

# DNA Codes for Nonadditive Stem Similarity

A. G. D'yachkov<sup>a</sup>, A. N. Kuzina<sup>a</sup>, N. A. Polyansky<sup>a</sup>,  
A. Macula<sup>b</sup>, and V. V. Rykov<sup>c</sup>

*Probability Theory Chair, Faculty of Mechanics and Mathematics,  
Lomonosov Moscow State University, Moscow, Russia*

agd-msu@yandex.ru    vorronina@gmail.com    nikitapolyansky@gmail.com

*Department of Mathematics, State University of New York at Geneseo, USA*

macula@geneseo.edu

*University of Nebraska Omaha, USA*

vrykov@unomaha.edu

Received June 24, 2013; in final form, December 16, 2013

**Abstract**—DNA sequences are sequences with elements from the quaternary DNA alphabet  $\{A, C, G, T\}$ . An important property of them is their directedness and ability to form duplexes as a result of hybridization process, i.e., coalescing two oppositely directed sequences. In biological experiments exploiting this property it is necessary to generate an ensemble of such sequences (DNA codes) consisting of pairs of DNA sequences referred to as Watson–Crick duplexes. Furthermore, for any two words of the DNA code that do not form a Watson–Crick duplex, hybridization energy—stability measure of a potential DNA duplex—is upper bounded by a constant specified by conditions of an experiment. This problem can naturally be interpreted in terms of coding theory. Continuing our previous works, we consider a nonadditive similarity function for two DNA sequences, which most adequately models their hybridization energy. For the maximum cardinality of DNA codes based on this similarity, we establish a Singleton upper bound and present an example of an optimal construction. Using ensembles of DNA codes with special constraints on codewords, which we call Fibonacci ensembles, we obtain a random-coding lower bound on the maximum cardinality of DNA codes under this similarity function.

**DOI:** 10.1134/S0032946014030041

## 1. INTRODUCTION

In [1–3], in the study of codes correcting insertions and deletions, which are important for many applications, there was introduced a similarity function for two  $n$ -sequences composed of symbols of a  $q$ -ary,  $q \geq 2$ , alphabet  $\mathcal{A}_q = \{0, 1, \dots, q-1\}$  which is uniquely defined as the *length of their largest common subsequence* and is referred to as *deletion similarity*. In contrast to the *Hamming similarity*, which underlies classical coding theory [4] and is *additively* computed as the number of positions where two  $q$ -ary  $n$ -sequences coincide, the deletion similarity can be viewed as an example of a *nonadditive similarity*. Obviously, a  $q$ -ary code of length  $n$  corrects  $D$  deletions,  $1 \leq D < n$ , if and only if the similarity for any pair of its words is  $\leq n - D + 1$ . In [1–3] for a fixed  $D$ ,  $1 \leq D < n$ , there were found upper bounds for the maximum cardinality of such codes; presently known and close-to-optimal constructions have been found for  $D = 1$  only and were proposed in [5]. For the rate defining the logarithmic asymptotics of the maximum cardinality of codes correcting a linearly growing number of deletions, best random-coding lower bounds were obtained in [6].

The present paper is devoted to developing combinatorial methods of coding theory for DNA sequences and studying DNA codes (see Section 2) introduced in [7, 8] for the *nonadditive stem  $w$ -similarity* defined by a collection of “thermodynamic weights”  $w = w(a, b) > 0$ ,  $a \in \mathcal{A}_4$ ,  $b \in \mathcal{A}_4$ ,

in the “nearest-neighbor” model, which adequately describes hybridization of DNA strands [9]. Information-theoretical and probabilistic methods for the analysis of DNA codes with *additive* stem  $w$ -similarity, interesting from the mathematical point of view, were considered in [10]. In [11, 12] there were also substantially developed applications and motivation for the definition of DNA codes for both the Hamming and deletion similarities, and for a linearly growing similarity there were constructed random-coding lower bounds for the rate of DNA codes.<sup>1</sup> In [12] it was shown that the optimal construction of [5] can be extended to the case of DNA codes.

In Section 2 we give key definitions and introduce the main notation required in the paper. In Section 3 we present a nontrivial example of an optimal construction of DNA codes for nonadditive stem similarity. In Section 4 we formulate a lower bound on the rate  $R^{(w)}(d)$  of DNA codes based on the nonadditive  $w$ -similarity called the random coding bound and obtained on ensembles of DNA codes for which admissible sequences are constrained in the same way as for binary Fibonacci sequences; this bound was previously announced in [14]. For this bound, proved in Section 6, we establish the main result of the paper: we compute the value  $d(w)$ , called the critical distance fraction for the random coding bound for the nonadditive stem  $w$ -distance, such that the rate  $R^{(w)}(d) > 0$  for  $0 < d < d(w)$ . In Section 5 we analyze and compare some known [9] experimental estimates of weights  $w = w(a, b)$ . The comparison is based on computing the values  $d(w)$  in the same way as it was done in [15] with the help of computing critical distance fractions  $T_w$  (using formulas from [10]) for optimal DNA codes with additive stem  $w$ -similarity.

## 2. NOTATION AND DEFINITIONS

Let  $\triangleq$  denote equality by definition, and let  $[n] \triangleq \{1, 2, \dots, n\}$  be the set of integers from 1 to  $n$ . Single DNA strands (or DNA sequences) are considered in this paper as sequences with elements from the DNA alphabet

$$\mathcal{A}_4 \triangleq \{A, T, C, G\} \triangleq \{0, 1, 2, 3\} \quad (1)$$

whose letters denote the corresponding nucleic acids (bases). For any  $x \in \mathcal{A}_4$  we define its *complementary* element  $\bar{x} \in \mathcal{A}_4$  by the rule  $\bar{A} \triangleq T$  ( $\bar{0} \triangleq 1$ ),  $\bar{C} \triangleq G$  ( $\bar{2} \triangleq 3$ ), and vice versa. Let

$$\mathbf{x} \triangleq (x_1, x_2, \dots, x_n) \in \mathcal{A}_4^n \quad \text{and} \quad \mathbf{y} \triangleq (y_1, y_2, \dots, y_n) \in \mathcal{A}_4^n$$

be arbitrary two DNA  $n$ -sequences. For a sequence  $\mathbf{x} \in \mathcal{A}_4^n$  define its *reverse complementary* sequence  $\tilde{\mathbf{x}} \triangleq (\bar{x}_n, \bar{x}_{n-1}, \dots, \bar{x}_2, \bar{x}_1) \in \mathcal{A}_4^n$ , also referred to as its Watson–Crick transform. If  $\mathbf{y} \triangleq \tilde{\mathbf{x}}$ , then  $\mathbf{x} = \tilde{\tilde{\mathbf{y}}}$  for any  $\mathbf{x} \in \mathcal{A}_4^n$ . If  $\mathbf{x} = \tilde{\mathbf{x}}$ , then  $\mathbf{x}$  is said to be a *self-reverse-complementary* sequence. If  $\mathbf{x} \neq \tilde{\mathbf{x}}$ , then a pair  $(\mathbf{x}, \tilde{\mathbf{x}})$  is referred to as a *pair of mutually reverse complementary* sequences.

By

$$\mathbf{z} = (z_1, z_2, \dots, z_\ell) \in \mathcal{A}_4^\ell, \quad \ell \in [n],$$

we denote a *common subsequence* [3] of length  $|\mathbf{z}| \triangleq \ell$  between  $\mathbf{x}$  and  $\mathbf{y}$ ; this means that there exist  $\ell$ -sequences of integers

$$1 \leq k_1 < k_2 < \dots < k_\ell \leq n, \quad 1 \leq j_1 < j_2 < \dots < j_\ell \leq n$$

such that  $z_u = x_{k_u} = y_{j_u}$ ,  $u \in [\ell]$ . By definition, the *empty* subsequence  $\mathbf{z}$  of length  $|\mathbf{z}| \triangleq 0$  is a common subsequence between any two sequences  $\mathbf{x}$  and  $\mathbf{y}$ .

**Definition 1.** Let  $2 \leq r \leq n$  be arbitrary integers. A DNA  $r$ -sequence  $\mathbf{a} = (a_1, a_2, \dots, a_r) \in \mathcal{A}_4^r$  is called a *common block for sequences  $\mathbf{x}$  and  $\mathbf{y}$*  (a *common  $(\mathbf{x}, \mathbf{y})$ -block*) of length  $r$  if both  $\mathbf{x}$  and  $\mathbf{y}$

<sup>1</sup> Note the most well-known survey [13] of constructions of linear quaternary ( $q = 4$ ) DNA codes with the Hamming similarity (distance).

contain  $\mathbf{a}$  as a subsequence consisting of  $r$  consecutive elements of  $\mathbf{x}$  and  $\mathbf{y}$ . We say that a common  $(\mathbf{x}, \mathbf{y})$ -block  $\mathbf{a}$  is *equivalent to  $r - 1$  common stems*  $a_i, a_{i+1}, i \in [r - 1]$ , containing two neighboring symbols of this  $(\mathbf{x}, \mathbf{y})$ -block.

**Definition 2.** Let  $\ell, 2 \leq \ell \leq n$ , be an integer. We say that a sequence  $\mathbf{z} = (z_1, z_2, \dots, z_\ell) \in \mathcal{A}_4^\ell$  is a *common block subsequence* of length  $|\mathbf{z}| \triangleq \ell$  between  $\mathbf{x}$  and  $\mathbf{y}$  if  $\mathbf{z}$  is an *ordered collection* of disjoint common  $(\mathbf{x}, \mathbf{y})$ -blocks and the length of each common  $(\mathbf{x}, \mathbf{y})$ -block in this collection is  $\geq 2$ . Let  $\mathcal{Z}(\mathbf{x}, \mathbf{y})$  denote the set of all common block subsequences between  $\mathbf{x}$  and  $\mathbf{y}$ . For any  $\mathbf{z} \in \mathcal{Z}(\mathbf{x}, \mathbf{y})$  denote by  $k(\mathbf{x}, \mathbf{y}, \mathbf{z}), 1 \leq k(\mathbf{z}, \mathbf{x}, \mathbf{y}) \leq |\mathbf{z}|/2$ , the *minimum number* of common  $(\mathbf{x}, \mathbf{y})$ -blocks *forming* this subsequence  $\mathbf{z}$ .

Note that the difference  $|\mathbf{z}| - k(\mathbf{x}, \mathbf{y}, \mathbf{z})$  equals the total number of stems containing neighboring symbols in common  $(\mathbf{x}, \mathbf{y})$ -blocks forming  $\mathbf{z} \in \mathcal{Z}(\mathbf{x}, \mathbf{y})$ .

**Definition 3.** For sequences  $\mathbf{x}, \mathbf{y} \in \mathcal{A}_4^n$ , the number

$$\mathcal{S}^{(1)}(\mathbf{x}, \mathbf{y}) \triangleq \max_{\mathbf{z} \in \mathcal{Z}(\mathbf{x}, \mathbf{y})} \{|\mathbf{z}| - k(\mathbf{x}, \mathbf{y}, \mathbf{z})\}, \quad \mathcal{S}^{(1)}(\mathbf{x}, \mathbf{y}) \geq 0, \quad (2)$$

is called the *nonadditive stem 1-similarity* between  $\mathbf{x}$  and  $\mathbf{y}$ . We say that  $\mathcal{S}^{(1)}(\mathbf{x}, \mathbf{y}) \triangleq 0$  if and only if  $\mathcal{Z}(\mathbf{x}, \mathbf{y}) = \emptyset$ . Clearly,

$$\mathcal{S}^{(1)}(\mathbf{x}, \mathbf{y}) = \mathcal{S}^{(1)}(\mathbf{y}, \mathbf{x}) \leq \mathcal{S}^{(1)}(\mathbf{x}, \mathbf{x}) = n - 1.$$

*Example.* For  $n = 10$  consider the sequences

$$\mathbf{x} = (A, T, \underbrace{T, A, A, A, A, T, T, A}, \quad \mathbf{y} \triangleq \tilde{\mathbf{x}} = (\underbrace{T, A, A, T, T, T, T, A, A, T}).$$

The maximizing common block subsequence between  $\mathbf{x}$  and  $\mathbf{y}$  is

$$\mathbf{z} \triangleq (\overbrace{T, A, A, T, T, A}) = \tilde{\mathbf{z}} = (x_3, x_4, x_5, x_8, x_9, x_{10}) = (y_1, y_2, y_3, y_6, y_7, y_8) \in \mathcal{Z}(\mathbf{x}, \mathbf{y}).$$

Then  $k(\mathbf{x}, \mathbf{y}, \mathbf{z}) = 2$ , and the corresponding nonadditive stem 1-similarity is

$$\mathcal{S}^{(1)}(\mathbf{x}, \mathbf{y}) \triangleq \max_{\mathbf{z} \in \mathcal{Z}(\mathbf{x}, \mathbf{y})} \{|\mathbf{z}| - k(\mathbf{x}, \mathbf{y}, \mathbf{z})\} = 6 - 2 = 4.$$

The maximum is attained at the self-reverse-complementary sequence  $\mathbf{z} \in \mathcal{Z}(\mathbf{x}, \mathbf{y})$  given above. Note that the same similarity value can also be obtained for the common block subsequence

$$\mathbf{z} \triangleq (\overbrace{A, T, T, T, T, A}) = (x_1, x_2, x_3, x_8, x_9, x_{10}) = (y_3, y_4, y_5, y_6, y_7, y_8) \in \mathcal{Z}(\mathbf{x}, \mathbf{y}),$$

which is not self-reverse-complementary.

Let  $w = w(a, b) > 0, a, b \in \mathcal{A}_4$ , be a *weight function* such that

$$w(a, b) = w(\bar{b}, \bar{a}), \quad a, b \in \mathcal{A}_4. \quad (3)$$

Condition (3) means that  $w(a, b)$  is invariant under the Watson–Crick transform. Table 1 presents an example of *unified weights* [9] that can be used as a biologically motivated weight function  $w(a, b)$  in the definition of a similarity function (4).

**Definition 4** [7, 8]. Let  $\mathbf{z} \in \mathcal{Z}(\mathbf{x}, \mathbf{y})$  be of the form

$$\mathbf{z} \triangleq (z^1, z^2, \dots, z^{k(\mathbf{x}, \mathbf{y}, \mathbf{z})}), \quad |\mathbf{z}| = \sum_{m=1}^{k(\mathbf{x}, \mathbf{y}, \mathbf{z})} |z^m| = \sum_{m=1}^{k(\mathbf{x}, \mathbf{y}, \mathbf{z})} r_m,$$

**Table 1.** Unified weights

$w(a, b)$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	1.00	1.44	<b>1.28</b>	0.88
$a = C$	1.45	1.84	2.17	<b>1.28</b>
$a = G$	1.30	2.24	1.84	1.44
$a = T$	0.58	1.30	1.45	1.00

where

$$\mathbf{z}^m \triangleq (z_1^m, z_2^m, \dots, z_{r_m}^m) \in \mathcal{A}_q^{r_m}, \quad m = 1, 2, \dots, k(\mathbf{x}, \mathbf{y}, \mathbf{z}),$$

is an ordered collection of common  $(\mathbf{x}, \mathbf{y})$ -blocks forming  $\mathbf{z}$  and  $r_m \triangleq |\mathbf{z}^m| \geq 2$  denotes the length of a block  $\mathbf{z}^m$ . For DNA sequences  $\mathbf{x}, \mathbf{y} \in \mathcal{A}_4^n$ , the number

$$\mathcal{S}^{(w)}(\mathbf{x}, \mathbf{y}) \triangleq \max_{\mathbf{z} \in \mathcal{Z}(\mathbf{x}, \mathbf{y})} \left\{ \sum_{m=1}^{k(\mathbf{x}, \mathbf{y}, \mathbf{z})} \sum_{i=1}^{r_m-1} w(z_i^m, z_{i+1}^m) \right\} \quad (4)$$

is called the *nonadditive stem  $w$ -similarity* between  $\mathbf{x}$  and  $\mathbf{y}$ . We assume that  $\mathcal{S}^{(w)}(\mathbf{x}, \mathbf{y}) \triangleq 0$  if and only if  $\mathcal{Z}(\mathbf{x}, \mathbf{y}) = \emptyset$ .

**Proposition 1.** For any  $\mathbf{x}, \mathbf{y} \in \mathcal{A}_4^n$ ,

$$\mathcal{S}^{(w)}(\mathbf{x}, \mathbf{y}) = \mathcal{S}^{(w)}(\mathbf{y}, \mathbf{x}) \leq \mathcal{S}^{(w)}(\mathbf{x}, \mathbf{x}). \quad (5)$$

Moreover,

$$\mathcal{S}^{(w)}(\mathbf{x}, \tilde{\mathbf{y}}) = \mathcal{S}^{(w)}(\mathbf{y}, \tilde{\mathbf{x}}), \quad \mathbf{x}, \mathbf{y} \in \mathcal{A}^n. \quad (6)$$

The symmetry property and inequality (5) are obvious. Equality (6) follows from definition (4) and condition (3). Identity (6) ensures a symmetry property for *hybridization energy*  $E(\mathbf{x}, \mathbf{y}) = E(\mathbf{y}, \mathbf{x})$  of two oppositely directed DNA sequences  $\mathbf{x}, \mathbf{y} \in \mathcal{A}^n$  which in the general case [11] is defined via the similarity function  $S(\mathbf{x}, \mathbf{y})$  as

$$E(\mathbf{x}, \mathbf{y}) = E(\mathbf{y}, \mathbf{x}) \triangleq S(\mathbf{x}, \tilde{\mathbf{y}}) = S(\mathbf{y}, \tilde{\mathbf{x}}).$$

Clearly, the stem similarity  $\mathcal{S}^{(1)}(\mathbf{x}, \mathbf{y})$  from Definition 3 corresponds to the case of a constant weight function:  $w(a, b) \equiv 1$  for any  $a, b \in \mathcal{A}_4$ .

**Definition 5** [7, 8]. The number

$$\mathcal{D}^{(w)}(\mathbf{x}, \mathbf{y}) \triangleq \mathcal{S}^{(w)}(\mathbf{x}, \mathbf{x}) - \mathcal{S}^{(w)}(\mathbf{x}, \mathbf{y}), \quad \mathbf{x}, \mathbf{y} \in \mathcal{A}_4^n, \quad (7)$$

is called the *nonadditive stem  $w$ -distance* between DNA sequences  $\mathbf{x}$  and  $\mathbf{y}$ .

From Proposition 1 we obtain

$$\mathcal{D}^{(w)}(\mathbf{x}, \mathbf{y}) \geq \mathcal{D}^{(w)}(\mathbf{x}, \mathbf{x}) = 0, \quad \mathbf{x}, \mathbf{y} \in \mathcal{A}_4^n. \quad (8)$$

*Remark 1.* As in [10], we consider a nonsymmetric distance function (7) for which in most cases  $\mathcal{D}^{(w)}(\mathbf{x}, \mathbf{y}) \neq \mathcal{D}^{(w)}(\mathbf{y}, \mathbf{x})$ ; it is motivated by biological experiments [8]. We could also define a symmetric distance function

$$\mathcal{D}_{\text{sym}}^{(w)}(\mathbf{x}, \mathbf{y}) \triangleq \frac{\mathcal{S}^{(w)}(\mathbf{x}, \mathbf{x}) + \mathcal{S}^{(w)}(\mathbf{y}, \mathbf{y})}{2} - \mathcal{S}^{(w)}(\mathbf{x}, \mathbf{y}), \quad \mathbf{x}, \mathbf{y} \in \mathcal{A}^n. \quad (9)$$

However, distance (7) is more useful from the point of view of biological applications described in [8]. Therefore, in what follows we study the nonsymmetric distance (7) only. Note that a similar

analysis of the symmetric distance (9) can also be made using methods developed in the present paper. Note also that coding theory [4] traditionally is based on the notion of a distance, not of a similarity. But in our case the initial biological construction assumes a similarity function as a mathematical model of hybridization energy, and distance is an artificial construction arising not from biological background but from needs of mathematical analysis of the obtained objects.

Let sequences  $\mathbf{x}(j) \triangleq (x_1(j), x_2(j), \dots, x_n(j)) \in \mathcal{A}_4^n$ ,  $j \in [N]$ , be *codewords* of a *code*  $\mathcal{X} = \{\mathbf{x}(1), \mathbf{x}(2), \dots, \mathbf{x}(N)\}$  of *length*  $n$  and *cardinality*  $N$ , where  $N = 2, 4, \dots$  is an even integer. Let  $D$ ,  $0 < D \leq \max_{\mathbf{x} \in \mathcal{A}_4^n} \mathcal{S}^{(w)}(\mathbf{x}, \mathbf{x})$ , be a positive integer. Taking into account (7) and (8), we give the following definition.

**Definition 6.** A code  $\mathcal{X}$  is called an  $(n, D)^{(w)}$  *code based on the nonadditive stem  $w$ -similarity*  $\mathcal{S}^{(w)}(\mathbf{x}, \mathbf{y})$  (an  $(n, D)^{(w)}$  *code*) if the following two conditions are satisfied:

- (i) For any  $j \in [N]$  there exists  $j' \in [N]$ ,  $j' \neq j$ , such that  $\mathbf{x}(j') = \widetilde{\mathbf{x}(j)} \neq \mathbf{x}(j)$ . In other words,  $\mathcal{X}$  is a collection of  $N/2$  pairs of mutually reverse complementary sequences;
- (ii) For any  $j, j' \in [N]$  with  $j \neq j'$ , the nonadditive stem  $w$ -distance

$$\mathcal{D}^{(w)}(\mathbf{x}(j), \mathbf{x}(j')) \geq D.$$

The following statement is obvious.

**Proposition 2.** Let a weight function  $w(a, b)$ ,  $a, b \in \mathcal{A}_4$ , satisfying condition (3) be constant, i.e.,

$$w(a, b) \equiv 1, \quad a, b \in \mathcal{A}_4. \quad (10)$$

The corresponding symmetric distance function  $\mathcal{D}^{(1)}(\mathbf{x}, \mathbf{y})$ ,  $\mathbf{x}, \mathbf{y} \in \mathcal{A}_4^n$ , is of the form

$$\mathcal{D}^{(1)}(\mathbf{x}, \mathbf{y}) = \mathcal{D}^{(1)}(\mathbf{y}, \mathbf{x}) \triangleq (n - 1) - S^{(1)}(\mathbf{x}, \mathbf{y}), \quad (11)$$

where the nonadditive stem 1-similarity  $S^{(1)}(\mathbf{x}, \mathbf{y})$  is defined in (2), and condition (ii) in Definition 6 for an  $(n, D)^{(1)}$  DNA code,  $0 < D \leq n - 1$ , is given by

$$S^{(1)}(\mathbf{x}(j), \mathbf{x}(j')) \leq (n - 1) - D, \quad j, j' \in [N], \quad j \neq j'. \quad (12)$$

**Definition 7.** Let  $N^{(w)}(n, D)$  denote the *maximum cardinality* of  $(n, D)^{(w)}$  DNA codes based on the nonadditive stem  $w$ -similarity. If  $d$ ,  $0 < d < \max_{(a,b) \in \mathcal{A}_4^2} w(a, b)$ , is a fixed positive number, then

$$R^{(w)}(d) \triangleq \lim_{n \rightarrow \infty} \frac{\log_4 N^{(w)}(n, dn)}{n}, \quad 0 < d < \max_{(a,b) \in \mathcal{A}_4^2} w(a, b), \quad (13)$$

is called the *rate* of  $(n, dn)^{(w)}$  DNA codes for *distance fraction*  $d$ .

### 3. BOUNDS AND CONSTRUCTIONS FOR $(n, D)^{(1)}$ CODES

**Theorem 1** [16]. If  $2 \leq D \leq n/2$ , then the maximum cardinality of an  $(n, D)^{(1)}$  DNA code satisfies the inequalities

$$N^{(1)}(n, D) \leq \begin{cases} \frac{4^{n-D+1} - 4^{(n-D+1)/2} + 4^{D-1}}{2} & \text{if } n - D + 1 \text{ is even,} \\ \frac{4^{n-D+1} + 4^{D-1}}{2} & \text{if } n - D + 1 \text{ is odd.} \end{cases} \quad (14)$$

In the next theorem we present a construction of an  $(n, 2)^{(1)}$  DNA code whose cardinality coincides with the upper bound.

**Theorem 2.** *If  $D = 2$  and  $n$  is even, then the maximum cardinality of an  $(n, 2)^{(1)}$  DNA code precisely equals the upper bound:*

$$N^{(1)}(n, 2) = \frac{4^{n-1} + 4}{2}. \quad (15)$$

**Proof of Theorem 1.** Let  $k = n - D + 1$ . Consider an arbitrary  $(n, D)^{(1)}$  DNA code

$$\mathcal{X} = \{\mathbf{x}(j) \in \mathcal{A}^n, j \in [N]\}.$$

Let  $\mathbf{a} = (a_1 a_2 \dots a_k) \in \mathcal{A}_4^k$  and  $\mathbf{b} = (b_1 b_2 \dots b_{n-k}) \in \mathcal{A}_4^{n-k}$  denote DNA sequences of lengths  $|\mathbf{a}| = k$  and  $|\mathbf{b}| = n - k = D - 1$ , respectively.

**Lemma 1.** *If  $k$  is even and  $\mathbf{a} = \tilde{\tilde{\mathbf{a}}}$  in  $\mathcal{A}_4^k$ , i.e.,  $\mathbf{a}$  is a self-reverse-complementary DNA sequence, then the  $(n, D)^{(1)}$  DNA code  $\mathcal{X}$  does not contain words of the form  $(\mathbf{a}\mathbf{b})$ ,  $\mathbf{b} \in \mathcal{A}_4^{n-k}$ .*

**Proof.** Assume that there exist  $j \in [N]$  and a word  $\mathbf{x}(j) = (\mathbf{a}\mathbf{b})$ . Then by property (i) of an  $(n, D)^{(1)}$  DNA code (see Definition 6) there is a number  $j' \in [N]$ ,  $j' \neq j$ , such that  $\mathbf{x}(j') = (\tilde{\tilde{\mathbf{b}}}\mathbf{a})$ . The words  $\mathbf{x}(j)$  and  $\mathbf{x}(j')$  have a common block  $\mathbf{a}$ ,  $|\mathbf{a}| = k$ . Hence we conclude that  $\mathcal{D}^{(1)}(\mathbf{x}(j), \mathbf{x}(j')) < D$ , which contradicts property (ii) of an  $(n, D)^{(1)}$  DNA code.  $\triangle$

Define integers  $m = 1, 2, \dots$  and  $r = 0, 1, \dots, D - 2$  satisfying the equation  $k = (D - 1)m + r$ . Any DNA sequence  $\mathbf{b} \in \mathcal{A}_4^{D-1}$  can be written as

$$\mathbf{b} = (\mathbf{b}_2 \mathbf{b}_1), \quad \mathbf{b}_1 \in \mathcal{A}^r, \quad \mathbf{b}_2 \in \mathcal{A}^{D-1-r}.$$

Let  $(\mathbf{a}, \tilde{\tilde{\mathbf{a}}})$ ,  $\mathbf{a} \neq \tilde{\tilde{\mathbf{a}}}$ , be a pair of mutually reverse complementary DNA sequences, and let there exist  $j, j' \in [N]$ ,  $j \neq j'$ , such that  $\mathbf{x}(j) = (\mathbf{a}\mathbf{b})$  and  $\mathbf{x}(j') = (\tilde{\tilde{\mathbf{a}}}\mathbf{b}')$ . By property (i) of an  $(n, D)^{(1)}$  DNA code, there exist  $j'' \in [N]$ ,  $j'' \neq j'$ , and a word  $\mathbf{x}(j'') = (\tilde{\tilde{\mathbf{b}}}\mathbf{a})$ . Then  $\mathcal{D}^{(1)}(\mathbf{x}(j), \mathbf{x}(j'')) < D$ , whence, using property (ii) of an  $(n, D)^{(1)}$  DNA code, we conclude that  $j'' = j$   $\mathbf{x}(j) = \mathbf{x}(j'')$ . Furthermore, it is easily seen that the word  $\mathbf{x}(j)$  is determined from the DNA sequence  $\mathbf{b} = (\mathbf{b}_2 \mathbf{b}_1) \in \mathcal{A}^{D-1}$  and is of the form

$$\mathbf{x}(j) = (\underbrace{\mathbf{b}_1 \mathbf{b}_2 \mathbf{b}_1 \mathbf{b}_2 \mathbf{b}_1 \dots \mathbf{b}_2 \mathbf{b}_1}_{\mathbf{a}}). \quad (16)$$

Define a set  $\mathcal{X}_1$ ,  $\mathcal{X}_1 \subseteq \mathcal{X}$ , containing all words of  $\mathcal{X}$  of the form (16). Let  $\mathcal{X}_2 \triangleq \mathcal{X} \setminus \mathcal{X}_1$ . The cardinality  $N$  of the DAN code  $\mathcal{X}$  is  $N = |\mathcal{X}| = |\mathcal{X}_1| + |\mathcal{X}_2|$ . It follows from Lemma 1 and equation (16) that the cardinalities  $|\mathcal{X}_i|$ ,  $i = 1, 2$ , satisfy the inequalities

$$|\mathcal{X}_1| \leq 4^{D-1} \quad \text{and} \quad |\mathcal{X}_2| \leq \begin{cases} \frac{4^k - 4^{k/2} - |\mathcal{X}_1|}{2} & \text{if } k \text{ is even,} \\ \frac{4^k - |\mathcal{X}_1|}{2} & \text{if } k \text{ is odd.} \end{cases}$$

Hence,

$$N = |\mathcal{X}_1| + |\mathcal{X}_2| \leq \begin{cases} \frac{4^k - 4^{k/2} + |\mathcal{X}_1|}{2} \leq \frac{4^k - 4^{k/2} + 4^D}{2} & \text{if } k \text{ is even,} \\ \frac{4^k + |\mathcal{X}_1|}{2} \leq \frac{4^k + 4^D}{2} & \text{if } k \text{ is odd.} \end{cases}$$

Theorem 1 is proved.  $\triangle$

**Proof of Theorem 2.** Let us interpret the DNA alphabet (1), i.e., the set  $\mathcal{A}_4 = \{0, 1, 2, 3\}$ , as a Galois field with a standard addition operation  $\oplus$ :

$$\begin{aligned} 0 \oplus 0 &\triangleq 1 \oplus 1 \triangleq 2 \oplus 2 \triangleq 3 \oplus 3 \triangleq 0, & 0 \oplus 1 &\triangleq 1 \oplus 0 \triangleq 2 \oplus 3 \triangleq 3 \oplus 2 \triangleq 1, \\ 0 \oplus 2 &\triangleq 2 \oplus 0 \triangleq 1 \oplus 3 \triangleq 3 \oplus 1 \triangleq 2, & 0 \oplus 3 &\triangleq 3 \oplus 0 \triangleq 1 \oplus 2 \triangleq 2 \oplus 1 \triangleq 3. \end{aligned}$$

If  $n$  is even, then for any  $a \in \{0, 3\} \subset \mathcal{A}_4$  the equations

$$x_1 = x_2 \oplus x_3 \oplus \dots \oplus x_{n-1} \oplus a, \quad x_n = x_1 \oplus 2 \quad (17)$$

define a set  $\mathcal{X}_a \in \mathcal{A}_4^n$  consisting of DNA sequences  $\mathbf{x} = (x_1 x_2 \dots x_n)$ . Two sets  $\mathcal{X}_a$ ,  $a \in \{0, 3\}$ , possess the obvious properties:

1.  $|\mathcal{X}_a| = 4^{n-2}$ ,  $a \in \{0, 3\}$ , and  $\mathcal{X}_0 \cap \mathcal{X}_3 = \emptyset$ ;
2. If  $x \in \mathcal{X}_0$ , then  $\tilde{x} \in \mathcal{X}_3$ , and vice versa.

To prove that  $\mathcal{X}_0 \cup \mathcal{X}_3$  is an  $(n, 2)^{(1)}$  DNA code of cardinality  $2 \cdot 4^{n-2}$ , it remains to show that  $\mathcal{D}^{(1)}(\mathbf{x}, \mathbf{y}) \geq 2$  for any  $\mathbf{x}, \mathbf{y} \in \mathcal{X}_0 \cup \mathcal{X}_3$ ,  $\mathbf{x} \neq \mathbf{y}$ . Assume the contrary, i.e., let  $\mathbf{x}$  and  $\mathbf{y}$  have a common block of length  $(n-1)$ . We consider two cases.

Case 1. Let  $\mathbf{x} \in \mathcal{X}_a$  and  $\mathbf{y} \in \mathcal{X}_a$ . Then, without loss of generality,

- either a common block is a subsequence  $(x_1 x_2 \dots x_{n-1}) = (y_2 y_3 \dots y_n)$ , which is impossible by (17) because of the contradiction

$$a = x_1 \oplus x_2 \oplus \dots \oplus x_{n-1} = y_2 \oplus y_3 \oplus \dots \oplus y_n = a \oplus 2,$$

- or  $(x_1 x_2 \dots x_{n-1}) = (y_1 y_2 \dots y_{n-1})$ , which is also impossible because of  $\mathbf{x} \neq \mathbf{y}$ .

Case 2. Let  $\mathbf{x} \in \mathcal{X}_0$  and  $\mathbf{y} \in \mathcal{X}_3$ . Then, without loss of generality,

- either a common block is a subsequence  $(x_1 x_2 \dots x_{n-1}) = (y_1 y_2 \dots y_{n-1})$ , which is impossible by (17) because of the contradiction

$$0 = x_1 \oplus x_2 \oplus \dots \oplus x_{n-1} = y_1 \oplus y_2 \oplus \dots \oplus y_{n-1} = 3,$$

- or  $(x_1 x_2 \dots x_{n-1}) = (y_2 y_3 \dots y_n)$ , which is also impossible by (17) because of the contradiction

$$0 = x_1 \oplus x_2 \oplus \dots \oplus x_{n-1} = y_2 \oplus y_3 \oplus \dots \oplus y_n = 1.$$

Define a set  $\mathcal{X}'$  of cardinality  $|\mathcal{X}'| = n + 2$  as the set containing words of the following form:

$$(\underbrace{00 \dots 0}_{2t} \underbrace{22 \dots 2}_{n-2t}), \quad (\underbrace{33 \dots 3}_{2t} \underbrace{11 \dots 1}_{n-2t}), \quad \text{for any } t \in \{0 \cup [n/2]\}.$$

Similarly, define a set  $\mathcal{X}''$  of cardinality  $|\mathcal{X}''| = n$  as the set containing words of the form

$$(\underbrace{00 \dots 0}_{2t-1} \underbrace{22 \dots 2}_{n-2t+1}), \quad (\underbrace{33 \dots 3}_{2t-1} \underbrace{11 \dots 1}_{n-2t+1}), \quad \text{for any } t \in [n/2].$$

It is clear that  $\mathcal{X}'$  and  $\mathcal{X}''$  are  $(n, 2)^{(1)}$  DNA codes and that

$$\mathcal{X}' \cap \{\mathcal{X}_0 \cup \mathcal{X}_3\} = \emptyset \quad \text{and} \quad \mathcal{X}'' \subseteq \{\mathcal{X}_0 \cup \mathcal{X}_3\}.$$

Now consider the set  $\mathcal{X} = \{\mathcal{X}_0 \cup \mathcal{X}_3 \cup \mathcal{X}'\} \setminus \mathcal{X}''$  of cardinality  $N = |\mathcal{X}| = \frac{4^{n-1} + 4}{2}$ . It is easily seen that it is an  $(n, 2)^{(1)}$  DNA code. Indeed, it suffices to check that  $\mathcal{D}^{(1)}(\mathbf{x}, \mathbf{y}) \geq 2$  for  $\mathbf{x} \in \{\mathcal{X}_0 \cup \mathcal{X}_3\} \setminus \mathcal{X}''$  and  $\mathbf{y} \in \mathcal{X}'$ . Assume the contrary. Without loss of generality, we may assume that

$$\mathbf{y} = (\underbrace{00 \dots 0}_{2t} \underbrace{22 \dots 2}_{n-2t}), \quad \mathbf{y} \in \mathcal{X}'$$

for some  $t \in \{0 \cup [n/2]\}$ . Then  $\mathbf{x}$  must be of the form

$$\mathbf{x} = (\underbrace{00 \dots 0}_k \underbrace{22 \dots 2}_{n-k});$$

i.e.,  $\mathbf{x} \in \mathcal{X}' \cup \mathcal{X}''$ . This is a contradiction, since  $\mathbf{x} \in \{\mathcal{X}_0 \cup \mathcal{X}_3\} \setminus \mathcal{X}''$ .  $\triangle$

## 4. RANDOM CODING BOUNDS

## 4.1. DNA Codes for Fibonacci Ensembles

In what follows, by  $L$  we denote any subset  $L \subset \mathcal{A}_4^2$  of 2-blocks of symbols of the DNA alphabet closed under the reverse complement transformation. Thus, for instance,  $L = L_k$ ,  $k = 0, 1, 2, 4$ , where

$$L_0 \triangleq \emptyset, \quad L_1 \triangleq \{TA\}, \quad L_2 \triangleq \{TA, AT\}, \quad L_4 \triangleq \{TA, AT, AA, TT\}. \quad (18)$$

Let  $DNA(n, L)$  (or, for brevity,  $[n, L]$ ) denote the set (ensemble) of all DNA sequences of length  $n$  that do not contain stems from  $L$ . Clearly,  $[n, L]$  is closed under the reverse complement transformation. We will call  $[n, L]$  the *Fibonacci  $L$ -ensemble*.<sup>2</sup> Denote by  $\lambda_L(n) \triangleq |DNA(n, L)| = |[n, L]|$  the *cardinality* of  $[n, L]$ .

**Definition 8.** Let  $N_L(n, D)$  denote the *maximum cardinality* of  $(n, D)^{(1)}$  DNA codes  $\mathcal{X} \subseteq DNA(n, L)$ . If distance fraction  $d > 0$  is a fixed number, then

$$R_L(d) \triangleq \overline{\lim}_{n \rightarrow \infty} \frac{\log_4 N_L(n, dn)}{n} \quad (19)$$

is called the *rate* of DNA codes for the Fibonacci  $L$ -ensemble.

For a weight function (3), define

$$\underline{w}_L \triangleq \min_{(a,b) \notin L} w(a, b). \quad (20)$$

For example, if  $w(a, b)$  is given by Table 1, then for the sets  $L$  from (18) we obtain

$$\underline{w}_L = \begin{cases} 0.58 & \text{if } L = L_0, \\ 0.88 & \text{if } L = L_1, \\ 1.00 & \text{if } L = L_2, \\ 1.28 & \text{if } L = L_4. \end{cases} \quad (21)$$

Construction of a lower bound for the rate  $R^{(w)}(d)$  is based on the following statement.

**Proposition 3.** Let  $\underline{w}_L$  be defined in (20), and let  $\mathcal{X} \subset DNA(n, L)$ . If  $\mathcal{X}$  is an  $(n, D)^{(1)}$  DNA code, then  $\mathcal{X}$  is also an  $(n, \underline{w}_L D)^{(w)}$  code. Thus, the rate (13) satisfies the inequality

$$R^{(w)}(d) \geq \max_L R_L\left(\frac{d}{\underline{w}_L}\right), \quad (22)$$

where  $R_L(d)$  is defined in (19).

**Proof of Proposition 3.** Following [8], let us show that for any sequences  $\mathbf{x}, \mathbf{y} \in [n, L]$  we have

$$\mathcal{D}^{(w)}(\mathbf{x}, \mathbf{y}) \geq \underline{w}_L \mathcal{D}^{(1)}(\mathbf{x}, \mathbf{y}). \quad (23)$$

For any common block subsequence  $\mathbf{z}$  between  $\mathbf{x}$  and  $\mathbf{y}$ , define two sets of stems (with repetitions). The first set consists of stems that form blocks of the sequence  $\mathbf{z}$  (see Definition 4):

$$\mathcal{M}_1(\mathbf{x}, \mathbf{y}, \mathbf{z}) \triangleq \{(z_i^j, z_{i+1}^j) \in \mathcal{A}_4^2, \quad j \in [k(\mathbf{x}, \mathbf{y}, \mathbf{z})], \quad i \in [r_j - 1]\}.$$

<sup>2</sup> Binary sequences over the alphabet  $\mathcal{A}_2 = \{0, 1\}$  that do not contain 2-stems of the form  $(1, 1)$  are known as Fibonacci sequences [17].

The second set,  $\mathcal{M}_2(\mathbf{x}, \mathbf{y}, \mathbf{z})$ , contains all other stems of  $\mathbf{x}$  that are not contained in  $\mathcal{M}_1(\mathbf{x}, \mathbf{y}, \mathbf{z})$ . For example, if

$$\mathbf{x} = (A, T, C, A, C, A), \quad \mathbf{y} = (T, C, A, G, C, A), \quad \mathbf{z} = (T, C, A, C, A),$$

then

$$\mathcal{M}_1(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \{(T, C), (C, A), (C, A)\}, \quad \mathcal{M}_2(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \{(A, T), (A, C)\}.$$

Then, obviously,

$$\mathcal{S}^{(w)}(\mathbf{x}, \mathbf{y}) = \sum_{(a,b) \in \mathcal{M}_1(\mathbf{x}, \mathbf{y}, \mathbf{z}(\mathbf{x}, \mathbf{y}))} w(a, b),$$

where  $\mathbf{z}(\mathbf{x}, \mathbf{y})$  is a block subsequence maximizing the similarity between  $\mathbf{x}$  and  $\mathbf{y}$  (see equation (4) in Definition 4). Moreover,

$$\mathcal{D}^{(w)}(\mathbf{x}, \mathbf{y}) = \sum_{(a,b) \in \mathcal{M}_2(\mathbf{x}, \mathbf{y}, \mathbf{z}(\mathbf{x}, \mathbf{y}))} w(a, b) \geq \underline{w}_L |\mathcal{M}_2(\mathbf{x}, \mathbf{y}, \mathbf{z}(\mathbf{x}, \mathbf{y}))|. \quad (24)$$

On the other hand, for any common block subsequence  $\mathbf{z}$  between  $\mathbf{x}$  and  $\mathbf{y}$ , we have  $|\mathcal{M}_1(\mathbf{x}, \mathbf{y}, \mathbf{z})| = |\mathbf{z}| - k(\mathbf{x}, \mathbf{y}, \mathbf{z})$ , where  $k(\mathbf{x}, \mathbf{y}, \mathbf{z})$  was introduced in Definition 2, and therefore,

$$|\mathcal{M}_2(\mathbf{x}, \mathbf{y}, \mathbf{z})| = (n - 1) - |\mathcal{M}_1(\mathbf{x}, \mathbf{y}, \mathbf{z})| = (n - 1) - (|\mathbf{z}| - k(\mathbf{x}, \mathbf{y}, \mathbf{z})).$$

By equality (2), the similarity  $\mathcal{S}^{(1)}(\mathbf{x}, \mathbf{y})$  is the maximum value of  $|\mathbf{z}| - k(\mathbf{x}, \mathbf{y}, \mathbf{z})$  over all common block subsequences  $\mathbf{z}$  between  $\mathbf{x}$  and  $\mathbf{y}$ . Therefore, the distance  $\mathcal{D}^{(1)}(\mathbf{x}, \mathbf{y})$  defined in (11) can be written as

$$\mathcal{D}^{(1)}(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{z} \in \mathcal{Z}(\mathbf{x}, \mathbf{y})} \{(n - 1) - (|\mathbf{z}| - k(\mathbf{x}, \mathbf{y}, \mathbf{z}))\},$$

where the minimum is over all common block subsequences  $\mathbf{z}$  between  $\mathbf{x}$  and  $\mathbf{y}$ . Thus,

$$\mathcal{D}^{(1)}(\mathbf{x}, \mathbf{y}) \leq |\mathcal{M}_2(\mathbf{x}, \mathbf{y}, \mathbf{z}(\mathbf{x}, \mathbf{y}))| \quad (25)$$

for any  $\mathbf{x}, \mathbf{y} \in \mathcal{A}_4^n$ . Combining (24) and (25), we obtain (23). Inequality (23) immediately implies (22).  $\triangle$

#### 4.2. On Cardinalities of Fibonacci $L$ -Ensembles

If  $L = L_0 = \emptyset$ , then  $\lambda_L(n) = 4^n$ . If  $L \neq \emptyset$ , then the cardinalities  $\lambda_L(1) = 4$  and  $\lambda_L(2) = 16 - |L|$  are determined. For nontrivial sets  $L$  defined in (18), we compute cardinalities  $\lambda_L(n)$ ,  $n \geq 3$ , using a well-known result for second-order linear recurrence equations [18] presented in Proposition 4.

**Proposition 4.** *Let  $a_1$  and  $a_2$  be arbitrary fixed numbers. If a sequence  $x_n$ ,  $n = 3, 4, \dots$ , satisfies a recurrence equation*

$$x_n = a_1 x_{n-1} + a_2 x_{n-2}, \quad (26)$$

then

$$x_n = c_1 r_1^n + c_2 r_2^n, \quad (27)$$

where  $r_1$  and  $r_2$ ,  $|r_1| > |r_2|$ , are roots of the characteristic equation

$$z^2 - a_1 z - a_2 = 0,$$

and  $c_i$ ,  $i = 1, 2$ , can be computed from the initial conditions

$$\begin{cases} x_1 = X_1, \\ x_2 = X_2. \end{cases}$$

Thus,

$$x_n \leq Cr^n [1 + \omega\alpha^n], \quad (28)$$

where

$$\begin{aligned} C &\triangleq |c_1| > 0, & r &\triangleq \max\{|r_1|, |r_2|\} = |r_1| > 0, \\ \omega &\triangleq \left| \frac{c_2}{C} \right| > 0, & \alpha &\triangleq \left| \frac{r_2}{r} \right| < 1. \end{aligned} \quad (29)$$

*Remark 2.* In our case, we additionally have  $c_1 > 0$  and  $r_1 > 0$  and also  $C = c_1$  and  $r = r_1$ , since it is clear that the number  $\lambda_L(n)$  of sequences in an  $L$ -ensemble  $DNA(n, L)$  is always nonnegative for any  $n$ .

**Lemma 2.** For the sets  $L_1$ ,  $L_2$ , and  $L_4$  defined in (18), we have the following statements.

1. If  $L = L_1$ , then  $\lambda_L(n)$  satisfies (26), where  $a_1 = 4$  and  $a_2 = -1$ . Thus, parameters (29) of the bound (28) are

$$\begin{aligned} C &= \frac{3 + 2\sqrt{3}}{6} = 1.08, & r &= 2 + \sqrt{3} = 3.73, \\ \omega &= 7 - 4\sqrt{3} = 0.0718, & \alpha &= 7 - 4\sqrt{3} = 0.0718. \end{aligned} \quad (30)$$

2. If  $L = L_2$ , then  $\lambda_L(n)$  satisfies (26), where  $a_1 = 3$  and  $a_2 = 2$ . Thus, parameters (29) of the bound (28) are

$$\begin{aligned} C &= \frac{17 + 5\sqrt{17}}{34} = 1.11, & r &= \frac{3 + \sqrt{17}}{2} = 3.56, \\ \omega &= \frac{21 - 5\sqrt{17}}{4} = 0.0961, & \alpha &= \frac{13 - 3\sqrt{17}}{4} = 0.158. \end{aligned} \quad (31)$$

3. If  $L = L_4$ , then  $\lambda_L(n)$  satisfies (26), where  $a_1 = 2$  and  $a_2 = 4$ . Thus, parameters (29) of the bound (28) are

$$\begin{aligned} C &= \frac{5 + 3\sqrt{5}}{10} = 1.17, & r &= 1 + \sqrt{5} = 3.24, \\ \omega &= \frac{7 - 3\sqrt{5}}{2} = 0.146, & \alpha &= \frac{3 - \sqrt{5}}{2} = 0.382. \end{aligned} \quad (32)$$

*Remark 3.* In Lemma 2 we collect results of computing the cardinalities of Fibonacci  $L$ -ensembles that we use below in maximization according to (22) for the analysis of all real weight functions known from [9] and obtained experimentally in different biology laboratories. Results for a number of other  $L$ -ensembles similar to Lemma 2 are presented in [19].

Let us introduce auxiliary notation. Let  $a, b \in \{A, C, G, T\}$  be arbitrary symbols from the DNA alphabet (1), and let

$$\begin{aligned} [n, L]_a &\triangleq \{\mathbf{x} : \mathbf{x} \in [n, L] \text{ and } x_n = a\}, \\ [n, L]_{a,b} &\triangleq \{\mathbf{x} : \mathbf{x} \in [n, L] \text{ and } x_{n-1} = a, x_n = b\} \end{aligned}$$

denote the corresponding subsets of  $[n, L]$ . If  $(a, b) \in L$ , then  $[n, L]_{a,b} = \emptyset$ . Furthermore,  $[n, L]_a$  and  $[n, L]$  can be represented as sums of the following disjoint subsets:

$$\begin{aligned} [n, L]_a &= [n, L]_{A,a} + [n, L]_{C,a} + [n, L]_{G,a} + [n, L]_{T,a}, \\ [n, L] &= [n, L]_A + [n, L]_C + [n, L]_G + [n, L]_T. \end{aligned} \quad (33)$$

It is also easily seen that the following two properties hold:

1. If for any  $b \in \{A, C, G, T\}$  the pair  $(b, a) \notin L$ , then

$$|[n, L]_a| = |[n - 1, L]| = \lambda_L(n - 1). \quad (34)$$

2. For any pair  $(a, b) \notin L$ ,

$$|[n, L]_{a,b}| = |[n - 1, L]_a|. \quad (35)$$

**Proof of Lemma 2.** 1. Let  $L = \{TA\}$ . Taking into account (33)–(35), we have

$$\begin{aligned}\lambda_L(n) &= 3\lambda_L(n-1) + |[n, L]_{A,A}| + |[n, L]_{C,A}| + |[n, L]_{G,A}| \\ &= 3\lambda_L(n-1) + |[n-1, L]_A| + |[n-1, L]_C| + |[n-1, L]_G| \\ &= 3\lambda_L(n-1) + 2\lambda_L(n-2) + |[n-1, L]_A|\end{aligned}$$

and

$$\lambda_L(n-1) = 3\lambda_L(n-2) + |[n-1, L]_A|.$$

From these formulas we obtain the recurrence equation

$$\lambda_L(n) = 4\lambda_L(n-1) - \lambda_L(n-2), \quad n = 3, 4, \dots$$

2. Let  $L = \{TA, AT\}$ . Taking into account (33)–(35), we have

$$\begin{aligned}\lambda_L(n) &= 2\lambda_L(n-1) + |[n, L]_{A,A}| + |[n, L]_{G,A}| + |[n, L]_{C,A}| + |[n, L]_{C,T}| + |[n, L]_{G,T}| + |[n, L]_{T,T}| \\ &= 2\lambda_L(n-1) + |[n-1, L]_A| + 2|[n-1, L]_C| + 2|[n-1, L]_G| + |[n-1, L]_T| \\ &= 2\lambda_L(n-1) + 4\lambda_L(n-2) + |[n-1, L]_A| + |[n-1, L]_T|\end{aligned}$$

and

$$\lambda_L(n-1) = 2\lambda_L(n-2) + |[n-1, L]_A| + |[n-1, L]_T|.$$

From these formulas we obtain the recurrence equation

$$\lambda_L(n) = 3\lambda_L(n-1) - 2\lambda_L(n-2), \quad n = 3, 4, \dots$$

3. Let  $L = \{TA, AT, AA, TT\}$ . Taking into account (33)–(35), for  $n = 3, 4, \dots$  we have

$$\begin{aligned}\lambda_L(n) &= 2\lambda_L(n-1) + |[n, L]_{C,A}| + |[n, L]_{G,A}| + |[n, L]_{C,T}| + |[n, L]_{G,T}| \\ &= 2\lambda_L(n-1) + 2|[n-1, L]_C| + 2|[n-1, L]_G| \\ &= 2\lambda_L(n-1) + 4\lambda_L(n-2). \quad \triangle\end{aligned}$$

#### 4.3. Random Coding Bound for a Fibonacci $L$ -Ensemble

Let

$$\rho_L \triangleq \log_4 r, \quad \rho'_L \triangleq \log_4 \frac{r}{C^3(1 + \omega\alpha^2)(1 + \omega\alpha)^2}, \quad (36)$$

where  $r = r(L)$ ,  $C = C(L)$ ,  $\alpha = \alpha(L)$ , and  $\omega = \omega(L)$  are introduced in Proposition 4 and are given by formulas (29). For sets  $L$  defined in (18), parameters (29) are computed by formulas (30)–(32). In Section 6, using the random coding method elaborated in [12], we prove the following result.

**Theorem 3.** For any distance fraction  $0 < d < d_L$  the rate (19) satisfies the inequality

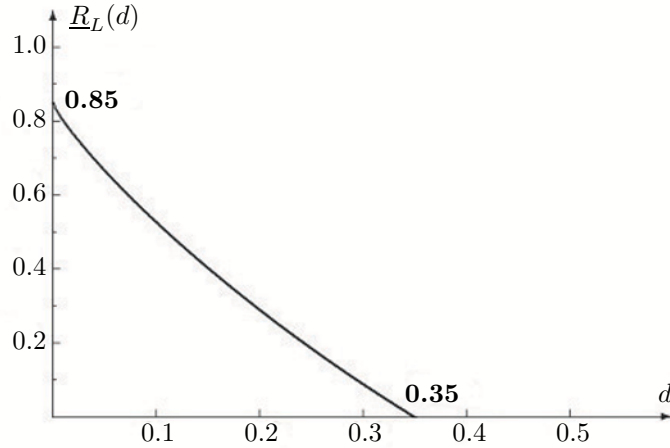
$$R_L(d) \geq \underline{R}_L(d) \triangleq (1-d)\rho_L - E_L(d) > 0, \quad (37)$$

where

$$\begin{aligned}E_L(u) &\triangleq \max_{0 \leq v \leq \min\{u, 1-u\}} E^L(v, u), \\ E^L(v, u) &\triangleq -\rho'_L v + (1-u)h_4\left(\frac{v}{1-u}\right) + 2uh_4\left(\frac{v}{u}\right), \\ h_4(u) &\triangleq -u \log_4 u - (1-u) \log_4 (1-u),\end{aligned} \quad (38)$$

and  $d_L$ ,  $0 < d_L < 1/2$ , is a unique root of the equation  $\underline{R}_L(d) = 0$ , or  $(1-d)\rho_L = E_L(d)$ , in the interval  $(0; 1/2)$ .

The function  $\underline{R}_L(d)$  is referred to as the *random coding bound* for  $R_L(d)$ . The number  $d_L$ ,  $0 < d_L < 1/2$ , will be referred to as the *critical distance fraction* for the random coding bound  $\underline{R}_L(d)$  for a  $DNA(n, L)$  ensemble. Figure 1 presents the graph of  $\underline{R}_L(d)$  for  $L = L_4$ .



**Fig. 1.** Random coding bound  $\underline{R}_L(d)$  for  $L = L_4 = \{TA, AT, AA, TT\}$ .

Our computations based on Lemma 2 yield the following numerical values of the critical distance fraction for the sets (18):

$$d_L = \begin{cases} 0.4794, & L = L_0 = \emptyset, \\ 0.4316, & L = L_1 = \{TA\}, \\ 0.4054, & L = L_2 = \{TA, AT\}, \\ 0.3487, & L = L_4 = \{TA, AT, AA, TT\}. \end{cases} \quad (39)$$

#### 4.4. Random Coding Bound for $(n, dn)^{(w)}$ DNA Codes

Let

$$d(w) \triangleq \max_L \{\underline{w}_L d_L\}, \quad \underline{w}_L \triangleq \min_{(a,b) \notin L} w(a,b). \quad (40)$$

The results of Proposition 3 and Theorem 3 imply the following fact.

**Theorem 4.** *If  $0 < d < d(w)$ , then the rate (13) of  $(n, dn)^{(w)}$  DNA codes is positive and we have a lower bound*

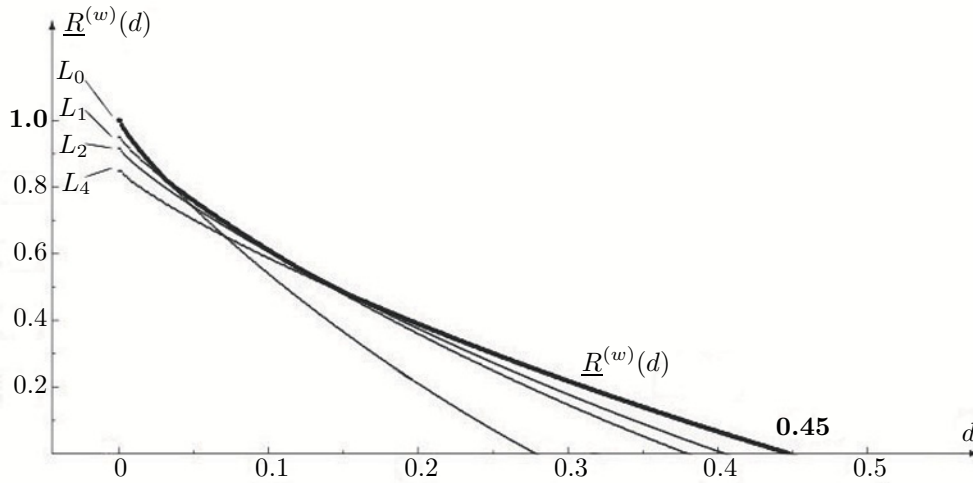
$$R^{(w)}(d) \geq \underline{R}^{(w)}(d) \triangleq \max_{L: d_L > d/\underline{w}_L} \left\{ \underline{R}_L\left(\frac{d}{\underline{w}_L}\right) \right\}, \quad 0 < d < d(w).$$

The function  $\underline{R}^{(w)}(d)$  is called the random coding bound for  $(n, dn)^{(w)}$  DNA codes. The number  $d(w)$  is called the *critical distance fraction* for the random coding bound  $\underline{R}^{(w)}(d)$ . Figure 2 presents graphs of the functions  $\underline{R}_L\left(\frac{d}{\underline{w}_L}\right)$  for  $L = L_0, L_1, L_2, L_4$  and the maximal bound  $\underline{R}^{(w)}(d)$  for the unified weights  $w(a, b)$ ,  $a, b \in \mathcal{A}_4$ , from Table 1.

## 5. ANALYSIS OF WEIGHT SAMPLES BASED ON THE CRITICAL DISTANCE FRACTION CRITERION

### 5.1. Tables of Weight Samples

In this subsection we discuss experimentally obtained *samples of weight functions* (or *weight samples*)  $w = w(a, b)$ ,  $a, b \in \mathcal{A}_4$ , borrowed from [9, Table 1]; they are presented on the left-hand



**Fig. 2.** Lower bound  $\underline{R}^{(w)}(d)$  and functions of the form  $\underline{R}_L\left(\frac{d}{\underline{w}_L}\right)$  for  $L = L_0, L_1, L_2, L_4$  for the weight sample  $w(a, b)$ ,  $a, b \in \mathcal{A}_4$ , from Table 1.

side of Tables 2–8. Right-hand sides of these tables contain collections of *relative* weights  $\tilde{w}(a, b)$  with the base  $w(A, A)$ ; i.e., for any  $a, b \in \mathcal{A}_4$ ,

$$\tilde{w} = \tilde{w}(a, b) \triangleq \frac{w(a, b)}{w(A, A)}, \quad \tilde{w}(a, b) = \tilde{w}(\bar{b}, \bar{a}).$$

Dimensionless numbers  $\tilde{w}(a, b)$  are more convenient to compare with each other and with unified weights from Table 1.

### 5.2. Analysis of Tables 1–8 for the Nonadditive Stem $w$ -Distance

Analysis of Tables 1–7. For Tables 1–4 and 7, the maximum in (40) is attained at the same set

$$L_4 = \{AT, TA, AA, TT\}, \quad \text{where } d_{L_4} = 0.3487. \quad (41)$$

For Tables 5 and 6, the maximum in (40) is attained at

$$L_2 = \{AT, TA\}, \quad \text{where } d_{L_2} = 0.4054. \quad (42)$$

Captions of Tables 1–7 present the computed critical distance fractions  $d(\tilde{w})$ , and weights  $\underline{w}_{L_4}$  or  $\underline{w}_{L_2}$  corresponding to them are given in the tables in bold.

For instance, for the weight sample  $\tilde{w} = \tilde{w}(a, b)$  from Table 1, from weights (21) and numbers (39) one obtains

$$\underline{w}_L d_L = \begin{cases} 0.28 & \text{if } L = L_0, \\ 0.38 & \text{if } L = L_1, \\ 0.41 & \text{if } L = L_2, \\ 0.45 & \text{if } L = L_4. \end{cases}$$

Thus, the corresponding critical distance fraction is

$$d(\tilde{w}) \triangleq \max_L \{\underline{w}_L d_L\} = \underline{w}_{L_4} d_{L_4} = 0.45.$$

In other words, for this weight sample and for the nonadditive stem distance we have shown that if occurrence of pairs from the set  $L_4 = \{TA, AT, AA, TT\}$  in admissible DNA sequences is forbidden,

**Table 2.**  $d(\tilde{w}) = 0.67$ 

$w(a, b)$	$b = A$	$b = C$	$b = G$	$b = T$	$\tilde{w}(a, b)$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	0.43	0.98	0.83	0.27	$a = A$	1.00	2.28	<b>1.93</b>	0.63
$a = C$	0.97	1.22	1.70	0.83	$a = C$	2.32	2.84	3.95	<b>1.93</b>
$a = G$	0.93	1.64	1.22	0.98	$a = G$	2.16	3.81	2.84	2.28
$a = T$	0.22	0.93	0.97	0.43	$a = T$	0.51	2.16	2.32	1.00

**Table 3.**  $d(\tilde{w}) = 0.47$ 

$w(a, b)$	$b = A$	$b = C$	$b = G$	$b = T$	$\tilde{w}(a, b)$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	0.89	1.35	1.16	0.81	$a = A$	1.00	<b>1.35</b>	1.52	0.91
$a = C$	1.37	1.64	1.99	1.16	$a = C$	1.54	1.84	2.24	1.52
$a = G$	1.25	1.96	1.64	1.35	$a = G$	1.40	2.20	1.84	<b>1.35</b>
$a = T$	0.76	1.25	1.37	0.89	$a = T$	0.85	1.40	1.54	1.00

**Table 4.**  $d(\tilde{w}) = 0.58$ 

$w(a, b)$	$b = A$	$b = C$	$b = G$	$b = T$	$\tilde{w}(a, b)$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	0.67	1.13	1.17	0.62	$a = A$	1.00	1.69	1.75	0.93
$a = C$	1.19	1.55	1.87	1.17	$a = C$	1.78	2.31	2.79	1.75
$a = G$	1.12	1.85	1.55	1.13	$a = G$	<b>1.67</b>	2.76	2.31	1.69
$a = T$	0.70	1.12	1.19	0.67	$a = T$	1.04	<b>1.67</b>	1.78	1.00

**Table 5.**  $d(\tilde{w}) = 0.41$ 

$w(a, b)$	$b = A$	$b = C$	$b = G$	$b = T$	$\tilde{w}(a, b)$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	0.93	1.52	1.03	0.83	$a = A$	1.00	1.63	<b>1.11</b>	0.89
$a = C$	1.26	1.67	1.65	1.03	$a = C$	1.35	1.80	1.77	<b>1.11</b>
$a = G$	1.56	2.44	1.67	1.52	$a = G$	1.68	2.62	1.80	1.63
$a = T$	0.70	1.56	1.26	0.93	$a = T$	0.75	1.68	1.35	1.00

**Table 6.**  $d(\tilde{w}) = 0.41$ 

$w(a, b)$	$b = A$	$b = C$	$b = G$	$b = T$	$\tilde{w}(a, b)$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	1.02	1.43	1.16	0.73	$a = A$	1.00	1.40	<b>1.14</b>	0.72
$a = C$	1.38	1.77	2.09	1.16	$a = C$	1.35	1.74	2.05	<b>1.14</b>
$a = G$	1.46	2.28	1.77	1.43	$a = G$	1.43	2.24	1.74	1.40
$a = T$	0.60	1.46	1.38	1.02	$a = T$	0.59	1.43	1.35	1.00

**Table 7.**  $d(\tilde{w}) = 0.44$ 

$w(a, b)$	$b = A$	$b = C$	$b = G$	$b = T$	$\tilde{w}(a, b)$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	1.20	1.50	1.50	0.90	$a = A$	1.00	1.25	<b>1.25</b>	0.75
$a = C$	1.70	2.10	2.80	1.50	$a = C$	1.42	1.75	2.33	<b>1.25</b>
$a = G$	1.50	2.30	2.10	1.50	$a = G$	1.25	1.92	1.75	1.25
$a = T$	0.90	1.50	1.70	1.20	$a = T$	0.75	1.25	1.42	1.00

**Table 8.**  $d(\tilde{w}) = 0.29$ 

$w(a, b)$	$b = A$	$b = C$	$b = G$	$b = T$	$\tilde{w}(a, b)$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	1.66	1.13	1.35	1.19	$a = A$	1.00	<b>0.68</b>	0.81	0.72
$a = C$	1.80	2.75	3.28	1.35	$a = C$	1.08	1.66	1.98	0.81
$a = G$	1.41	2.82	2.75	1.13	$a = G$	0.85	1.70	1.66	<b>0.68</b>
$a = T$	0.76	1.41	1.80	1.66	$a = T$	0.46	0.85	1.08	1.00

Table 9

	T1	T2	T3	T4	T5	T6	T7	T8*
$L$ for (40)	$L_4$	$L_4$	$L_4$	$L_4$	$L_2$	$L_2$	$L_4$	$L_1$
$d(\tilde{w})$	0.45	0.67	0.47	0.58	0.41	0.41	0.44	0.29
$T_{\tilde{w}}$	1.58	2.60	1.61	1.97	1.58	1.55	1.50	—

then the critical distance fraction  $d(\tilde{w})$  of the random coding bound for  $(n, dn)^{(\tilde{w})}$  DNA codes can be increased from 0.28 to 0.45.

Analysis of Table 8. The maximum in (40) is attained at the set  $L_1 = \{TA\}$ ; for this set,  $d_{L_1} = 0.4316$  and

$$\underline{w}_{L_1} = \min_{(a,b) \notin L_1} \tilde{w}(a,b) = 0.68, \quad d(\tilde{w}) = \underline{w}_{L_1} d_{L_1} = 0.29.$$

### 5.3. Conclusion

In a summarizing Table 9 we present descriptive analysis and comparison of critical parameters for regular [10] weight functions from Tables 1–7 (T1–T7) and for the irregular weight function from Table 8 (T8\*). Here,  $L_4$  is defined in (41),  $L_2$  in (42), and  $L_1 \triangleq \{TA\}$  in (18). The number  $d(\tilde{w})$  is called the *critical distance fraction* for the random coding bound  $\underline{R}^{(\tilde{w})}(d)$  of DNA codes for the *nonadditive stem distance*. The parameter  $T_{\tilde{w}}$  in the last row of the table was introduced in [10]; it is an important numerical combinatorial characteristic of the weight function  $\tilde{w}$ , referred to as the critical distance fraction for the theoretical rate of DNA codes based of the *additive* stem  $\tilde{w}$ -distance. The values of  $T_{\tilde{w}}$  given in Table 9 are obtained in [15] with the use of calculation formulas established in [10].

These results justify and provide a mathematical ground for the main conclusion of [9] on satisfactory concordance between samples of weight functions presented in Tables 2–7 and obtained in six different laboratories. Also, it is worth mentioning a very good concordance between weight samples of Tables 3 and 5–7 and unified parameters from Table 1 proposed in [9]. At least, this is seen from the fact that in five tables, namely T1, T3, and T5–T7, the range of estimates of the critical distance is by order less than the range of the tabular weights themselves.

## 6. PROOF OF THEOREM 3

### 6.1. General Scheme of the Proof

Let  $S^{(1)}(\mathbf{x}, \mathbf{y})$  be the nonadditive stem **1**-similarity (2) (i.e., the nonadditive stem  $w$ -similarity for the constant weight function). For an arbitrary  $s \in [n-1]$ , define two sets,

$$\mathcal{P}_L(n, s) \triangleq \{(\mathbf{x}, \mathbf{y}) \in [n, L] \times [n, L] : S^{(1)}(\mathbf{x}, \mathbf{y}) = s\}$$

and

$$\bar{\mathcal{P}}_L(n, s) \triangleq \{\mathbf{x} \in [n, L] : S^{(1)}(\mathbf{x}, \tilde{\mathbf{x}}) = s\}.$$

Let  $\mathbf{u}, \mathbf{v} \in \text{DNA}(n, L)$  be random sequences of length  $n$  equiprobably chosen from the set  $\text{DNA}(n, L)$ . Then probability distributions of the random variables  $S^{(1)}(\mathbf{u}, \mathbf{v})$  and  $S^{(1)}(\mathbf{u}, \tilde{\mathbf{u}})$  for  $0 \leq s \leq n-1$  are, respectively,

$$\begin{aligned} \Pr\{S^{(1)}(\mathbf{u}, \mathbf{v}) = s\} &= \frac{|\mathcal{P}_L(n, s)|}{(\lambda_L(n))^2}, \\ \Pr\{S^{(1)}(\mathbf{u}, \tilde{\mathbf{u}}) = s\} &= \frac{|\bar{\mathcal{P}}_L(n, s)|}{\lambda_L(n)}. \end{aligned}$$

Applying the random coding method for DNA codes developed in [12], we obtain the following lower bound on the cardinality  $N_L(n, D)$ :

For any  $D$ ,  $1 \leq D \leq n - 1$ , we have

$$N_L(n, D) \geq \left\lfloor \frac{1/2 - P_1^{(1)}(n, D; L)}{2P_2^{(1)}(n, D; L)} \right\rfloor - 1, \quad (43)$$

where

$$P_1^{(1)}(n, D; L) \triangleq \Pr\{S^{(1)}(\mathbf{u}, \tilde{\mathbf{u}}) \geq n - D\} = \lambda_L(n)^{-1} \sum_{t=1}^D |\bar{\mathcal{P}}_L(n, n - t)|,$$

$$P_2^{(1)}(n, D; L) \triangleq \Pr\{S^{(1)}(\mathbf{u}, \mathbf{v}) \geq n - D\} = \lambda_L(n)^{-2} \sum_{t=1}^D |\mathcal{P}_L(n, n - t)|.$$

In Section 6.2 we prove the following result.

**Lemma 3.** For any  $s \in [n - 1]$  we have

$$|\mathcal{P}_L(n, s)| \leq \sum_{j=1}^{\min\{s, n-s\}} C^{3j+2} r^{2n-s-j} \binom{s-1}{j-1} \binom{n-s}{j}^2 \times (1 + \omega\alpha^2)^{j-1} (1 + \omega\alpha^{s-j+2}) (1 + \omega\alpha)^{2j} (1 + \omega\alpha^{n-s-2j})^2, \quad (44)$$

where  $r = r(L)$ ,  $C = C(L)$ ,  $\alpha = \alpha(L)$ , and  $\omega = \omega(L)$  are defined in (29).

In Section 6.3 we prove the following result.

**Lemma 4.** For any  $s \in [n - 1]$  we have

$$|\bar{\mathcal{P}}_L(n, s)| \leq \sum_{j=1}^{\min\{s, n-s\}} C^{\lceil j/2 \rceil + j + 1} r^{\frac{2n-s-j+1}{2}} \binom{\lceil (s+2)/2 \rceil}{\lceil (j-2)/2 \rceil} \binom{n-s}{j} \times (1 + \omega\alpha^2)^{(j-1)/2} (1 + \omega\alpha^{(s-j+2)/2}) (1 + \omega\alpha)^j (1 + \omega\alpha^{n-s-2j}), \quad (45)$$

where  $r = r(L)$ ,  $C = C(L)$ ,  $\alpha = \alpha(L)$ , and  $\omega = \omega(L)$  are defined in (29).

One can easily check the following estimate for any  $j \in [s]$ :

$$\frac{\binom{\lfloor s/2 \rfloor}{\lfloor j/2 \rfloor}^2}{\binom{s}{j}} \leq s. \quad (46)$$

For a fixed parameter  $u$ ,  $0 \leq u \leq 1$ , define the functions

$$\mathfrak{p}_L(u) \triangleq \overline{\lim}_{n \rightarrow \infty} \frac{\log_4 |\mathcal{P}_L(n, \lceil (1-u)n \rceil)|}{n},$$

$$\bar{\mathfrak{p}}_L(u) \triangleq \overline{\lim}_{n \rightarrow \infty} \frac{\log_4 |\bar{\mathcal{P}}_L(n, \lceil (1-u)n \rceil)|}{n}. \quad (47)$$

Note that according to (28) we have

$$\lim_{n \rightarrow \infty} \frac{\log_4 \lambda_L(n)}{n} = \log_4 r = \rho_L,$$

where  $r = r(L)$  and  $\rho_L$  are defined in (29) and (36). Hence (see [12]); formula (43) leads to the following *random coding bound* for the rate (19) of  $(n, dn)^{(1)}$  DNA codes:

Let  $d$ ,  $0 < d < 1$ , be a fixed number. If  $\min_{0 \leq u \leq d} \{\rho_L - \bar{\rho}_L(u)\} > 0$ , then

$$R_L(d) \geq \min_{0 \leq u \leq d} \{2\rho_L - \mathbf{p}_L(u)\}, \quad (48)$$

where  $\rho_L$  is defined in (36), and  $\mathbf{p}_L(u)$  in (47).

For  $s \in [n-1]$ , define the numbers

$$\Psi_L(n, s) \triangleq \max_{1 \leq j \leq \{s; n-s\}} \left[ r^{-j} C^{3j} \binom{s}{j} \binom{n-s}{j}^2 (1 + \omega\alpha^2)^j (1 + \omega\alpha)^{2j} \right],$$

where  $r = r(L)$ ,  $C = C(L)$ ,  $\alpha = \alpha(L)$ , and  $\omega = \omega(L)$  are defined in (29), and let

$$\tau_L(u) \triangleq \lim_{n \rightarrow \infty} \frac{\log_4 \Psi_L(n, \lceil (1-u)n \rceil)}{n}, \quad 0 < u < 1.$$

Then, taking into account (46), (44) and (45), we have

$$\begin{aligned} \mathbf{p}_L(u) &\leq (1+u)\rho_L + \tau_L(u), \\ \bar{\mathbf{p}}_L(u) &\leq \frac{1}{2} [(1+u)\rho_L + \tau_L(u)], \end{aligned} \quad (49)$$

where  $\rho_L$  is defined in (36).

Combining (48) and (49), we obtain the following result:

For any distance fraction  $d > 0$ , the rate (19) satisfies

$$R_L(d) \geq \underline{R}_L^*(d) \triangleq \min_{0 \leq u \leq d} \{(1-u)\rho_L - E_L(u)\},$$

where  $E_L(u)$ ,  $u \geq 0$ , is defined in (38).

We complete our arguments with the following lemma, proved in Section 6.4.

**Lemma 5.** For any distance fraction  $0 < d < d_L$  we have

$$\underline{R}_L^*(d) = \underline{R}_L(d) > 0, \quad (50)$$

where  $\underline{R}_L(d)$  is defined in (37) and  $d_L$ ,  $0 < d_L < 1/2$ , is a unique root of the equation  $(1-d)\rho_L = E_L(d)$  in the interval  $(0; 1/2)$ .

## 6.2. Proof of Lemma 3

For arbitrary integers  $M \geq 1$  and  $B \geq 1$ , denote by  $W_0(M; B)$  the total number of ways to distribute  $M$  identical balls into  $B$  boxes provided that boxes are allowed to be empty. Let  $W_2(M; B)$ ,  $M \geq 2B$ , denote the total number of ways to put  $M$  identical balls into  $B$  boxes so that each of the  $B$  boxes contains at least two balls. It is well known that

$$\begin{aligned} W_0(M; B) &= \binom{M+B-1}{B-1}, \\ W_2(M; B) &= \binom{M-B-1}{B-1}. \end{aligned} \quad (51)$$

Consider a pair  $(\mathbf{x}, \mathbf{y}) \in \mathcal{P}_L(n, s)$ . Then there exist  $\mathbf{z} \in \mathcal{Z}(\mathbf{x}, \mathbf{y})$  and an integer  $j = k(\mathbf{x}, \mathbf{y}, \mathbf{z})$  such that the following estimates hold:

$$|\mathbf{z}| - j = s, |\mathbf{z}| \geq 2j \Rightarrow j \leq s; \quad |\mathbf{z}| - j = s, |\mathbf{z}| \leq n \Rightarrow j \leq n - s.$$

Hence we obtain

$$1 \leq j \leq \min \{s; n - s\}. \quad (52)$$

As is seen from (51), the number of ways to put  $|\mathbf{z}|$  identical balls into  $j$  boxes so that each of the  $j$  boxes contains at least two balls equals

$$W_2(|\mathbf{z}|; j) = \binom{s-1}{j-1}.$$

Furthermore, the number of ways to arrange  $n - |\mathbf{z}|$  balls into  $j + 1$  boxes provided that boxes may be empty equals

$$W_0(n - |\mathbf{z}|; j + 1) = W_0(n - s - j; j + 1) = \binom{n-s}{j}.$$

Let  $1 \leq j \leq b \leq n$  be some fixed integers, and let

$$\{b_\ell\} \triangleq (b_1, b_2, \dots, b_\ell, \dots, b_j), \quad b_\ell \geq 1,$$

be an ordered tuple of  $j$  integers. For  $m = 1, 2$ , introduce two sets

$$\Omega_j^m(b) \triangleq \left\{ \{b_\ell\} : \sum_{\ell=1}^j b_\ell = b, b_\ell \geq m \right\}, \quad m = 1, 2, \quad (53)$$

and define

$$\tilde{\lambda}_L^m(j, b) \triangleq \max_{\{b_\ell\} \in \Omega_j^m(b)} \left\{ \prod_{\ell=1}^j \lambda_L(b_\ell) \right\}. \quad (54)$$

Applying the above definitions and formulas, one can easily obtain the following bound on the cardinality of  $\mathcal{P}_L(n, s)$  for any  $s \in [n - 1]$ :

$$|\mathcal{P}_L(n, s)| \leq \sum_{j=1}^{\min\{s; n-s\}} \tilde{\lambda}_L^2(j, s+j) \binom{s-1}{j-1} \left[ \max_{0 \leq j' \leq j} \left\{ \tilde{\lambda}_L^1(j' + 1, n - s - j) \right\} \binom{n-s}{j} \right]^2. \quad (55)$$

Further analysis of the bound (55) is based on upper bounding the number  $\tilde{\lambda}_L^m(j, b)$ ,  $m = 1, 2$ . To obtain such a bound, we show that for any integers  $x \geq y > t > 0$  we have

$$(1 + \omega \alpha^x)(1 + \omega \alpha^y) \leq (1 + \omega \alpha^{x+t})(1 + \omega \alpha^{y-t}), \quad (56)$$

where  $\alpha = \alpha(L)$  and  $\omega = \omega(L)$  are defined in (29). Indeed, transform the left-hand side of (56):

$$(1 + \omega \alpha^x)(1 + \omega \alpha^y) = 1 + \omega^2 \alpha^{x+y} + \omega(\alpha^x + \alpha^y),$$

whereas the right-hand side of (56) equals

$$1 + \omega^2 \alpha^{x+y} + \omega(\alpha^{x+t} + \alpha^{y-t}).$$

It is easily seen that

$$\alpha^x + \alpha^y - \alpha^{x+t} - \alpha^{y-t} = \alpha^x (1 - \alpha^t) - \alpha^{y-t} (1 - \alpha^t) = (\alpha^x - \alpha^{y-t}) (1 - \alpha^t) < 0,$$

since  $\alpha < 1$ . Thus, (56) is proved.

Definitions (53) and (54), upper bound (28), and inequality (56) imply for  $m = 1, 2$  the bound

$$\begin{aligned} \tilde{\lambda}_L^m(j, b) &\leq \max_{\{b_\ell\} \in \Omega_j^m(b)} \left\{ \prod_{\ell=1}^j [Cr^{b_\ell} (1 + \omega\alpha^{b_\ell})] \right\} \\ &= C^j r^b \max_{\{b_\ell\} \in \Omega_j^m(b)} \left\{ \prod_{\ell=1}^j [1 + \omega\alpha^{b_\ell}] \right\} \\ &= r^b C^j (1 + \omega\alpha^m)^{j-1} (1 + \omega\alpha^{b-m(j-1)}). \end{aligned} \quad (57)$$

Combining these inequalities with (55) and taking into account that  $C = C(L) \geq 1$ , we arrive at (44).  $\triangle$

### 6.3. Proof of Lemma 4

Consider an arbitrary  $\mathbf{x} \in \bar{\mathcal{P}}_L(n, s)$  and its reverse complement  $\tilde{\mathbf{x}}$ . Let  $\mathbf{y}$  be a block subsequence of  $\mathbf{x}$ . Then it is easily seen that  $\mathbf{y}$  is also a block subsequence of  $\tilde{\mathbf{x}}$  if and only if the reverse complement  $\tilde{\mathbf{y}}$  is a block subsequence of  $\mathbf{x}$ . Therefore, the sequence  $\tilde{\mathbf{z}}$ , the reverse complement to the maximal common block subsequence  $\mathbf{z}$  of  $\mathbf{x}$  and  $\tilde{\mathbf{x}}$  (see Definition 2), is also a block subsequence of  $\mathbf{x}$ .

Let  $\mathbf{z}$  consist of  $j = k(\mathbf{x}, \tilde{\mathbf{x}}, \mathbf{z})$   $\mathbf{x}$ -blocks. Denote

$$\mathbf{z} = (\mathbf{z}^1, \mathbf{z}^2, \dots, \mathbf{z}^j) \quad \text{and} \quad \tilde{\mathbf{z}} = (\tilde{\mathbf{z}}^1, \tilde{\mathbf{z}}^2, \dots, \tilde{\mathbf{z}}^j).$$

By the definition of a common block subsequence, for any  $i \in [j]$  the  $\mathbf{x}$ -blocks  $\mathbf{z}^i$  and  $(\tilde{\mathbf{z}})^{j-i+1}$  have the same length and are mutually reverse complementary.

Without loss of generality, let the head of the subsequence  $\mathbf{z}$  be to the left of the head of  $\tilde{\mathbf{z}}$  in  $\mathbf{x}$  (or, possibly, they coincide). Let us find the largest index  $J$  such that the head of the block  $\mathbf{z}^J$  lies in the sequence  $\mathbf{x}$  to the left of the head of the block  $(\tilde{\mathbf{z}})^{j-J+1} = \widetilde{\mathbf{z}^J}$ ; i.e., if

$$\mathbf{z}^J = (x_{i_1}, x_{i_1+1}, \dots, x_{i_1+k}) \quad \text{and} \quad (\tilde{\mathbf{z}})^{j-J+1} = (x_{i_2}, x_{i_2+1}, \dots, x_{i_2+k}),$$

then  $i_1 \leq i_2$ . Then two cases are possible.

Case 1. Let  $i_1 + k \geq i_2$ . This means that the intersection of the blocks  $\mathbf{z}^J$  and  $(\tilde{\mathbf{z}})^{j-J+1} = \widetilde{\mathbf{z}^J}$  is the sequence

$$\mathbf{v}_2 \triangleq (x_{i_2} \dots x_{i_1+k}).$$

Introduce also the sequence

$$\mathbf{v}_1 \triangleq (x_{i_1} \dots x_{i_2+k}).$$

Clearly,  $\mathbf{v}_1$  and  $\mathbf{v}_2$  are self-reverse-complementary. Then consider the following two self-reverse-complementary block subsequences of  $\mathbf{x}$ :

$$\begin{aligned} \mathbf{z}_1 &= (\mathbf{z}^1, \mathbf{z}^2, \dots, \mathbf{z}^{J-1}, \mathbf{v}_1, (\tilde{\mathbf{z}})^{j-J+2}, \dots, (\tilde{\mathbf{z}})^{j-1}, (\tilde{\mathbf{z}})^j) \\ &= (\mathbf{z}^1, \mathbf{z}^2, \dots, \mathbf{z}^{J-1}, \mathbf{v}_1, \widetilde{\mathbf{z}^{J-1}}, \dots, \widetilde{\mathbf{z}^2}, \widetilde{\mathbf{z}^1}) \end{aligned}$$

and

$$\begin{aligned} z_2 &= \left( (\widetilde{z})^1, (\widetilde{z})^2, \dots, (\widetilde{z})^{j-J}, v_2, z^{J+1}, z^{J+2}, \dots, z^{j-1}, z^j \right) \\ &= \left( \widetilde{z^j}, \widetilde{z^{j-1}}, \dots, \widetilde{z^{J+1}}, v_2, z^{J+1}, z^{J+2}, \dots, z^{j-1}, z^j \right). \end{aligned}$$

Case 2. Let  $i_1 + k < i_2$ . Then consider the following two self-reverse-complementary block subsequences of  $\mathbf{x}$ :

$$\begin{aligned} z_1 &= \left( z^1, z^2, \dots, z^J, (\widetilde{z})^{j-J+1}, \dots, (\widetilde{z})^{j-1}, (\widetilde{z})^j \right) \\ &= \left( z^1, z^2, \dots, z^J, \widetilde{z^J}, \dots, \widetilde{z^2}, \widetilde{z^1} \right) \end{aligned}$$

and

$$\begin{aligned} z_2 &= \left( (\widetilde{z})^1, (\widetilde{z})^2, \dots, (\widetilde{z})^{j-J}, z^{J+1}, z^{J+2}, \dots, z^{j-1}, z^j \right) \\ &= \left( \widetilde{z^j}, \widetilde{z^{j-1}}, \dots, \widetilde{z^{J+1}}, z^{J+1}, z^{J+2}, \dots, z^{j-1}, z^j \right). \end{aligned}$$

Note that in both cases  $z_1$  and  $z_2$  are common block subsequences between  $\mathbf{x}$  and  $\widetilde{\mathbf{x}}$ . Denote their lengths by  $|z_i| = \ell_i$  and denote by  $k_i$  the minimal number of blocks in the corresponding subsequences. Then we have the following system:

$$\begin{cases} \ell_1 + \ell_2 = 2(s + j), \\ k_1 + k_2 \leq 2j. \end{cases} \quad (58)$$

Since  $\mathbf{x} \in \bar{\mathcal{P}}_L(n, s)$ , we have

$$\begin{cases} \ell_1 - k_1 \leq s, \\ \ell_2 - k_2 \leq s. \end{cases}$$

Taking into account (58), we obtain

$$\begin{cases} \ell_1 - k_1 = s, \\ \ell_2 - k_2 = s. \end{cases} \quad (59)$$

Thus, we have proved that if the maximal common block subsequence  $\mathbf{z}$  of the sequences  $\mathbf{x}$  and  $\widetilde{\mathbf{x}}$  consists of  $j$  common  $(\mathbf{x}, \widetilde{\mathbf{x}})$ -blocks, i.e.,  $k(\mathbf{x}, \widetilde{\mathbf{x}}, \mathbf{z}) = j$ , then there exists a self-reverse-complementary block subsequence  $z_1$  of  $\mathbf{x}$  of length  $\ell_1$  consisting of  $k_1$   $\mathbf{x}$ -blocks and such that the corresponding  $\mathbf{x}$ -blocks  $((z_1)^1$  and  $(z_1)^{k_1}$ ,  $(z_1)^2$  and  $(z_1)^{k_1-1}$ , etc.) have the same length and are mutually reverse complementary. Since (59) holds, we obtain that  $z_1$  is also a maximal common block subsequence of  $\mathbf{x}$  and  $\widetilde{\mathbf{x}}$ ; i.e., the maximal nonadditive stem similarity as attained at it (2). Therefore, taking into account (52), we have

$$k_1 \leq \min\{s; n - s\}.$$

taking into account this result and using the same definitions and arguments as in the proof of Lemma 3, one can show that

$$\begin{aligned} |\bar{\mathcal{P}}_L(n, s)| &\leq \sum_{k=1}^{\min\{s; n-s\}} W_2(\lceil (s+k)/2 \rceil; \lceil k/2 \rceil) \widetilde{\lambda}_L^2(\lceil k/2 \rceil, \lceil (s+k)/2 \rceil) \\ &\quad \times [W_0(n-s-k; k+1) \widetilde{\lambda}_L^1(k+1, n-s-k)]. \end{aligned}$$

The proof is completed by transforming this inequality taking into account the explicit form of  $W_i(M; B)$ ,  $i = 0, 2$ , and estimate (57).  $\triangle$

## 6.4. Proof of Lemma 5

Denote

$$F_L(u) \triangleq (1-u)\rho_L - E_L(u), \quad 0 < u < 1/2,$$

where the function  $E_L(u)$  is defined in (38). To prove Lemma 5, let us show that for any set  $L$  we have

$$1. \lim_{u \rightarrow 0+} F_L(u) = \rho_L > 0; \quad (60)$$

$$2. \frac{\partial F_L(u)}{\partial u} < 0, \quad \text{for } 0 < u < 1/2. \quad (61)$$

It is well known that

$$\lim_{u \rightarrow 0+} [u \log_4 u] = \lim_{t \rightarrow +\infty} \frac{-\log_4 t}{t} = 0.$$

Therefore,

$$\begin{aligned} \lim_{u \rightarrow 0+} F_L(u) &= \lim_{u \rightarrow 0+} \left[ (1-u)\rho_L - \max_{0 \leq v \leq \min\{u, 1-u\}} \left\{ -\rho'_L v + (1-u)h_4\left(\frac{v}{1-u}\right) + 2uh_4\left(\frac{v}{u}\right) \right\} \right] \\ &= \rho_L + \lim_{u \rightarrow 0+} \max_{0 \leq v \leq u} \left\{ v \log_4\left(\frac{v}{1-u}\right) + (1-u-v) \log_4\left(\frac{1-u-v}{1-u}\right) \right. \\ &\quad \left. + 2v \log_4\left(\frac{v}{u}\right) + 2(u-v) \log_4\left(\frac{u-v}{1-u}\right) \right\} = \rho_L, \end{aligned}$$

which proves (60).

One can easily check that

$$G(v, u) \triangleq \frac{\partial E^L(v, u)}{\partial v} = -\rho'_L + \log_4\left(\frac{1-u-v}{v}\right) + 2 \log_4\left(\frac{u-v}{v}\right).$$

This function has the following obvious properties:

$$\lim_{v \rightarrow 0+} G(v, u) = +\infty, \quad \lim_{v \rightarrow u-} G(v, u) = -\infty, \quad \frac{\partial G(v, u)}{\partial v} < 0.$$

Therefore, to find  $v = v(u)$  that maximizes (38), one has to solve the equation  $G(v(u), u) = 0$ . Hence we obtain

$$G(v(u), u) = -\rho'_L + \log_4\left(\frac{1-u-v(u)}{v(u)}\right) + 2 \log_4\left(\frac{u-v(u)}{v(u)}\right) = 0. \quad (62)$$

Using the implicit function theorem, we compute the derivative

$$\left. \frac{\partial v(u)}{\partial u} \right|_{u_0} = - \frac{\left. \frac{\partial G(v, u)}{\partial u} \right|_{u=u_0}}{\left. \frac{\partial G(v, u)}{\partial v} \right|_{v=v(u_0)}} = - \frac{v(u)(2-3u-v(u))}{3u^2 + uv(u) - 3u + v(u)} \Big|_{u=u_0}. \quad (63)$$

Hence, taking into account that  $v \in [0, u]$  and  $u \in (0; 1/2)$ , we obtain that  $v = v(u)$  is an increasing function. Expressing  $\rho'_L$  from equation (62) and substituting it into (38), we obtain the function  $F_L(u)$  in the form

$$\begin{aligned} F_L(u) &= (1-u)\rho_L - \left[ -v(u) \left( \log_4 \frac{1-u-v(u)}{v(u)} + 2 \log_4 \frac{u-v(u)}{v(u)} \right) \right. \\ &\quad \left. + (1-u)h_4\left(\frac{v(u)}{1-u}\right) + 2uh_4\left(\frac{v(u)}{u}\right) \right] \\ &= (1-u)\rho_L + (1-u) \log_4\left(1 - \frac{v(u)}{1-u}\right) + 2u \log_4\left(1 - \frac{v(u)}{u}\right). \end{aligned} \quad (64)$$

To prove (61), we compute

$$\begin{aligned} \frac{\partial F_L(u)}{\partial u} = & -\rho_L - \log_4 \left( 1 - \frac{v(u)}{1-u} \right) + 2 \log_4 \left( 1 - \frac{v(u)}{u} \right) \\ & + \log_4 e \cdot \frac{-v'(u)(1-u) - v(u)}{1-u-v(u)} + 2 \log_4 e \cdot \frac{-uv'(u) + v(u)}{u-v(u)}. \end{aligned}$$

Since  $v \in [0, u]$  and  $u \in (0; 1/2)$ , we have

$$-\log_4 \left( 1 - \frac{v(u)}{1-u} \right) + 2 \log_4 \left( 1 - \frac{v(u)}{u} \right) < 0.$$

Using this inequality and (63), let us show that  $F_L(u)$  is decreasing:

$$\begin{aligned} \frac{\partial F_L(u)}{\partial u} & < -\rho_L + 2 \log_4 e \frac{-uv'(u) + v(u)}{u-v(u)} \\ & = -\rho_L - 2 \log_4 e \frac{v(u)}{3u^2 + uv(u) - 3u + v(u)} < -\rho_L + \log_4 e < 0. \end{aligned}$$

This is true, since  $\rho_L = \log_4 r > \log_4 3 > \log_4 e$ .

At the point  $u = 1/2$ ,  $F_L(u) < 0$  for each set  $L$  that we consider. Thus, from (60) and (61) we see that  $F_L(u)$  is monotone decreasing from the point  $F_L(0) = \rho_L > 0$  to some point  $0 < d_L \leq 1/2$  where  $F_L(d_L) = 0$  and further to negative values of  $F_L(u)$ . The form of the graph of this function is analogous to the graph of the random coding bound given in Fig. 1. Thus, the point  $d_L$  is a unique root of the equation  $(1-d)\rho_L = E_L(d)$  for  $0 < d < 1/2$ , and

$$\underline{R}_L^*(d) = \min_{0 \leq u \leq d} \{(1-u)\rho_L - E_L(u)\} = (1-d)\rho_L - E_L(d) = \underline{R}_L(d)$$

for  $0 < d < d_L$ .  $\triangle$

## REFERENCES

1. Levenshtein, V.I., Binary Codes Capable of Correcting Deletions, Insertions, and Reversals, *Dokl. Akad. Nauk SSSR*, 1965, vol. 163, no. 4, pp. 845–848 [*Soviet Phys. Dokl.* (Engl. Transl.), 1966, vol. 10, no. 8, pp. 707–710].
2. Levenshtein, V.I., Elements of Coding Theory, in *Diskretnaya matematika i matematicheskie voprosy kibernetiki* (Discrete Mathematics and Mathematical Problems of Cybernetics), Moscow: Nauka, 1974, pp. 207–305.
3. Levenshtein, V.I., Efficient Reconstruction of Sequences from Their Subsequences and Supersequences, *J. Combin. Theory, Ser. A*, 2001, vol. 93, no. 2, pp. 310–332.
4. MacWilliams, F.J. and Sloane, N.J.A., *The Theory of Error-Correcting Codes*, Amsterdam: North-Holland, 1977. Translated under the title *Teoriya kodov, ispravlyayushchikh oshibki*, Moscow: Svyaz', 1979.
5. Tenengolts, G.M., Nonbinary Codes, Correcting Single Deletions or Insertions, *IEEE Trans. Inform. Theory*, 1984, vol. 30, no. 5, pp. 766–769.
6. Dancik, V., Expected Length of Longest Common Subsequence, *PhD Thesis*, Univ. of Warwick, UK, 1994.
7. D'yachkov, A.G., Macula, A.J., Pogozelski, W.K., Renz, T.E., Rykov, V.V., and Torney, D.C., A Weighted Insertion–Deletion Stacked Pair Thermodynamic Metric for DNA Codes, *DNA Computing (Proc. 10th Int. Workshop on DNA Computing, Milan, Italy, June 7–10, 2004)*, Ferretti, C., Mauri, G., and Zandron, C., Eds., Lect. Notes Comp. Sci, vol. 3384, Berlin: Springer, 2005, pp. 90–103.

8. Bishop, M.A., D'yachkov, A.G., Macula, A.J., Renz, T.E., and Rykov, V.V., Free Energy Gap and Statistical Thermodynamic Fidelity of DNA Codes, *J. Comput. Biol.*, 2007, vol. 14, no. 8, pp. 1088–1104.
9. SantaLucia J., Jr., A Unified View of Polymer, Dumbbell, and Oligonucleotide DNA Nearest-Neighbor Thermodynamics, *Proc. Natl. Acad. Sci. USA*, 1998, vol. 95, no. 4, pp. 1460–1465.
10. D'yachkov, A.G. and Voronina, A.N., DNA Codes for Additive Stem Similarity, *Probl. Peredachi Inf.*, 2009, vol. 45, no. 2, pp. 56–77 [*Probl. Inf. Trans.* (Engl. Transl.), 2009, vol. 45, no. 2, pp. 124–144].
11. D'yachkov, A.G., Erdős, P.L., Macula, A.J., Rykov, V.V., Torney, D.C., Tung, C.-S., Vilenkin, P.A., and White, P.S., Exordium for DNA Codes, *J. Comb. Optim.*, 2003, vol. 7, no. 4, pp. 369–379.
12. D'yachkov, A.G., Vilenkin, P.A., Ismagilov, I.K., Sarbaev, R.S., Macula, A., Torney, D., and White, S., On DNA Codes, *Probl. Peredachi Inf.*, 2005, vol. 41, no. 4, pp. 57–77 [*Probl. Inf. Trans.* (Engl. Transl.), 2005, vol. 41, no. 4, pp. 349–367].
13. King, O.D. and Gaborit, P., Linear Constructions for DNA Codes, *Theoret. Comput. Sci.*, 2005, vol. 334, no. 1–3, pp. 99–113.
14. D'yachkov, A.G., Macula, A.J., Renz, T.E., and Rykov, V.V., Random Coding Bounds for DNA Codes Based on Fibonacci Ensembles of DNA Sequences, in *Proc. 2008 IEEE Int. Sympos. on Information Theory, Toronto, Canada, July 6–11, 2008*, pp. 2292–2296.
15. D'yachkov, A.G., Voronina, A.N., Macula, A.J., Renz, T.E., and Rykov, V.V., On Critical Relative Distance of DNA Codes for Additive Stem Similarity, in *Proc. 2010 IEEE Int. Sympos. on Information Theory (ISIT'2010), Austin, Texas, USA, June 13–18, 2010*, P. 1325–1329.
16. Dyachkov, A.G., Voronina, A.N., Volkova, J.A., and Polyanskii, N.A., On Optimal DNA Codes for Additive and Non-Additive Stem Similarity, in *Proc. 7th Int. Workshop on Coding and Cryptography (WCC'2011), Paris, France, April 11–15, 2011*, pp. 313–322.
17. Cameron, P.J., *Combinatorics: Topics, Techniques, Algorithms*, Cambridge, UK: Cambridge Univ. Press, 1994.
18. Reingold, E.M., Nievergelt, J., and Deo, N., *Combinatorial Algorithms: Theory and Practice*, Englewood Cliffs, N.J.: Prentice-Hall, 1977. Translated under the title *Kombinatornye algoritmy: teoriya i praktika*, Moscow: Mir, 1980.
19. Voronina, A.N., *Probability-Theoretic and Combinatorial Problems in DNA Sequence Coding Theory, Cand. Sci. (Phys.-Math.) Dissertation*, Moscow: Moscow State Univ., 2010.