University of Nebraska - Lincoln

DigitalCommons@University of Nebraska - Lincoln

2010

On Critical Relative Distance of DNA Codes for Additive Stem Similarity

A. D'yachkov

A. Voronina

Anthony Macula

T. Renz

Vyacheslav Rykov

On Critical Relative Distance of DNA Codes for Additive Stem Similarity

A. D'yachkov, A. Voronina
Department of Probability Theory,
Faculty of Mechanics and Mathematics,
Moscow State University,
Moscow, 119992, Russia,
Email: agd-msu@yandex.ru,
vorronina@gmail.com.

A. Macula, T. Renz
Air Force Res. Lab.,
IFTC, Rome Research Site,
Rome NY 13441, USA,
Email: macula@geneseo.edu,
thomas.renz@rl.af.mil.

and V. Rykov
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 nearest-neighbor (stem) similarity model which adequately reflects the "hybridization potential" of two DNA sequences. Our aim is to present a survey of bounds on the rate of DNA codes with respect to a thermodynamically motivated similarity measure called an additive stem similarity. These results yield a method to analyze and compare known samples of the nearest neighbor "thermodynamic weights" associated to stacked pairs that occurred in DNA secondary structures.

I. INTRODUCTION

Single strands of DNA are represented by oriented sequences with elements from alphabet $\mathcal{A} \triangleq \{A, C, G, T\}$. 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 $x = (x_1 x_2 \ldots x_{n-1} x_n) \in \mathcal{A}^n = \{A, C, G, T\}^n$, let

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

denote its reverse complement. If $y = \widetilde{x}$, then $x = \widetilde{y}$ for any $x \in \mathcal{A}^n$. If $x = \widetilde{x}$, then x is called a *self reverse complementary* sequence. If $x \neq \widetilde{x}$, then a pair $(x, \widetilde{x})$ is called a *pair of mutually reverse complementary* sequences. A (perfect) *Watson-Crick duplex* is the joining of oppositely directed x and $\widetilde{x}$ so that every letter of one strand is paired with its complementary letter on the other strand in the double helix structure, i.e., x and $\widetilde{x}$ 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.

In general, crosshybridization is undesirable as it usually leads to experimental error. To increase the accuracy and throughput of the applications listed in [1]-[4], 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 of coding theory [5].

DNA nanotechnology often requires collections of DNA strands called *free energy gap codes* [6] 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* [3] and *DNA codes* [6]-[12].

The best known to date *biological* model, which is commonly utilized to estimate hybridization energy is the "nearest-neighbor thermodynamics" (see, references in [1]). Roughly, it implies that hybridization energy for any two DNA strands should be calculated as a sum of *thermodynamic weights* of all *stems* that were formed in the process of hybridization. Stem is defined as a pair of consecutive DNA letters of either of the strands, which coalesced with a pair of consecutive DNA letters of the other DNA strand. This biological model leads to a special *similarity function* on the space $\mathcal{A}^n$.

First known to authors constructions of DNA codes were suggested in [8]-[9]. They were based on conventional Hamming distance codes. Some methods of combinatorial coding theory have been developed [13]-[14] as a means by which such DNA codes can be found. From the very beginning it was understood that hybridization energy for DNA strands should be somehow simulated with the similarity function for sequences from $\mathcal{A}^n$. But it can be easily noticed, that Hamming similarity does not in the proper degree inherit the idea of "nearest-neighbor" similarity model. Thus there is no wonder that further exploration activities primarily focused on the search of appropriate similarity function.

One example of such function was proposed in [15], where it was calculated as the sum of weights of all elements, constituting the longest common Hamming subsequence. Later attempts included deletion similarity [7], which was earlier introduced by Levenshtein [16] and block similarity [11]-[12]. Both functions are non-additive which allowed for consideration of such cases as shifts of DNA sequences along each other. Nevertheless, all of them still did not catch the point of "nearest-neighbor" similarity model.

In 2008 we published our first work [17], devoted to the study of *stem* similarity functions. There we considered the simplest case, when similarity between two sequences from $\mathcal{A}^n$ is equal to the number of stems in the longest common Hamming subsequence between these two sequences. The common stem is understood as a block of length 2 which contains two adjacent elements of both of the initial sequences.

In [18], we introduced the concept of an additive stem w-similarity for an arbitrary weight function $w = w(a,b) > 0$, defined for all 16 elements $(ab) \in \mathcal{A}^2$, called stems. To calculate the additive stem w-similarity between two DNA sequences one should add up weights of all stems in the longest common Hamming subsequence between them (see, below Definition 1). Finally, our recent works [19]-[20] deal with non-additive stem w-similarity function, previously introduced in [6]. The given model also implies counting the weights of all formed stems between two DNA sequences with only difference that these stems are contained not in Hamming common subsequence but in subsequence in sense of Levenstein insertion-deletion metric. To find more detailed discussion of applicability of proposed constructions for modeling DNA hybridization assays please refer to work [6].

In current report we will summarize main results of [18] in study of asymptotic behavior of DNA codes maximal size for additive stem w-similarity function. We will show how these results lead to the development of possible criteria called a *critical relative w-distance of DNA codes* for distinguishing between weight samples $w(a,b)$ found in different experiments. We will also explain, how our consideration prompts the algorithms for composing DNA ensembles of optimal size for the given length of DNA strands.

II. Additive Stem w-Similarity Model

A. Notations and 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 $w = w(a,b) > 0$, $a, b \in \mathcal{A}$, be a weight function such that

$$w(a,b) = w(\bar{b}, \bar{a}), \quad a, b \in \mathcal{A}. \tag{2}$$

Condition (2) means that $w(a,b)$ is an invariant function under Watson-Crick transformation.

Definition 1: [6],[18]. For $x, y \in \mathcal{A}^n$, the number

$$\mathcal{S}_w(x,y) \triangleq \sum_{i=1}^{n-1} s_i^w(x,y), \quad \text{where}$$

$$s_i^w(x,y) \triangleq \begin{cases} w(a,b) & \text{if } x_i = y_i = a, \ x_{i+1} = y_{i+1} = b, \\ 0 & \text{otherwise}, \end{cases} \tag{3}$$

is called an *additive stem w-similarity* between x and y.

Function $\mathcal{S}_w(x, \widetilde{y})$ is used to model a *thermodynamic similarity* (*hybridization energy*) between DNA sequences x and y. In virtue of (2)-(3) the function

$$\mathcal{S}_w(x,y) = \mathcal{S}_w(y,x) \leq \mathcal{S}_w(x,x), \quad x, y \in \mathcal{A}^n \tag{4}$$

In addition,

$$\mathcal{S}_w(x,\widetilde{y}) = \mathcal{S}_w(y,\widetilde{x}), \qquad x, y \in \mathcal{A}^n. \tag{5}$$

Identity (5) implies the symmetry property of hybridization energy between DNA sequences x and y [6]-[12].

Example 1: In [17] we considered *constant weights* $w = w(a,b) \equiv 1$, $a, b \in \mathcal{A}$, for which the additive stem 1-similarity $\mathcal{S}_1(x,y)$, $0 \leq \mathcal{S}_1(x,y) \leq \mathcal{S}_1(x,x) = n - 1$, is the above-mentioned number of *stems* in the longest common Hamming subsequence between x and y.

Example 2: Table 1 shows a biologically motivated collection of weights $w(a,b) \triangleq U(a,b)$ called [1] *unified weights*:

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

Table 1: Unified weights $U(a,b)$, 1998.

The given values $U(a,b)$ are based on weight samples which come from [1] and [4] and are the nearest neighbor "thermodynamic weights" (e.g., free energy of formation) associated to stacked pairs that occurred in DNA secondary structures. See [2] for an introduction to the nearest neighbor model.

Taking into account inequality (4), we give

Definition 2: [6],[18]. The number

$$\mathcal{D}_w(x,y) \triangleq \mathcal{S}_w(x,x) - \mathcal{S}_w(x,y) = \sum_{i=1}^{n-1} \eta_i^w(x,y),$$

$$\eta_i^w(x,y) \triangleq s_i^w(x,x) - s_i^w(x,y) \geq 0, \tag{6}$$

is called an *additive stem w-distance* between $x, y \in \mathcal{A}^n$.

Let $x(j) \triangleq (x_1(j)x_2(j) \ldots x_n(j)) \in \mathcal{A}^n$, $j \in [N]$, be *codewords* of a q-ary *code* $\mathbf{X} = \{x(1), x(2), \ldots, x(N)\}$ of *length* n and *size* N, where $N = 2, 4, \ldots$ is an *even* number. Let D, $0 < D \leq \max_{x \in \mathcal{A}^n} \mathcal{S}_w(x,x)$, be an arbitrary positive number.

Definition 3: [6],[18]. A code $\mathbf{X}$ is called a DNA code of distance D for additive stem w-similarity (3) (or a $(n, D)_w$-*code*) if the following two conditions are fulfilled. (i). For any integer $j \in [N]$, there exists $j' \in [N]$, $j' \neq j$, such that $x(j') = \widetilde{x(j)} \neq x(j)$. In other words, X is a collection of $N/2$ pairs of mutually reverse complementary sequences. (ii). The minimal w-distance of code X is

$$\mathcal{D}_w(X) \triangleq \min_{j \neq j'} \mathcal{D}_w(x(j), x(j')) \geq D. \tag{7}$$

Let $N_w(n, D)$ be the *maximal* size of DNA $(n, D)_w$-codes for distance (6). If $d > 0$ is a fixed number, then

$$R_w(d) \triangleq \varlimsup_{n \to \infty} \frac{\log_4 N_w(n, nd)}{n}, \qquad d > 0, \tag{8}$$

is called a *rate* of DNA $(n, nd)_w$-codes for the *relative distance* $d > 0$.

B. Construction

Theorem 1: If $n = 2t + 1$, $t = 1, 2, \ldots$, then

$$N_{\mathbf{1}}(n, n - 1) = 16.$$

Proof: Codewords of $(n, n-1)_{\mathbf{1}}$-code should not contain any common stems with each other. Note, that $|\mathcal{A}^2| = 16$ and hence for any $(n, n-1)_{\mathbf{1}}$-code $\mathbf{X} = \{x(1), \ldots x(N)\}$

$$|\{(x_1(u) x_2(u)), \ u \in [N]\}| \leq |\mathcal{A}^2| = 16.$$

Thus,

$$N_{\mathbf{1}}(n, n - 1) \leq 16.$$

Obviously, for odd n, the set $\mathcal{A}^n$ doesn't contain self reverse complementary words. For stem $\mathbf{a} \triangleq (a_1 a_2) \in \mathcal{A}^2$, define $\hat{x}(\mathbf{a}) \triangleq (a_1 a_2 a_1 a_2 \ldots a_2 a_1 a_2 a_1) \in \mathcal{A}^n$. Code

$$\hat{\mathbf{X}}_n \triangleq \{\hat{x}(\mathbf{a}), \ \mathbf{a} \in \mathcal{A}^2\}, \quad n = 2t + 1, \ t = 1, 2, \ldots$$

constitute a DNA $(n, n - 1)_{\mathbf{1}}$-code of size $|\hat{\mathbf{X}}_n| = 4^2 = 16$ for additive stem 1-similarity. ∎

Example 3: For instance, if $n = 5$, $D = n - 1 = 4$, then 8 pairs of mutually reverse complementary codewords of code $\hat{\mathbf{X}}_n$ are:

$$(AAAAA, TTTTT), \quad (ACACA, TGTGT),$$
$$(CCCCC, GGGGG), \quad (CACAC, GTGTG),$$
$$(AGAGA, TCTCT), \quad (ATATA, TATAT),$$
$$(CGCGC, GCGCG), \quad (CTCTC, GAGAG).$$

Remark 1: Note that for any weight function w, the additive stem w-similarity $\mathcal{S}_w(\hat{x}(\mathbf{a}), \hat{x}(\mathbf{b})) = 0$, $\mathbf{a}, \mathbf{b} \in \mathcal{A}^2$, $\mathbf{a} \neq \mathbf{b}$. Hence, the minimal w-distance (7) of code $\hat{\mathbf{X}}_n$ is

$$\mathcal{D}_w(\hat{\mathbf{X}}_n) = \min_{\mathbf{a} \in \mathcal{A}^2} \mathcal{S}_w(\hat{x}(\mathbf{a}), \hat{x}(\mathbf{a})) \geq (n - 1) \cdot \underline{w},$$

where $\underline{w} \triangleq \min_{a, b \in \mathcal{A}} w(a, b)$. Thus, for any weight function w, the code $\hat{\mathbf{X}}_n$ is also a $(n, (n - 1) \cdot \underline{w})_w$-code. For example, for the additive stem U-similarity of Example 2, the number $\underline{w} = 1$. Therefore, the code $\hat{\mathbf{X}}_n$ is a $(n, n - 1)_U$-code.

C. Bounds on Rate $R_w(d)$

Let $\mathsf{p} \triangleq \{p(a, b), \ a, b \in \mathcal{A}\}$ be an arbitrary joint probability distribution on the set of stems $(ab) \in \mathcal{A}^2$, i.e.,

$$\sum_{a, b \in \mathcal{A}} p(a, b) = 1, \qquad p(a, b) \geq 0 \quad \text{for any} \quad a, b \in \mathcal{A}.$$

To describe bounds on the rate $R_w(d)$, we will consider joint probability distributions p, such that the corresponding marginal probabilities coincide, i.e., for any $a \in \mathcal{A}$

$$p_1(a) \triangleq \sum_{b \in \mathcal{A}} p(a, b) = \sum_{b \in \mathcal{A}} p(b, a) \triangleq p_2(a) > 0 \quad (9)$$

and, in addition, function $p(a, b)$, as well as weight function (2), is invariant under Watson-Crick transformation, i.e.,

$$p(a, b) = p(\overline{b}, \overline{a}) \quad \text{for any} \quad a, b \in \mathcal{A}. \quad (10)$$

Let

$$p_1(b|a) \triangleq \frac{p(a, b)}{p_1(a)}, \quad p_2(b|a) \triangleq \frac{p(b, a)}{p_2(a)}$$

denote the corresponding *conditional* probabilities. It is easy to check, that for distributions p with properties (9)-(10), and for the corresponding conditional probabilities, the following equalities hold true for any $a, b \in \mathcal{A}$:

$$p_1(a) = p_2(a) = p_1(\overline{a}) = p_2(\overline{a}), \quad p_1(b|a) = p_2(\overline{b}|\overline{a}). \quad (11)$$

For a fixed weight function (2), introduce values

$$T_w \triangleq \max_{(9)} T_w(\mathsf{p}),$$

$$T_w(\mathsf{p}) \triangleq \sum_{a, b \in \mathcal{A}} \left(p(a, b) - p^2(a, b) \right) w(a, b), \quad (12)$$

where the maximum is taken over all distributions p for which condition (9) hold true. Note, that if weight function is invariant under Watson-Crick transformation, then maximizing distribution of (12) will satisfy conditions (10)-(11).

Applying an analog of the conventional Plotkin bound [5], one can prove

Theorem 2: [18] If $d \geq T_w$, then $R_w(d) = 0$.

Let $x = (x_1 x_2 \ldots x_n) \in \mathcal{A}^n$ be the stationary Markov chain with initial distribution $p_1(a)$, $a \in \mathcal{A}$, and transition matrix $P = \|p_1(b|a)\|$, $a, b \in \mathcal{A}$, i.e.

$$\Pr\{x_i = a\} \triangleq p_1(a), \ \Pr\{x_{i+1} = b | x_i = a\} \triangleq p_1(b|a) \quad (13)$$

for any $a, b \in \mathcal{A}$ and $i \in [n - 1]$.

Let a distribution p satisfy (9) and let also the following *Markov condition* $\mathcal{M}$ be fulfilled: *transition matrix P must define such Markov chain $x = (x_1 x_2 \ldots x_n)$, that for any pair of states $a, b \in \mathcal{A}$ there exists an integer $m \in [4]$ such that the conditional probability $\Pr\{x_{m+1} = b | x_1 = a\} > 0$.*

Theorem 3: [18] For any probability distribution p, satisfying condition (9) and Markov condition $\mathcal{M}$, and any relative distance d, $0 < d < T_w(\mathsf{p})$, the rate $R_w(d) > 0$.

Theorem 3 is established using the ensemble of random codes where independent codewords $x = (x_1 x_2 \ldots x_n)$ are identically distributed in accordance with the Markov chain (13) and, in virtue of (11), the corresponding reverse complement codewords $\widetilde{x} = (\overline{x}_n \overline{x}_{n-1} \ldots \overline{x}_2 \overline{x}_1)$ have the same distribution (13) as well. In addition, the proof of Theorem 3 is based on the Perron-Frobenius theorem (see [21], Theorem 3.1.1).

Let $T_w(\mathsf{p})$ be defined by (12) and

$$T_w^{\mathcal{M}} \triangleq \max_{(9), \mathcal{M}} T_w(\mathsf{p}). \quad (14)$$

If $T_w = T_w^{\mathcal{M}}$, then the corresponding weight function $w = w(a, b)$ is called *regular*, and non-regular otherwise. If a weight function $w = w(a, b)$ is regular, then T_w is called the *critical relative distance* of $(n, dn)_w$-codes.

From Theorem 2 and 3 it follows

Corollary 1: [18] If a weight function $w = w(a, b)$ is regular, then the maximal size of $(n, nd)_w$-codes increases exponentially with increasing n if and only if $0 < d < T_w$.

Remark 2: Results of Theorem 3 prompts an idea, that the ensemble of optimal random DNA codes for non-additive stem w-similarity should be based on a generator of independent Markov chains with transition matrix P and initial distribution $p_1(a)$, such that the corresponding distribution p affords maximum in (14).

III. Weight Sample Analysis Based on Criterion of Critical Relative Distance

In this section, we will discuss *samples of weight function* (or, briefly, *weight samples*) $w = w(a,b)$, $a,b \in \mathcal{A}$, taken from SantaLucia (1998) (see Table 1 in [1]). In Tables 2-8, we present weights $w(A,A) = w(T,T)$ and samples of *relative weights* $\widetilde{w}(a,b)$ with respect to $w(A,A)$, i.e., for any $a,b \in \mathcal{A}$,

$$\widetilde{w} = \widetilde{w}(a,b) \triangleq \frac{w(a,b)}{w(A,A)}, \quad \widetilde{w}(a,b) = \widetilde{w}(\bar{b},\bar{a}). \quad (15)$$

Pure numbers $\widetilde{w}(a,b)$ are comfortable for a mutual comparison and for the comparison with unified weights of Table 1.

$w(A,A) = 0.43$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	1.00	2.28	1.93	0.63
$a = C$	2.32	2.84	3.95	1.93
$a = G$	2.16	3.81	2.84	2.28
$a = T$	0.51	2.16	2.32	1.00

Table 2: Gotoh, 1981.

$w(A,A) = 0.89$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	1.00	1.35	1.52	0.91
$a = C$	1.54	1.84	2.24	1.52
$a = G$	1.40	2.20	1.84	1.35
$a = T$	0.85	1.40	1.54	1.00

Table 3: Vologodskii, 1984.

$w(A,A) = 0.67$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	1.00	1.69	1.75	0.93
$a = C$	1.78	2.31	2.79	1.75
$a = G$	1.67	2.76	2.31	1.69
$a = T$	1.04	1.67	1.78	1.00

Table 4: Blake, 1991.

$w(A,A) = 0.93$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	1.00	1.63	1.11	0.89
$a = C$	1.35	1.80	1.77	1.11
$a = G$	1.68	2.62	1.80	1.63
$a = T$	0.75	1.68	1.35	1.00

Table 5: Benight, 1992.

$w(A,A) = 1.02$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	1.00	1.40	1.14	0.72
$a = C$	1.35	1.74	2.05	1.14
$a = G$	1.43	2.24	1.74	1.40
$a = T$	0.59	1.43	1.35	1.00

Table 6: SantaLucia, 1996.

$w(A,A) = 1.20$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	1.00	1.25	1.25	0.75
$a = C$	1.42	1.75	2.33	1.25
$a = G$	1.25	1.92	1.75	1.25
$a = T$	0.75	1.25	1.42	1.00

Table 7: Sugimoto, 1996.

$w(A,A) = 1.66$	$b = A$	$b = C$	$b = G$	$b = T$
$a = A$	1.00	0.68	0.81	0.72
$a = C$	1.08	1.66	1.98	0.81
$a = G$	0.85	1.70	1.66	0.68
$a = T$	0.46	0.85	1.08	1.00

Table 8: Breslauer, 1986.

A. Analysis of Tables 1-8 for Additive $\widetilde{w}$-Distance

Analysis of Table 1 and Tables 3-7: The given weight samples are regular and the maximum in (12) is attained when $p(a,b) = 0$ if stem $(ab) \in L_4$, where the set L_4 of forbidden stems in the Markov chain (13) maximizing (12) has the form

$$L_4 \triangleq \{(AT),(TA),(AA),(TT)\}. \quad (16)$$

Below, in Table 1' and Tables 3'-7', we present the estimated values of joint probabilities $p(a,b)$ and marginal probabilities $p_1(a)$ for which the maximum in (12) is attained. Values of the critical relative distance $T_{\widetilde{w}}$ are given as well.

$p(a,b)$	$b = A$	$b = C$	$b = G$	$b = T$	$p_1(a)$
$a = A$	0	.0589	.0081	0	.067
$a = C$	.0610	.1544	.2095	.0081	.433
$a = G$	.0060	.2136	.1544	.0589	.433
$a = T$	0	.0060	.0610	0	.067

Table 1': Unified weights $U(a,b)$. $T_U = 1.58$.

$p(a,b)$	$b = A$	$b = C$	$b = G$	$b = T$	$p_1(a)$
$a = A$	0	.0706	.0080	0	.078
$a = C$	.0638	.1411	.2087	.0080	.422
$a = G$	.0147	.1951	.1411	.0706	.422
$a = T$	0	.0147	.0638	0	.078

Table 3': Vologodskii, 1984. $T_{\widetilde{w}} = 1.61$.

$p(a,b)$	$b = A$	$b = C$	$b = G$	$b = T$	$p_1(a)$
$a = A$	0	.0331	.0346	0	.068
$a = C$	.0406	.1535	.2037	.0346	.432
$a = G$	.0270	.2188	.1535	.0331	.432
$a = T$	0	.0270	.0406	0	.068

Table 4': Blake, 1991. $T_{\widetilde{w}} = 1.97$.

$p(a,b)$	$b = A$	$b = C$	$b = G$	$b = T$	$p_1(a)$
$a = A$	0	.0675	.0144	0	.082
$a = C$	.0478	.1326	.2234	.0144	.418
$a = G$	.0340	.1841	.1326	.0675	.418
$a = T$	0	.0340	.0478	0	.082

Table 5': Benight, 1992. $T_{\widetilde{w}} = 1.58$.

$p(a,b)$	$b=A$	$b=C$	$b=G$	$b=T$	$p_1(a)$
$a=A$	0	.0608	.0095	0	.070
$a=C$	.0616	.1499	.2087	.0095	.430
$a=G$	.0087	.2102	.1499	.0608	.430
$a=T$	0	.0087	.0616	0	.070

Table 6': SantaLucia, 1996. $T_{\widetilde{w}} = 1.52$.

$p(a,b)$	$b=A$	$b=C$	$b=G$	$b=T$	$p_1(a)$
$a=A$	0	.0507	.0140	0	.065
$a=C$	.0444	.1551	.2217	.0140	.435
$a=G$	.0203	.2091	.1551	.0507	.435
$a=T$	0	.0203	.0444	0	.065

Table 7': Sugimoto, 1996. $T_{\widetilde{w}} = 1.50$.

Analysis of Table 2: The given weight sample is regular and the maximum in (12) is attained when $p(a,b) = 0$ if stem $(ab) \in L_6$, where the set L_6 of forbidden stems in the Markov chain (13) maximizing (12) has the form

$$L_6 = \{(AT),(TA),(AA),(TT),(AG),(CT)\}. \qquad (17)$$

Below, in Table 2', we present the estimated values of joint $p(a,b)$ and marginal $p_1(a)$ probabilities for which the maximum in (12) is attained. The estimated value of critical relative distances $T_{\widetilde{w}} = 2.60$ is given as well.

$p(a,b)$	$b=A$	$b=C$	$b=G$	$b=T$	$p_1(a)$
$a=A$	0	.0593	0	0	.059
$a=C$	.0466	.1427	.2515	0	.441
$a=G$	.0127	.2261	.1427	.0593	.441
$a=T$	0	.0127	.0466	0	.059

Table 2': Gotoh, 1981. $T_{\widetilde{w}} = 2.60$.

Analysis of Table 8: The given weight sample $\widetilde{w}$ is a non-regular weight sample because the maximum in (12) is attained (with the maximal value $T_{\widetilde{w}} = 1.70$) for probability distribution $p'(a,b)$, $(ab) \in \mathcal{A}^2$, which does not satisfy *Markov condition* $\mathcal{M}$ and has the form:

$p'(a,b)$	$b=A$	$b=C$	$b=G$	$b=T$	$p_1'(a)$
$a=A$	.0344	0	0	0	.034
$a=C$	0	.2190	.2466	0	.466
$a=G$	0	.2466	.2190	0	.466
$a=T$	0	0	0	.0344	.034

Table 8': Breslauer, 1986. $T'_{\widetilde{w}} = 1.70$.

This implies that for weight sample $\widetilde{w}$ from Table 8, we cannot estimate the critical relative distance of optimal DNA codes based on additive stem $\widetilde{w}$-similarity.

B. Conclusion

For regular weight samples from Tables 1-7 (T1-T7), the descriptive analysis and comparison of critical parameters are summarized as follows:

	T1	T2	T3	T4	T5	T6	T7
L	L_4	L_6	L_4	L_4	L_4	L_4	L_4
$T_{\widetilde{w}}$	1.58	2.60	1.61	1.97	1.58	1.52	1.50

These results confirm the main conclusion of paper [1] about a consensus agreement among the parameters determined from six laboratories (see, Table 2-7). In addition, there is an excellent agreement between parameters of Table 3, Tables 5-7 and unified parameters (see, Table 1) suggested in [1].

REFERENCES

[1] J. SantaLucia, "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics," *Proc. National Academy of Sciences USA*, vol. 95, pp. 1460–1465, 1998.

[2] M. Zuker, D. Mathews, D. Turner, "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.

[3] L. Kaderali, A. Deshpande, J. Nolan, P. White, "Primer-design for multiplexed genotyping," *Nucleic Acids Res.*, vol. 31, pp. 1796–1802, 2003.

[4] J. SantaLucia, D. Hicks, "The thermodynamics of DNA structural motifs," *Annu. Rev. Biophys. Biomol. Struct.*, vol. 33, pp. 415–440, 2004.

[5] F. J. MacWilliams, N. J. A. Sloane, *The Theory of Error-correcting Codes*, Amsterdam, The Netherlands: North Holland, 1977.

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

[7] A. G. D'yachkov, P. A. Vilenkin, D. C. Torney, P. S. White, "Reverse-Complement Similarity Codes for DNA Sequences," // in *Proc. 2000 IEEE Int. Symp. Information Theory*, Sorrento, Italy, 2000, pp. 330.

[8] V.V. Rykov , A.J. Macula , C.M.Korzelius, D.C. Engelhart, D.C. Torney, P.C. White, "DNA Sequences Constructed on the Basis of Quaternary Cyclic Codes". *Proceedings of 4-th World Multiconference on Systemics, Cybernetics and Informatics*, Orlando, Florida, USA, July 2000.

[9] A. Marathe, A. E. Condon, R. M. Corn, "On combinatorial DNA design," *J. Comp. Biol.*, vol. 8, pp. 201–219, 2001.

[10] A. G. D'yachkov, P. L. Erdos, A. J. Macula, V. V. Rykov, D. C. Torney, C. S. Tung, P. A. Vilenkin, P. S. White, "Exordium for DNA Codes," *J. Comb. Optimization*, vol. 7, n. 4, pp. 369–379, 2003.

[11] A. G. D'yachkov, A. J. Macula, T. E. Renz, P. A. Vilenkin, I. K. Ismagilov, "New Results on DNA Codes," in *Proc.* 2005 *IEEE Int. Symp. Information Theory*, Adelaide, South Australia, Australia, 2005, pp. 283–288.

[12] A. G. D'yachkov, A. J. Macula, D. C. Torney, P. A. Vilenkin, P. S. White, I. K. Ismagilov, R. S. Sarbayev, "On DNA Codes," *Problems of Information Transmission*, vol. 41, n. 4, pp. 349–367, 2005.

[13] O. Milenkovic, N. Kashyap, "New Constructions of Codes for DNA computing," *Proc. 2005 International Workshop on Coding and Cryptography (WCC 2005)*, Bergen, Norway, 2005, pp. 204-213.

[14] T. Abualrub, A. Ghrayeb, X. N. Zeng, "Construction of cyclic codes over $GF(4)$ for DNA computing," *Journal of the Franklin Institute*, vol. 343, n. 4-5, pp. 448–457, 2006.

[15] A. G. D'yachkov, D. C. Torney, "On similarity codes," *IEEE Trans. Inform. Th.*, vol. 46, n. 4, pp. 1558–1664, 2000.

[16] V. I. Levenshtein, "Efficient Reconstruction of Sequences from Their Subsequences and Supersequences," *J. Comb. Th., Ser. A*, vol. 93, pp. 310–332, 2001.

[17] A. G. D'yachkov, A. N. Voronina, "DNA Codes Based on Stem Hamming Similarity," in *Proc. 11th Int. Workshop Algebraic and Combinatorial Coding Theory*, Pamporovo, Bulgaria, 2008, pp. 85–91.

[18] A. G. D'yachkov, A. N. Voronina, "DNA Codes for Additive Stem Similarity," *Problems of Information Transmission*, vol. 45, n. 2, pp. 348–367, 2009.

[19] A. G. D'yachkov, A. J. Macula, T. E. Renz, V. V. Rykov, "Random Coding Bounds for DNA Codes Based on Fibonacci Ensembles of DNA Sequences," in *2008 IEEE Int. Symp. Information Theory*, Toronto, Canada, 2008, pp. 2292–2296.

[20] A. G. D'yachkov, A. N. Voronina, A. J. Macula, T. E. Renz, V. V. Rykov, "DNA Codes for the Nearest-Neighbor Similarity," submitted to *IEEE Trans. Inform. Th.*.

[21] Dembo, A., Zeitouni, O., *Large Deviations Techniques and Applications*, Boston, MA: Jones and Bartlett, 1993.