2008

Random Coding Bounds for DNA Codes Based on Fibonacci Ensembles of DNA Sequences

A. D'yachkov

A. Macula

T. Renz

V. Rykov

2008 IEEE International Symposium on Information Theory - Proceedings 4595399, pp. 2292-2296

Random Coding Bounds for DNA Codes Based on Fibonacci Ensembles of DNA Sequences[0]

A. D'yachkov*, A. Macula†, T. Renz† and V. Rykov‡

*Department of Probability Theory, Faculty of Mechanics and Mathematics
Moscow State University, Moscow, 119992, Russia, Email: agd-msu@yandex.ru
†Air Force Res. Lab., IFTC, Rome Research Site, Rome NY 13441, USA,
Email: macula@geneseo.edu, thomas.renz@rl.af.mil
‡Department of Mathematics, University of Nebraska at Omaha,
6001 Dodge St., Omaha, NE 68182-0243, USA, E-mail: vrykov@mail.unomaha.edu

Abstract— **We consider DNA codes based on the concept of a weighted 2-stem similarity measure which reflects the "hybridization potential" of two DNA sequences. A random coding bound on the rate of DNA codes with respect to a thermodynamic motivated similarity measure is proved. Ensembles of DNA strands whose sequence composition is restricted in a manner similar to the restrictions in binary Fibonacci sequences are introduced to obtain the bound.**

I. Introduction

Single strands of DNA are represented by $\{A, C, G, T\}$ – sequences that are oriented. The *reverse-complement* (Watson-Crick transformation) of a DNA strand is defined by first reversing the order of the letters and then substituting each letter x for its complement $\bar{x}$, namely: A for T, C for G and vice-versa. For example, the reverse complement of *AACG* is *CGTT*. For strand $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \{A, C, G, T\}^n$, let

$$\tilde{\tilde{\mathbf{x}}} = (\bar{x}_n, \bar{x}_{n-1}, \ldots, \bar{x}_2, \bar{x}_1) \in \{A, C, G, T\}^n \quad (1)$$

denote its reverse complement. If $\mathbf{y} = \tilde{\tilde{\mathbf{x}}}$, then $\mathbf{x} = \tilde{\tilde{\mathbf{y}}}$ for any $\mathbf{x} \in \{A, C, G, T\}^n$. If $\mathbf{x} = \tilde{\tilde{\mathbf{x}}}$, then $\mathbf{x}$ is called a *self reverse complementary* sequence. If $\mathbf{x} \neq \tilde{\tilde{\mathbf{x}}}$, then a pair $(\mathbf{x}\,, \tilde{\tilde{\mathbf{x}}})$ is called a *pair of mutually reverse complementary* sequences. A (perfect) *Watson-Crick duplex* is the joining of $\mathbf{x}$ and $\tilde{\tilde{\mathbf{x}}}$ so that every letter of one strand is paired with its complementary letter on the other strand in the double helix structure, i.e., $\mathbf{x}$ and $\tilde{\tilde{\mathbf{x}}}$ are "perfectly compatible." However, when two, not necessarily complementary, oppositely directed DNA strands are "sufficiently compatible," they too are capable of coalescing into a double stranded DNA duplex. The process of forming DNA duplexes from single strands is referred to as *DNA hybridization. Crosshybridization* occurs when two oppositely directed and non-complementary DNA strands form a duplex. Crosshybridization doesn't always occur, but there is a potential for it to happen. In general, crosshybridization is undesirable as it usually leads to experimental error. To increase the accuracy and throughput of the applications listed in [1]-[7], there is a desire to have collections of DNA strands, as large and as mutually incompatible as possible, so that no crosshybridization can take place. It is straightforward to view this problem as one in coding theory.

DNA nanotechnology often requires collections of DNA strands called *free energy gap codes* [8] that will correctly "self-assemble" into Watson-Crick duplexes and do not produce erroneous crosshybridizations. When these collections consist entirely of pairs of mutually reverse complementary DNA strands they are called *DNA tag-antitag systems* [1] and DNA codes [9], [10].

Statistical thermodynamics is applied [5]-[7] to model competitive multiplexing hybridization. In paper [8], a weighted 2-stem similarity function (see, below Definition 4) is introduced which provides a more accurate estimation of the hybridization energy than other similarity functions current in use, e.g., Hamming, insertion-deletion or edit [2]-[4]. The model in [8] argues that the probability that a DNA code correctly assembles (called the fidelity of DNA codes) is a function of the corresponding distance measure (see, below Definition 5).

In the given paper, the techniques of [9], [10] are extended to obtain a random coding bound on the rate of DNA codes defined in [8]. For applications [5], the bound shows that, asymptotically, dramatically improved DNA codes exist and yields an asymptotic behavior for the fidelity of DNA codes.

II. Statement of Problem

A. Notations and Auxiliary Definitions

The symbol $\triangleq$ denotes definitional equalities and the symbol $[n] \triangleq \{1, 2, \ldots, n\}$ denotes the set of integers from 1 to n. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, where $\mathbf{x}, \mathbf{y} \in \{A, C, G, T\}^n$, be two arbitrary DNA n-sequences. By symbol $\mathbf{z} = (z_1, z_2, \ldots, z_\ell) \in \{A, C, G, T\}^\ell$, $\ell \in [n]$, we will denote a *common subsequence* [11] of length $|\mathbf{z}| \triangleq \ell$ between $\mathbf{x}$ and $\mathbf{y}$. The *empty* subsequence $\mathbf{z}$ of length $|\mathbf{z}| \triangleq 0$ is a common subsequence between any sequences $\mathbf{x}$ and $\mathbf{y}$.

Definition 1. Let $2 \le r \le n$ be an arbitrary integer. A fixed DNA r-sequence $\mathbf{a} = (a_1, a_2, \ldots, a_r) \in \{A, C, G, T\}^r$, is called a *common block for sequences* $\mathbf{x}$ *and* $\mathbf{y}$ (briefly, *common* $(\mathbf{x}, \mathbf{y})$*-block*) of length r if sequences $\mathbf{x}$ and $\mathbf{y}$ (simultaneously) contain $\mathbf{a}$ as a subsequence consisting of r consecutive elements of $\mathbf{x}$ and $\mathbf{y}$. We will say that a common $(\mathbf{x}, \mathbf{y})$-block $\mathbf{a}$ *yields* $r - 1$ *common* 2*-stems* a_i, a_{i+1}, $i \in [r-1]$, containing 2 adjacent symbols of the given common $(\mathbf{x}, \mathbf{y})$-block.

[0]The work was supported by AFOSR – FA8750-07-C-0089

Definition 2. Let $2 \le \ell \le n$ be an integer. A sequence $\mathbf{z} = (z_1, z_2, \ldots, z_\ell) \in \{A, C, G, T\}^\ell$ is called a *common block subsequence* of length $|\mathbf{z}| \triangleq \ell$ between $\mathbf{x}$ and $\mathbf{y}$ if $\mathbf{z}$ is an *ordered collection* of non-overlapping (separated) common $(\mathbf{x}, \mathbf{y})$-blocks and the length of each common $(\mathbf{x}, \mathbf{y})$-block in this collection is ≥ 2. Let $\mathcal{Z}(\mathbf{x}, \mathbf{y})$ be the set of all common block subsequences between $\mathbf{x}$ and $\mathbf{y}$. For any $\mathbf{z} \in \mathcal{Z}(\mathbf{x}, \mathbf{y})$, we denote by $k(\mathbf{z}, \mathbf{x}, \mathbf{y})$, $1 \le k(\mathbf{z}, \mathbf{x}, \mathbf{y}) \le |\mathbf{z}|/2$, the *minimal number* of common $(\mathbf{x}, \mathbf{y})$–blocks which *constitute* the given subsequence $\mathbf{z}$.

Note that the difference $|\mathbf{z}| - k(\mathbf{z}, \mathbf{x}, \mathbf{y})$, $\mathbf{z} \in \mathcal{Z}(\mathbf{x}, \mathbf{y})$, is a total number of common 2-stems containing adjacent symbols in common $(\mathbf{x}, \mathbf{y})$-blocks constituting $\mathbf{z} \in \mathcal{Z}(\mathbf{x}, \mathbf{y})$.

Definition 3. [8] For sequences $\mathbf{x}, \mathbf{y} \in \{A, C, G, T\}^n$, the number

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

is called an 2*-stem similarity between* $\mathbf{x}$ *and* $\mathbf{y}$. Obviously, $S(\mathbf{x}, \mathbf{y}) = S(\mathbf{y}, \mathbf{x}) \le S(\mathbf{x}, \mathbf{x}) = n - 1$.

Example. Let $n = 10$ and

$$\mathbf{x} = (A, T, \underbrace{T, A, A}, A, A, \underbrace{T, T, A}),$$

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

A common block subsequence $\mathbf{z}$ between $\mathbf{x}$ and $\mathbf{y} = \tilde{\tilde{\mathbf{x}}}$ is

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

The value $k(\mathbf{z}, \mathbf{x}, \mathbf{y}) = 2$ and the corresponding 2–stem similarity is

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

The maximal value is achieved for the above self reverse complementary sequence $\mathbf{z} \in \mathcal{Z}(\mathbf{x}, \mathbf{y})$.

B. Weighted Stem Similarity and Distance

Let $w = w(a, b) \ge 0$, $a, b \in \{A, C, G, T\}$, be a weight function such that

$$w(a, b) = w(\bar{b}, \bar{a}), \quad a, b \in \{A, C, G, T\}. \quad (3)$$

Condition (3) means that $w(a, b)$ is an invariant function under Watson-Crick transformation.

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

$$\mathbf{z} \triangleq \left(\mathbf{z}^1, \mathbf{z}^2, \ldots, \mathbf{z}^{k(\mathbf{z}, \mathbf{x}, \mathbf{y})}\right),$$

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

where

$$\mathbf{z}^m \triangleq \left(z_1^m, z_2^m, \ldots, z_{r_m}^m\right) \in \{A, C, G, T\}^{r_m},$$
$$m = 1, 2, \ldots, k(\mathbf{z}, \mathbf{x}, \mathbf{y}),$$

is an ordered collection of common $(\mathbf{x}, \mathbf{y})$-blocks constituting $\mathbf{z}$ and $r_m \triangleq |\mathbf{z}^m| \ge 2$ is the length of block $\mathbf{z}^m$. For DNA sequences $\mathbf{x}, \mathbf{y} \in \{A, C, G, T\}^n$, the number

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

is called a *weighted* 2*-stem similarity between* $\mathbf{x}$ *and* $\mathbf{y}$. We will say that $\mathcal{S}^{(w)}(\mathbf{x}, \mathbf{y}) \triangleq 0$ iff the set $\mathcal{Z}(\mathbf{x}, \mathbf{y}) = \varnothing$.

Function $\mathcal{S}^{(w)}(\mathbf{x}, \tilde{\tilde{\mathbf{y}}})$ is used to model [8]-[10] a *thermodynamic similarity* (*hybridization energy*) between DNA sequences $\mathbf{x}$ and $\mathbf{y}$.

Proposition 1. *For any* $\mathbf{x}, \mathbf{y} \in \{A, C, G, T\}^n$*, the function*

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

In addition,

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

The symmetry property and inequality (5) are evident. Equality (6) follows from definitions (1),(4) and condition (3). Identity (6) means the symmetry property of hybridization energy between DNA sequences $\mathbf{x}$ and $\mathbf{y}$ [8]-[10].

One can easily check that 2-stem similarity $S(\mathbf{x}, \mathbf{y})$ from Definition 3 corresponds to the uniform weight function: $w(a, b) \equiv 1$ for any $a, b \in \{A, C, G, T\}$. Table 1 shows an example of values for $w(a, b)$ which satisfy (3) and have a significant biological motivation:

$w(a,b)$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	1.02	1.46	1.29	0.88
$a = C$	1.46	1.83	2.17	1.29
$a = G$	1.32	2.24	1.83	1.46
$a = T$	0.60	1.32	1.46	1.02

Table 1.

These values come from [5] and are the nearest neighbor ”thermodynamic weights” (e.g., free energy of formation) associated to stacked pairs that occurred in DNA secondary structures. See [7] for an introduction to the nearest neighbor model.

Definition 5. [8] The number

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

is called a *weighted* 2*-stem distance between* $\mathbf{x}$ *and* $\mathbf{y}$.

Typically, $\mathcal{D}^{(w)}(\mathbf{x}, \mathbf{y}) \ne \mathcal{D}^{(w)}(\mathbf{y}, \mathbf{x})$, i.e., function (7) is not symmetric. Proposition 1 gives:

$$\mathcal{D}^{(w)}(\mathbf{x}, \mathbf{y}) \ge \mathcal{D}^{(w)}(\mathbf{x}, \mathbf{x}) = 0. \quad (8)$$

C. DNA Codes based on Stem Similarity

Let $\mathbf{x}(j) \triangleq (x_1(j), x_2(j), \ldots, x_n(j)) \in \{A, C, G, T\}^n$, $j \in N$, be *codewords* of a *code* $X = \{\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N)\}$ of *length* n and *size* N, where $N = 2, 4, \ldots$ is an even integer. Let D, $0 < D \le \max_{\mathbf{x}} \mathcal{S}^{(w)}(\mathbf{x}, \mathbf{x})$, be an arbitrary positive number. Taking into account (7) and (8), we give

Definition 6. A code X is called a DNA (n, D, w)-*code based on weighted* 2*-stem similarity* $\mathcal{S}^{(w)}(\mathbf{x}, \mathbf{y})$ (briefly,

(n, D, w)-code) if the following two conditions are fulfilled. (i). For any number $j \in [N]$ there exists $j' \in [N]$, $j' \neq j$, such that $\mathbf{x}(j') = \widetilde{\mathbf{x}(j)} \neq \mathbf{x}(j)$. In other words, X is a collection of $N/2$ pairs of mutually reverse complementary sequences. (ii). For any $j, j' \in [N]$, where $j \neq j'$, the distance $\mathcal{D}^{(w)}(\mathbf{x}(j), \mathbf{x}(j')) \geq D$.

The following statement is obvious.

Proposition 2. *Let* (3) *be the uniform weight function, i.e.,*

$$w(a, b) \equiv 1, \qquad a, b \in \{A, C, G, T\}.$$

The corresponding symmetric distance function $\mathcal{D}^{(\equiv 1)}(\mathbf{x}, \mathbf{y})$, $\mathbf{x}, \mathbf{y} \in \{A, C, G, T\}^n$ *has the form*

$$\mathcal{D}^{(\equiv 1)}(\mathbf{x}, \mathbf{y}) = \mathcal{D}^{(\equiv 1)}(\mathbf{y}, \mathbf{x}) = (n-1) - S(\mathbf{x}, \mathbf{y}), \quad (9)$$

where 2*-stem similarity* $S(\mathbf{x}, \mathbf{y})$ *is defined by* (2), *and the definition of DNA* $(n, D, \equiv 1)$*-code,* $0 < D \leq n-1$*, is identified by inequality*

$$S(\mathbf{x}(j), \mathbf{x}(j')) \leq (n-1) - D, \quad j, j' \in [N],\ j \neq j'. \quad (10)$$

One reason for a considering (n, D, w)-code X of size N can be found by noting that the statistical thermodynamic model for DNA code self assembly given in [8] indicates that: given two identical copies of X the probability $\mathcal{F}$ (called the *fidelity* of DNA code X) that only Watson-Crick duplexes form and no crosshybridization duplexes exist is

$$\mathcal{F} \geq \left(\frac{1}{1 + N \exp_4\{-D \cdot \mathcal{K}\}} \right)^N, \quad \mathcal{K} \triangleq \frac{\log_4 e}{\mathcal{R}\mathcal{T}}, \quad (11)$$

where $\mathcal{R}$ is the universal gas constant, $\mathcal{T}$ is the Kelvin temperature and $e = 2.71828$ is the base of natural logarithm.

Definition 7. Let $N^{(w)}(n, D)$ be the *maximal* size of DNA (n, D, w)-codes based on weighted 2-stem similarity. If $d > 0$ is a fixed number, then

$$R^{(w)}(d) \triangleq \overline{\lim_{n \to \infty}} \frac{\log_4 N^{(w)}(n, nd)}{n} \quad (12)$$

is called a *rate* of DNA (n, nd, w)-codes for a *distance fraction* d.

D. DNA Codes for Fibonacci Ensembles

Let L be a collection of 2-strings of DNA letters, *closed under reverse complement transformation*. For instance,

$$L = \varnothing, \quad L = \{TA\}, \quad L = \{TA, AT\}$$

$$L = \{TA, AT, AA, TT\}. \quad (13)$$

Denote by $DNA(n, L)$ (briefly, $[n, L]$) the set (ensemble) of all DNA sequences which *do not contain* 2*-stems from* L. We will say that $[n, L]$ is the *Fibonacci L-ensemble*[1]. Denote by $\lambda_L(n) \triangleq |DNA(n, L)| = |[n, L]|$ the *cardinality* of $[n, L]$.

[1]Binary $0, 1$-sequences which do not contain 2-stems of the form $(1, 1)$ are known as the Fibonacci sequences [12].

Definition 8. Let $N_L(n, D)$ be the *maximal size* of DNA $(n, D, \equiv 1)$-codes $X \subseteq DNA(n, L)$. If the distance fraction $d > 0$ is a fixed number, then

$$R_L(d) \triangleq \overline{\lim_{n \to \infty}} \frac{\log_4 N_L(n, nd)}{n} \quad (14)$$

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

For a weight function (3), introduce numbers

$$\underline{w}_L \triangleq \min_{(a,b) \notin L} w(a, b). \quad (15)$$

For instance, if the values of $w = w(a, b)$ are given by Table 1, then

$$\underline{w}_L = \begin{cases} 0.60 & \text{if } L = \varnothing, \\ 0.88 & \text{if } L = \{TA\}, \\ 1.02 & \text{if } L = \{TA, AT\}, \\ 1.29 & \text{if } L = \{TA, AT, AA, TT\}. \end{cases} \quad (16)$$

One can easily check [8] that the distance

$$\mathcal{D}^{(w)}(\mathbf{x}, \mathbf{y}) \geq \underline{w}_L \cdot \mathcal{D}^{(\equiv 1)}(\mathbf{x}, \mathbf{y}) \quad \text{if} \quad \mathbf{x}, \mathbf{y} \in DNA(n, L).$$

In virtue of (9) and (10), this gives

Proposition 3. *Let* $\underline{w}_L$ *be a number defined by* (15) *and a code* $X \subset DNA(n, L)$*. If* X *is a DNA* $(n, D, \equiv 1)$*-code, then* X *is a DNA* $(n, \underline{w}_L \cdot D, w)$*-code. Hence, rate* (12) *satisfies inequality*

$$R^{(w)}(d) \geq \max_L R_L \left(\frac{d}{\underline{w}_L} \right), \quad (17)$$

where $R_L(d)$ *is defined by* (14).

In the rest part of paper, we obtain a random coding bound on $R_L(d)$ for L defined by (13). Then applying (17), we get a random coding bound on the rate $R^{(w)}(d)$ of DNA (n, nd, w)-codes based on weighted 2-stem similarity.

III. Random Coding Bounds

A. On Cardinalities of Fibonacci L-Ensembles

If $L = \varnothing$, then $\lambda_L(n) = 4^n$. If $L \neq \varnothing$, then cardinalities $\lambda_L(1) = 4$ and $\lambda_L(2) = 16 - |L|$ are given. For sets L defined by (13), we calculate cardinalities $\lambda_L(n)$, $n = 3, 4 \ldots$, using the following well known result [12] from the theory of recurrent sequences.

Proposition 4. *Let* $f_1 \neq 0$ *and* $f_2 \neq 0$ *be arbitrary fixed numbers. If sequence* $\lambda_L(n)$*,* $n = 3, 4, \ldots$*, satisfies recurrent equation*

$$\lambda_L(n) = f_1 \lambda_L(n-1) + f_2 \lambda_L(n-2), \quad (18)$$

then

$$\lambda_L(n) = C_1 r_1^n + C_2 r_2^n, \quad n = 1, 2, \ldots, \quad (19)$$

where $r_1 = r_1(L)$ *and* $r_2 = r_2(L)$ *are roots of the characteristic equation* $r^2 - f_1 r - f_2 = 0$ *and* $C_1 = C_1(L)$, $C_2 = C_2(L)$ *are calculated from initial conditions*: $4 = C_1 r_1 + C_2 r_2$, $16 - |L| = C_1 r_1^2 + C_2 r_2^2$.

Formula (19), obviously, leads to

Proposition 5. *If* r_1, r_2 *are real numbers,* $r_1 > 0$ *and* $r_1 > |r_2|$*, then* $\lambda_L(n)$*,* $n = 1, 2, \ldots$*, satisfies inequalities*

$$C r^n \left[1 - \beta \alpha^n\right] \leq \lambda_L(n) \leq C r^n \left[1 + \beta \alpha^n\right], \quad (20)$$

where

$$r = r_1 \triangleq \max\{r_1, r_2\}, \quad C \triangleq C_1,$$

$$\alpha \triangleq \left|\frac{r_2}{r_1}\right| < 1, \quad \beta \triangleq \left|\frac{C_2}{C_1}\right|. \tag{21}$$

Remark. For the case $L = \varnothing$, bounds (20) will be true as well (with the sign of equality) if we formally define $r_1 = 4$, $C_1 = 1$ and $r_2 = C_2 = 0$, i.e., $C = 1$, $r = 4$ and $\alpha = \beta = 0$.

Lemma 1. *If $L = \{TA\}$, then $\lambda_L(n)$ satisfies* (18), *where* $f_1 = 4,\ f_2 = -1$. *Hence, parameters* (21) *of bounds* (20) *are:*

$$r = 2 + \sqrt{3} = 3.73, \quad C = \frac{3 + 2\sqrt{3}}{6} = 1.08,$$

$$\alpha = \beta = 7 - 4\sqrt{3} = .0718. \tag{22}$$

Lemma 2. *If $L = \{TA, AT\}$, then $\lambda_L(n)$ satisfies* (18), *where* $f_1 = 3,\ f_2 = 2$. *Hence, parameters* (21) *of bounds* (20) *are:*

$$r = \frac{3 + \sqrt{17}}{2} = 3.56, \quad C = \frac{17 + 5\sqrt{17}}{34} = 1.11,$$

$$\alpha = \frac{13 - 3\sqrt{17}}{4} = .158, \quad \beta = \frac{21 - 5\sqrt{17}}{4} = .0961. \tag{23}$$

Lemma 3. *If $L = \{TA, AT, AA, TT\}$, then $\lambda_L(n)$ satisfies* (18), *where* $f_1 = 2,\ f_2 = 4$. *Hence parameters* (21) *of bounds* (20) *are:*

$$r = 1 + \sqrt{5} = 3.24, \quad C = \frac{5 + 3\sqrt{5}}{10} = 1.17,$$

$$\alpha = \frac{3 - \sqrt{5}}{2} = .382, \quad \beta = \frac{7 - 3\sqrt{5}}{2} = .146. \tag{24}$$

Proof of Lemmas 1-3. Let $a, b \in \{A, C, G, T\}$ denote arbitrary letters of DNA alphabet and

$$[n, L]_a \triangleq \{\mathbf{x} : \mathbf{x} \in [n, L] \quad \text{and} \quad x_n = a\},$$

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

denote the corresponding subsets of ensemble $[n, L]$. If a pair $(a, b) \in L$, then subset $[n, L]_{a,b} = \varnothing$. Note that $[n, L]_a$ and $[n, L]$ can be written as sums of non-intersecting subsets:

$$[n, L]_a = [n, L]_{A,a} + [n, L]_{C,a} + [n, L]_{G,a} + [n, L]_{T,a}$$

$$[n, L] = [n, L]_A + [n, L]_C + [n, L]_G + [n, L]_T. \tag{25}$$

In addition, one can easily see the following two properties.
1) If for any $b \in \{A, C, G, T\}$, pair $(b, a) \notin L$, then the cardinality

$$|[n, L]_a| = |[n-1, L]| = \lambda_L(n-1). \tag{26}$$

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

$$|[n, L]_{a,b}| = |[n-1, L]_a|. \tag{27}$$

Applying (25)-(27), one can check all recurrent equations formulated in Lemmas 1–3.

B. Random Coding Bound for Fibonacci L-Ensemble

Let

$$\rho_L \triangleq \log_4 r, \quad \rho_L' \triangleq \log_4 \frac{r}{C^3(1 + \beta\alpha^2)(1 + \beta\alpha)^2}.$$

where $r = r(L)$, $C = C(L)$, $\alpha = \alpha(L)$ and $\beta = \beta(L)$ are introduced in Propositions 4 and 5 and given by formulas (21). For sets L defined by (13), parameters (21) are calculated by formulas (22)-(24). In Sect. IV, using a random coding method [10], we present a brief proof of

Theorem 1. *For any distance fraction $d > 0$, the rate* (14) *satisfies inequality*

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

where

$$E_L(u) \triangleq \max_{0 \leq v \leq \min\{u, 1-u\}} E^L(v, u),$$

$$E^L(v, u) \triangleq -\rho_L' \cdot v + (1-u)\, h_4\left(\frac{v}{1-u}\right) + 2\, u\, h_4\left(\frac{v}{u}\right),$$

$$h(u) \triangleq -u \log_4 u - (1-u) \log_4 (1-u).$$

Let a number d_L, $0 < d_L < 1$, be the unique root of equation $\underline{R}_L(d) = 0$ or $(1-d)\rho_L = E_L(d)$. Obviously, if $0 < d < d_L$, then $R_L(d) > 0$ and the following lower bound

$$R_L(d) \geq \underline{R}_L(d) \triangleq (1-d)\rho_L - E_L(d), \qquad 0 < d < d_L,$$

holds. Function $\underline{R}_L(d)$ is called a *random coding bound* on the rate $R^L(d)$. We will say that the number d_L, $0 < d_L < 1$, is a *critical distance fraction* of the random coding bound $\underline{R}_L(d)$ for $DNA(n, L)$-ensemble.

For sets (13), our calculations based on Lemmas 1-3 give the following numerical values for critical distance fractions:

$$d_L = \begin{cases} 0.4794 & \text{if } L = \varnothing, \\ 0.4316 & \text{if } L = \{TA\}, \\ 0.4054 & \text{if } L = \{TA, AT\}, \\ 0.3487 & \text{if } L = \{TA, AT, AA, TT\}. \end{cases} \tag{28}$$

C. Random Coding Bound for DNA (n, dn, w)-Codes

Let $R^{(w)}(d)$, $d > 0$, be the rate (12) of DNA (n, dn, w)-codes and d_L, $0 < d_L < 1$, is the critical distance fraction of random coding bound $\underline{R}_L(d)$ for Fibonacci L-ensemble. Proposition 3 and Theorem 1 lead to

Theorem 2. *If* $0 < d < d^{(w)} \triangleq \max_L \{\underline{w}_L \cdot d_L\}$, *then the rate* $R^{(w)}(d) > 0$ *and lower bound*

$$R^{(w)}(d) \geq \underline{R}^{(w)}(d) \triangleq \max_L \left\{\underline{R}_L\left(\frac{d}{\underline{w}_L}\right)\right\}$$

holds.

Function $\underline{R}^{(w)}(d)$ is called a random coding bound for DNA (n, dn, w)-codes. The number $d^{(w)} > 0$ is called a *critical distance fraction* of the random coding bound $\underline{R}^{(w)}(d)$.

Obviously, inequality (11) and Theorem 2 yield the asymptotic ($n \to \infty$) existence of DNA (n, dn, w)-codes of size $N \geq \exp_4\{n\underline{R}^{(w)}(d)\}$ and fidelity

$$\mathcal{F} \geq 1 - \exp_4\left\{-n\left[\mathcal{K}d - 2\underline{R}^{(w)}(d)\right]\right\}, \quad \underline{d}^{(w)} \leq d \leq d^{(w)},$$

where number $\underline{d}^{(w)}$, $0 < \underline{d}^{(w)} < d^{(w)}$, is the unique root of equation $\mathcal{K}\, d = 2\,\underline{R}^{(w)}(d)$. Note that $\mathcal{K}\, d - 2\,\underline{R}^{(w)}(d) > 0$ if $\underline{d}^{(w)} < d < d^{(w)}$.

IV. Proof of Theorem 1

Let $S(\mathbf{x}, \mathbf{y})$ be 2-stem similarity (2) for the uniform weight function. For an arbitrary integer $s \in [n-1]$, define the set $\mathcal{P}_L(n,s) \triangleq \{(\mathbf{x},\mathbf{y}) \in [n,L] \times [n,L] \,:\, S(\mathbf{x},\mathbf{y}) = s\}$.

Lemma 4. *The size*

$$|\mathcal{P}_L(n,s)| \le \sum_{j=1}^{\min\{s,n-s\}} r^{s+j} \binom{s-1}{j-1} \left[C\,(1 + \beta\,\alpha^2)\right]^j \times$$

$$\times \left\{ r^{n-s-j}\, \left[C\,(1+\beta\,\alpha)\right]^{j+1} \binom{n-s}{j} \right\}^2, \qquad (29)$$

where $r = r(L)$, $C = C(L)$, $\alpha = \alpha(L)$ and $\beta = \beta(L)$ were introduced in the formulation of Theorem 1.

The random coding method [10], Lemma 4 and an asymptotic analysis of the right-hand side (29) yield Theorem 1. To complete the proof of Theorem 1, we give

Proof of Lemma 4. Consider a pair $(\mathbf{x},\mathbf{y}) \in A^n \times A^n$ for which $S(\mathbf{x},\mathbf{y}) = s$. Then there exists $\mathbf{z} \in \mathcal{Z}(\mathbf{x},\mathbf{y})$, $|\mathbf{z}| \le n$, and the integer $j = k(\mathbf{z},\mathbf{x},\mathbf{y}) \le |\mathbf{z}|/2$ for which equalities

$$s = |\mathbf{z}| - j \iff |\mathbf{z}| = s + j \iff n - |\mathbf{z}| = n - s - j$$

take place. It follows that for any $\mathbf{z} \in \mathcal{Z}(\mathbf{x},\mathbf{y})$, the number $j = k(\mathbf{z},\mathbf{x},\mathbf{y})$ satisfies inequalities $1 \le j \le \min\{s\,;\, n-s\}$.

Obviously, the number of all ways to distribute $|\mathbf{z}|$ indistinguishable marbles in j boxes provided that each of j boxes contains ≥ 2 marbles is $\binom{s-1}{j-1}$. In addition, the number of all ways to distribute $n - |\mathbf{z}|$ indistinguishable marbles in $j+1$ boxes if empty boxes are accepted is $\binom{n-s}{j}$.

Let $1 \le j \le b \le n$ be fixed integers and

$$\{b_\ell\} \triangleq (b_1, b_2, \ldots, b_\ell, \ldots, b_j), \quad b_\ell \ge 1,$$

is an ordered collection of integers. For $m = 1, 2$, introduce two sets

$$(\{b_\ell\})_m \triangleq \left\{ \{b_\ell\} \,:\, \sum_{\ell=1}^{j} b_\ell = b, \quad b_\ell \ge m \right\} \qquad (30)$$

and define numbers

$$\widetilde{\lambda}_L^m(j,b) \triangleq \max_{(\{b_\ell\})_m} \left\{ \prod_{\ell=1}^{j} \lambda_L(b_\ell) \right\}. \qquad (31)$$

Applying above formulas and notations, one can see that for any $s \in [n-1]$, the cardinality

$$|\mathcal{P}_L(n,s)| \le \sum_{j=1}^{\min\{s\,;\,n-s\}} \widetilde{\lambda}_L^2\,(j\,,\, s+j) \cdot \binom{s-1}{j-1} \times$$

$$\times \left[\widetilde{\lambda}_L^1\,(j+1\,,\, n-s-j) \binom{n-s}{j} \right]^2. \qquad (32)$$

From definition (30)-(31) and upper bound (20) it follows that for $m = 1, 2$,

$$\widetilde{\lambda}_L^m(j,b) \le \max_{(\{b_\ell\})_m} \left\{ \prod_{\ell=1}^{j} \left[C\, r^{b_\ell}\, \left(1 + \beta\,\alpha^{b_\ell}\right) \right] \right\} =$$

$$= C^j\, r^b \max_{(\{b_\ell\})_m} \left\{ \prod_{\ell=1}^{j} \left[1 + \beta\alpha^{b_\ell}\right] \right\} \le r^b\, \left[C(1+\beta\alpha^m)\right]^j.$$

These inequalities and (32) lead to (29).

Lemma 4 is proved.

V. Conclusion

Let weight function $w = w(a,b)$ be defined [5] by Table 1. Then for sets (13), numbers (16) and (28) give:

$$\underline{w}_L \cdot d_L = \begin{cases} 0.29 & \text{if } L = \varnothing, \\ 0.38 & \text{if } L = \{TA\}, \\ 0.41 & \text{if } L = \{TA, AT\}, \\ 0.45 & \text{if } L = \{TA, AT, AA, TT\}. \end{cases}$$

Therefore, the corresponding critical distance fraction is $d^{(w)} \triangleq \max\limits_{L}\{\underline{w}_L \cdot d_L\} = 0.45$. In other words, for the given weight function, we have shown that by restricting the allowed sequence composition in DNA strands to avoid occurrences of $L = \{TA, AT, AA, TT\}$ the thermodynamically weighted critical distance fraction of the random coding bound for DNA (n, dn, w)-codes can be improved from 0.29 to 0.45.

References

[1] *Kaderali L.*, *Deshpande A.*, *Nolan J.*, *White P.*, Primer-design for multiplexed genotyping // Nucleic Acids Res., 2003, V. 31, P. 1796–1802.

[2] *Li X.*, *He Z*, *Zhou J.*, Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation // Nucleic Acids Res., 2005, V. 33, N. 19, P. 6114–6123.

[3] *Nordberg E. K.*, YODA: selecting signature oligonucleotides // Bioinformatics, 2005, V. 21, N. 8, P. 1365–1370.

[4] *Wu C. T.*, *Liao C. Y.*, *Su H. J.*, IMPORT - Integrated Massive Probe's Optimal Recognition Tools // Genome Informatics, 2003, V. 14, P. 478–479.

[5] *SantaLucia J.*, *Hicks D.*, The thermodynamics of DNA structural motifs // Annu. Rev. Biophys. Biomol. Struct., 2004, V. 33, P. 415–440.

[6] *Dirks R.*, *Bois J.*, *Schaeffer J., et al.*, Thermodynamic analysis of interacting nucleic acid strands // SIAM Rev., 2007, V. 49, P. 65-88.

[7] *Zuker M.*, *Mathews D*, *Turner, D.*, Algorithms and Thermodynamics for RNA Secondary Structure Prediction: A Practical Guide // In RNA Biochemistry and Biotechnology, J. Barciszewski & B.F.C. Clark, eds., NATO ASI Series, Kluwer Academic Publishers, 1999, also available at http://www.bioinfo.rpi.edu/~zukerm.

[8] *Bishop M.A.*,*D'yachkov A.G.*, *Macula A.J.*, *Renz T.E.*, *Rykov V.V.*, Free Energy Gap and Statistical Thermodynamic Fidelity of DNA Codes // Journal of Computational Biology, 2007, V. 14, N. 8, P. 1088–1104.

[9] *D'yachkov A.G.*,*Macula A.J.*, *Renz T.E.*,*Vilenkin P.A*,*Ismagilov I.K*, New Results on DNA Codes // Proc. of the 2005 IEEE International Symposium on Information Theory, Adelaide, South Australia, Australia, September 4-9, 2005, P. 283–288.

[10] *D'yachkov A.G.*, *Macula A.J.*, *Torney D.C.*, *Vilenkin P.A.*, *White P.S.*, *Ismagilov I.K.*, *Sarbayev R.S.*, On DNA Codes // Probl. Peredachi Informatsii, 2005, V. 41, N. 4, P. 57–77, (in Russian). English translation: Problems of Information Transmission, V. 41, N. 4, 2005, P. 349–367.

[11] *Levenshtein V.I.*, Efficient Reconstruction of Sequences from Their Subsequences and Supersequences // J. Comb. Th., Ser. A, V. 93, 2001, P. 310–332.

[12] Cameron P.J., *Combinatorics: Topics, Techniques, Algorithms*, Cambridge University Press, 1994.