

The lexicographic constructions described above do not scale well to large n. One can avoid the burden of explicitly computing distances between all pairs of codewords (and also the burden of explicitly listing all codewords) by using modifications of algebraic constructions such as linear codes. For example, a DNA code with minimum Hamming distance at least d and constant GC-content w can be constructed by taking any linear code over Z_4 (or the Galois field $\mathbf{F}_4$ [5] or the Kleinian four-group [10]) that has minimum Hamming distance d, and selecting only those codewords with exactly w occurrences of two fixed characters.

In this section we give lower bounds for DNA codes that are constructed from binary codes, binary constant-weight codes, and ternary constant-weight codes, for which a variety of algebraic constructions are known (e.g. [16, 4, 19]).

Note that the reverse-complement operator RC can be viewed as the composition of two (commuting) operators R and C, where R maps $x_1 \cdots x_n$ to $x_n \cdots x_1$ and C replaces each coordinate x_i with its complement $\overline{x_i}$. We state the product bounds below in terms of constraints on R rather than on RC to make the arguments cleaner. (This approach was used in [17].) The values $A_q^R(n,d,w)$ and $A_4^{GC,R}(n,d,w)$ are defined in the same manner as $A_q^{RC}(n,d,w)$ and $A_4^{GC,RC}(n,d,w)$, but with the constraint that $H(\mathbf{x},\mathbf{y}^R) \geq d$ for all codewords $\mathbf{x}$ and $\mathbf{y}$ in place of the constraint that $H(\mathbf{x},\mathbf{y}^{RC}) \geq d$. Bounds on $A_4^{GC,R}(n,d,w)$ can be used to derive bounds for $A_4^{GC,RC}(n,d,w)$ using the following result:

Proposition 12 *For $0 \leq d \leq n$ and $0 \leq w \leq n$,*

$$A_4^{GC,RC}(n,d,w) = A_4^{GC,R}(n,d,w) \text{ if } n \text{ is even}, \tag{16}$$

$$A_4^{GC,R}(n,d+1,w) \leq A_4^{GC,RC}(n,d,w) \leq A_4^{GC,R}(n,d-1,w) \text{ if } n \text{ is odd}. \tag{17}$$

Proof. The analogous result for DNA codes with unrestricted GC-content was proved in [17], and essentially the same proof works here. Given a set of codewords of length n, if we replace all the entries in any subset of the positions by their complements, the GC-content of each codeword is preserved, as is the Hamming distance between any pair of codewords. The Hamming distance between a codeword and the reverse or reverse-complement of another codeword is not in general preserved, but if n is even and we replace the first $n/2$ coordinates of each codeword $\mathbf{x}_i$ by their complements to form a new word $\mathbf{y}_i$, then $H(\mathbf{x}_i,\mathbf{x}_j^R) = H(\mathbf{y}_i,\mathbf{y}_j^{RC})$ for all codewords $\mathbf{x}_i$ and $\mathbf{x}_j$. Similarly, if n is odd and we replace the first $(n-1)/2$ coordinates of each codeword $\mathbf{x}_i$ by their complements to form $\mathbf{y}_i$, then $|H(\mathbf{x}_i,\mathbf{x}_j^R) - H(\mathbf{y}_i,\mathbf{y}_j^{RC})| \leq 1$.

Theorem 13 *For $0 \leq d \leq n$ and $0 \leq w \leq n$,*

$$A_4^{GC}(n,d,w) \geq A_2(n,d,w) \cdot A_2(n,d) \tag{18}$$
$$A_4^{GC,R}(n,d,w) \geq A_2^R(n,d,w) \cdot A_2(n,d) \tag{19}$$
$$A_4^{GC,R}(n,d,w) \geq A_2(n,d,w) \cdot A_2^R(n,d) \tag{20}$$
$$A_4^{GC}(n,d,w) \geq A_3(n,d,w) \cdot A_2(n-w,d) \tag{21}$$
$$A_4^{GC,R}(n,d,w) \geq A_3^R(n,d,w) \cdot A_2(n-w,d) \tag{22}$$
$$A_4^{GC,R}(n,d,w) \geq A_3(n,d,w) \cdot A_2^R(n-w,d) \tag{23}$$

Proof. For (18) and (19), note that if $\mathcal{B}_1$ is a set of binary words with length n, Hamming weight w and minimum Hamming distance d, and if $\mathcal{B}_2$ is a set of binary words with length n and minimum Hamming distance d, then $\mathcal{D} = \{\mathbf{x} \odot \mathbf{y} : \mathbf{x} \in \mathcal{B}_1 \text{ and } \mathbf{y} \in \mathcal{B}_2\}$ is a set of DNA words with length n, GC-content w and minimum Hamming distance d. If, in addition, $H(\mathbf{x}_1, \mathbf{x}_2^R) \geq d$ for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{B}_1$, then $H(\mathbf{z}_1, \mathbf{z}_2^R) \geq d$ for all $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{D}$ as well, since $H(\mathbf{x}_1 \odot \mathbf{y}_1, (\mathbf{x}_2 \odot \mathbf{y}_2)^R) = H(\mathbf{x}_1 \odot \mathbf{y}_1, \mathbf{x}_2^R \odot \mathbf{y}_2^R) \geq H(\mathbf{x}_1, \mathbf{x}_2^R) \geq d$. Inequality (20) is proved in the same manner as (19).

For (21)–(23) we first define a function $\oslash$ that maps a pair consisting of ternary word $\mathbf{x}$ of length n and Hamming weight w, and a binary word $\mathbf{y}$ of length $n-w$, to a DNA word $\mathbf{z} = \mathbf{x} \oslash \mathbf{y}$ of length n. This map is defined by $z_i = C$ if $x_i = 1$; $z_i = G$ if $x_i = 2$; $z_i = A$ if x_i is the j-th zero-entry in $\mathbf{x}$ and $y_j = 0$; and $z_i = T$ if x_i is the j-th zero-entry in $\mathbf{x}$ and $y_j = 1$. The argument now proceeds as for (18)–(20).

Remark 14 Lower bounds for $A_2(n, d, w)$ can be found in [4], lower bounds for $A_2(n, d)$ in [3, 15], and lower bounds for $A_3(n, d, w)$ in [19]. The bounds on ternary constant-weight codes in [19] also apply directly to DNA codes with constant C-content over the three-letter alphabet $\{A, C, T\}$. This restricted alphabet is used by some researchers to reduce the probability of individual codewords having "secondary structure" such as hairpin loops [18, 8] — note also that if $\mathbf{x}$ and $\mathbf{y}$ are DNA words over $\{A, C, T\}$ with C-content at least d, the reverse-complement Hamming distance constraint $H(\mathbf{x}, \mathbf{y}^{RC}) \geq d$ is automatically satisfied.

Remark 15 Inequalities (18)–(20) are analogues of the product bounds for DNA codes with unrestricted GC-content in [17]; (18) is also a generalization of the "template-map" construction used in [14] for codes with constant GC-content — in that construction, a constant-weight binary code acts as the "template" (corresponding to the first factor in (18)), and the same constant-weight binary code, with at most two words of other weights added in, acts as the "map" (corresponding to the second factor in (18)). This gives a DNA code of size no larger than $A_2(n, d, w) \cdot A_2(n, d)$, and when $A_2(n, d, w) + 2 < A_2(n, d)$ this gives a strictly smaller code (e.g., $A_2(n, 2, w) = \binom{n}{w}$, which can be much less than $A_2(n, 2) = 2^{n-1}$). But for the parameters $w = d \approx n/2$ considered in [14], this difference can be inconsequential; in particular, $A_2(n, n/2, n/2) = A_2(n, n/2) - 2 = 2n - 2$ whenever a Hadamard matrix of order n exists [21], i.e. for all n divisible by 4 up to at least $n = 424$. Note that even when optimal binary codes are used as factors, the lower bounds derived from product codes are not in general tight — for instance, $A_2(12, 6, 6) \cdot A_2(12, 6) = 22 \cdot 24 = 528$, while we constructed a lexicographic code showing that $A_4^{GC}(12, 6, 6) \geq 736$. In fact, product codes do not even meet the Gilbert-type lower bound for $A_4^{GC}(2w, w, w)$ when w is sufficiently large: replacing the denominator in (12) with the upper-bound $w\binom{2w}{w-1}3^{w-1}$ for the number of words with Hamming distance at most $w-1$ from a fixed codeword gives $A_4^{GC}(2w, w, w) \geq 3(4/3)^w(w+1)/w^2$; the product-code construction gives a code of size at most $A_2(2w, w, w) \cdot A_2(2w, w) \leq (4w-2)4w$. (The "template-code" construction used in [1, 13] is similar to the template-map construction discussed above, but with an additional constraint to prevent codewords from hybridizing to concatenations of other codewords.)

Below we show that product codes can be optimal when $d = 2$:

Theorem 16 *For* $0 \le w \le n$,

$$A_4^{GC}(n, 2, w) = \binom{n}{w} 2^{n-1}. \tag{24}$$

Proof. In one direction we have $A_4^{GC}(n, 2, w) \ge A_2(n, 2, w) \cdot A_2(n, 2)$ by (18). Note that $A_2(n, 2, w) = \binom{n}{w}$ since the Hamming distance between two distinct binary words of the same weight is at least two; note also that $A_2(n, 2) = 2^{n-1}$, since the first $n-1$ coordinates can be arbitrary with the last coordinate used as a parity check bit (see e.g. [20]).

In the other direction, $A_4^{GC}(w, 2, w) = A_4^{GC}(w, 2, 0) = A_2(w, 2) = 2^{w-1} = \binom{w}{w}2^{w-1}$, and if $A_4^{GC}(n, 2, w) \le \binom{n}{w}2^{n-1}$ for some $n \ge w$ then by (8) we have $A_4^{GC}(n + 1, 2, w) \le 2(n + 1 - w)/(n + 1)\binom{n}{w}2^{n-1} = \binom{n+1}{w}2^n$. Hence by induction $A_4^{GC}(n, 2, w) \le \binom{n}{w}2^{n-1}$ for all $n \ge w$.

Theorem 17 *For* $0 \le w \le n$ *and* n *even,*

$$A_4^{GC,RC}(n, 2, w) = \binom{n}{w} 2^{n-2}. \tag{25}$$

Proof. By (12), $A_4^{GC,RC}(n, 2, w) \le \frac{1}{2}A_4^{GC}(n, 2, w) = \frac{1}{2}\binom{n}{w}2^{n-1} = \binom{n}{w}2^{n-2}$. For n even, $A_4^{GC,RC}(n, 2, w) = A_4^{GC,R}(n, 2, w)$ by (16), and $A_2^R(n, 2) = 2^{n-2}$ by Theorem 4.5 of [17]. Thus by the product bound $A_4^{GC,R}(n, d, w) \ge A_2(n, 2, w) \cdot A_2^R(n, 2) = \binom{n}{w}2^{n-2}$. (Here is an alternate argument showing $A_2^R(n, 2) = 2^{n-2}$ for n even: when n is even, the set of all 2^{n-1} binary words of odd Hamming weight contains no palindromes, and the reverse of a binary word of odd weight has odd weight, so these 2^{n-1} words break up into 2^{n-2} pairs $\{\mathbf{x}, \mathbf{x}^R\}$; taking one word from each pair shows that $A_2^R(n, 2) \ge 2^{n-2}$, since the Hamming distance between two distinct binary words of odd weight is at least two; equality follows from a halving bound, $A_2^R(n, 2) \le \frac{1}{2}A_2(n, 2) = 2^{n-2}$ [17].

Tables

Lower bounds for $A_4^{GC,RC}(n, d, w)$, derived from codes constructed using stochastic local search, are given in [23] and [24] for $n \le 12$ (n even) with $d \le n$ and $w = n/2$. In Tables 1 and 2 we give lower bounds for $A_4^{GC,RC}(n, d, w)$ and $A_4^{GC}(n, d, w)$ derived from lexicographic constructions for these same parameters. Our bounds are at least as large as those in [14, 23, 24] for all parameters except the five cases marked with asterisks; those that are strictly larger (or for which no bounds were given) are underlined. (Our bound on $A_4^{GC}(n, d, w)$ is not underlined if it is equal to twice the bound on $A_4^{GC,RC}(n, d, w)$ given in [14, 23, 24], since the former bound is then implied by the latter using the halving bound.) Entries followed by periods are optimal, as the lower bounds are equal to the upper bounds computed using Theorems 2, 4 and 7 (the Johnson-type bounds and the halving bound), except where noted.

Guide to superscripts in Tables 1 and 2:
a. Not explicitly constructed lexicographically; value from Theorem 16.
b. Not explicitly constructed lexicographically; value from Theorem 15.
c. Optimal since $A_4^{GC}(12, 11, 6) \le A_4(12, 11) \le A_4(11, 10) = 4$ by the Plotkin bound [2].
$*$. Larger code constructed using stochastic local search in [24] (size given in superscript).

Table 1. Lower bounds for $A_4^{GC,RC}(n, d, w)$ with $n \le 12$ (n even), $d \le n$ and $w = n/2$.

$n\backslash d$	2	3	4	5	6	7	8	9	10	11	12
4	24.	6.	2.	-	-	-	-	-	-	-	-
6	$\underline{320}$.	39^{*41}	$\underline{16}$	4.	2.	-	-	-	-	-	-
8	$\underline{4480}$.	384^{*390}	112	25^{*26}	10^{*12}	2.	2.	-	-	-	-
10	$\underline{64512}$.	$\underline{4084}$	$\underline{795}$	$\underline{166}$	$\underline{46}$	15	6	2.	2.	-	-
12	$\underline{946176}.^a$	$\underline{49764}$	$\underline{8704}$	$\underline{1362}$	$\underline{306}$	$\underline{81}$	$\underline{27}$	$\underline{10}$	4.	2.	2.

Table 2. Lower bounds for $A_4^{GC}(n, d, w)$ with $n \le 12$ (n even), $d \le n$ and $w = n/2$.

$n\backslash d$	2	3	4	5	6	7	8	9	10	11	12
4	48.	12.	4.	-	-	-	-	-	-	-	-
6	$\underline{640}$.	$\underline{96}$	$\underline{40}$.	8.	4.	-	-	-	-	-	-
8	$\underline{8960}$.	$\underline{832}$	224	$\underline{56}$	20^{*24}	$\underline{5}$.	4.	-	-	-	-
10	$\underline{129024}$.	$\underline{9344}$	$\underline{1676}$	$\underline{360}$	$\underline{96}$	$\underline{32}$	$\underline{16.}$	$\underline{5}$.	4.	-	-
12	$\underline{1892352}.^b$	$\underline{112640}$	$\underline{17408}$	$\underline{2992}$	$\underline{736}$	$\underline{177}$	$\underline{68}$	$\underline{22}$	8.	$4.^c$	4.

Remark 18 In [23] and [14], lower bounds for $A_4^{GC}(n, d, w)$ are given for $4 \le n \le 20$ (n odd or even) with $w = d = \lfloor n/2 \rfloor$. Though not covered in Table 2, we also improved upon these bounds for $n = 5$, 7, 9, 11 and 13–20 using lexicographic constructions.

Tables 3 and 4 record the nucleotide-orderings and the offsets used in constructing the lexicographic codes whose sizes are given in Tables 1 and 2. Entries are written either in the form $\text{offset}_1 \odot \text{offset}_2$ for the "factored" lexicographic variant, or as offset^k, with the superscript $k \in \{1, 2, 3, 4, 5, 6\}$ indicating the nucleotide-ordering used, as follows: $1 = (A < C < G < T)$; $2 = (C < G < A < T)$; $3 = (A < T < C < G)$; $4 = (C < A < T < G)$; $5 = (C < A < G < T)$; $6 = (A < C < T < G)$. Note that we list all offsets in base-16 rather than base-4 or base-2 for compactness, and that the offset need not itself be a codeword since it may not satisfy the GC-content constraint or it may be too close to its own reverse-complement. (We re-used seeds for our random-number generator, which is why some of the same "random" offsets appear for more than one entry.)

Table 3. Offsets used to generate lexicographic codes giving lower bounds in Table 1.

$n \backslash d$	2	3	4	5	6	7	8	9	10	11	12
4	59^1	59^2	0^1	–	–	–	–	–	–	–	–
6	0^1	42d^4	$12 \odot 19$	bfc^2	0^1	–	–	–	–	–	–
8	5021^1	44dd^2	$4\text{e} \odot 95$	d3de^5	$90\text{a}5^5$	0^1	0^1	–	–	–	–
10	0^1	0^5	bfc99^1	0^5	0^1	c0d96^1	c54c6^2	0^1	0^1	–	–
12	–	$0 \odot 0$	$0 \odot 0$	0^2	4121c8^4	0^5	0^2	96c697^1	96c697^1	0^1	0^1

Table 4. Offsets used to generate lexicographic codes giving lower bounds in Table 2.

$n \backslash d$	2	3	4	5	6	7	8	9	10	11	12
4	0^1	0^1	0^1	–	–	–	–	–	–	–	–
6	0^1	0^2	434^1	0^1	0^1	–	–	–	–	–	–
8	0^1	5021^2	$0 \odot 0$	$2\text{d} \odot 23$	90f6^1	0^1	0^1	–	–	–	–
10	0^1	$0 \odot 0$	0^2	$0 \odot 0$	$0 \odot 0$	$0 \odot 0$	c8e60^5	3792d^2	0^1	–	–
12	–	$0 \odot 0$	$0 \odot 0$	$0 \odot 0$	$0 \odot 0$	c8e605^1	$994 \odot \text{70b}$	$0 \odot 0$	0^2	0^1	0^1

References

[1] M. Arita and S. Kobayashi. DNA sequence design using templates. New Generation Computing, vol. 20 (2002), 263–277.

[2] G. T. Bogdanova, A. E. Brouwer, S. N. Kapralov, and P. R. J. Östergård. Error-correcting codes over an alphabet of four elements. Designs, Codes and Cryptography, vol. 23 (2001), 333–342.

[3] A. E. Brouwer. Bounds on the size of linear codes. In *Handbook of Coding Theory* (editors V. S. Pless and W. C. Huffman), North-Holland, 1998, pp. 295–461,

[4] A. E. Brouwer, J. B. Shearer, N. J. A. Sloane, and W. D. Smith. A new table of constant weight codes. IEEE Trans. Inform. Theory, vol. 36 (1990), 1334–1380.

[5] A. R. Calderbank, E. M. Rains, P. W. Shor, and N. J. A. Sloane. Quantum error correction via codes over $GF(4)$, IEEE Trans. Inform. Theory, vol. 44 (1998) 1369–1387.

[6] J. H. Conway and N. J. A. Sloane. Lexicographic codes: error-correcting codes from game theory. IEEE Trans. Inform. Theory, vol. 32 (1986), 337–348.

[7] T. Etzion. Optimal constant weight codes over Z_k and generalized designs. Discrete Math., vol. 169 (1997), 55–82.

[8] D. Faulhammer, A. R. Cukras, R. J. Lipton, and L. F. Landweber. Molecular computation: RNA solutions to chess problems. PNAS, vol. 97 (2000), 1385–1389.

[9] A. G. Frutos, Q. Liu, A. J. Thiel, A. M. W. Sanner, A. E. Condon, L. M. Smith and R. M. Corn. Demonstration of a word design strategy for DNA computing on surfaces. Nucleic Acids Research, vol. 25 (1997), 4748–4757.

[10] G. Höhn. Self-dual codes over the Kleinian four group. Preprint, available electronically at `arXiv:math.CO/0005266` .

[11] S. M. Johnson. A new upper bound for error-correcting codes. IRE Trans. Inform. Theory, vol. 8 (1962), 203-207.

[12] S. M. Johnson. Upper bounds for constant weight error-correcting codes. Discrete Math., vol. 3 (1972), 109-124.

[13] S. Kobayashi, T. Kondo, M. Arita. On template method for DNA sequence design. In *DNA Computing: 8th International Workshop on DNA-Based Computers* (editors M. Hagiya and A. Ohuchi), Springer LNCS vol. 2568, 2002, pp. 205–214.

[14] M. Li, H. J. Lee, A. E. Condon, and R. M. Corn. DNA word design strategy for creating sets of non-interacting oligonucleotides for DNA microarrays. Langmuir, vol. 18 (2002), 805–812.

[15] S. Litsyn. An updated tables of the best binary codes known. In *Handbook of Coding Theory* (editors V. S. Pless and W. C. Huffman), North-Holland, 1998, pp. 463–498,

[16] F. J. MacWilliams and N. J. A. Sloane. *The Theory of Error-Correcting Codes*, North Holland, 1977.

[17] A. Marathe, A. E. Condon, and R. M. Corn. On combinatorial DNA word design. Journal of Computational Biology, vol. 8 (2001), 201–220.

[18] K. U. Mir. A restricted genetic alphabet for DNA computing. In *DNA Based Computers II* (editors L. F. Landweber and E. B. Baum), AMS/DIMACS, 1999, pp. 243-246.

[19] P. R. J. Östergård and M. Svanström. Ternary constant weight codes. Electronic Journal of Combinatorics, vol. 9 (2002), R41, 23pp.

[20] V. S. Pless, W. C. Huffman and R. A. Brualdi. An introduction to algebraic codes. In *Handbook of Coding Theory* (editors V. S. Pless and W. C. Huffman), North-Holland, 1998, pp. 3–139,

[21] N. V. Semakov and V. A. Zinoviev. Balanced codes and tactical configurations. Problems Inform. Transmission, vol. 5 (1969), 22–28.

[22] M. Svanström, P. R. J. Östergård, and G. T. Bogdanova. Bounds and constructions for ternary constant-composition codes. IEEE Trans. Inform. Theory, vol. 48 (2002), 101–111.

[23] D. C. Tulpan, H. H. Hoos, and A. E. Condon. Stochastic local search algorithms for DNA word design. In *DNA Computing: 8th International Workshop on DNA-Based Computers* (editors M. Hagiya and A. Ohuchi), Springer LNCS vol. 2568, 2002, pp. 229–241.

[24] D. C. Tulpan and H. H. Hoos. Hybrid randomised neighbourhoods improve stochastic local search for DNA Code design. In *Advances in Artificial Intelligence: 16th Conference of the Canadian Society for Computational Studies of Intelligence* (editors Y. Xiang and B. Chaib-draa), Springer LNCS vol. 2671, 2003, pp. 418–433.

[25] R. J. M. Vaessens, E. H. L. Aarts, and J. H. van Lint. Genetic algorithms in coding theory - a table for $A_3(n, d)$. Discrete Applied Math., vol. 45 (1993), 71–87.