DNA Codes Based on Stem Similarities Between DNA Sequences

Arkadii D'yachkov[1], Anthony Macula[2], Vyacheslav Rykov[3], and Vladimir Ufimtsev[3]

[1] Moscow State University, Moscow 119992, Russia
agd-msu@yandex.ru
[2] Air Force Res. Lab., IFTC, Rome Research Site, Rome NY 13441, USA
macula@geneseo.edu
[3] University of Nebraska at Omaha, 6001 Dodge St., Omaha, NE 68182-0243 USA
vrykov@mail.unomaha.edu

Abstract. DNA codes consisting of DNA sequences are necessary for DNA computing. The minimum distance parameter of such codes is a measure of how dissimilar the codewords are, and thus is indirectly a measure of the likelihood of undetectedable or uncorrectable errors occurring during hybridization. To compute distance, an abstract metric, for example, longest common subsequence, must be used to model the actual bonding energies of DNA strands. In this paper we continue the development [1,2,3] of similarity functions for q-ary n-sequences The theoretical lower bound on the maximal possible size of codes, built on the space endowed with this metric, is obtained. that can be used (for $q = 4$) to model a thermodynamic similarity on DNA sequences. We introduce the concept of a stem similarity function and discuss DNA codes [2] based on the stem similarity. We suggest an optimal construction [2] and obtain random coding bounds on the maximum size and rate for such codes.

1 Introduction

In order to accomplish DNA computing, it is necessary to have DNA libraries, also known as DNA codes, of large size and small energies of hybridization between the DNA sequences. The ultimate criterion for the value of a metric for DNA codes is the degree to which it approximates actual bonding energies, which in turn determines the degree to which distance approximates the likelihood of one codeword mistakenly binding to the reverse complement of another codeword. We can use a branch of mathematics known as coding theory, that was initiated around the same time that the structure of DNA was discovered, to study the space of DNA sequences endowed with a measure of distance (metric). The introduced measure of distance between DNA sequences has an immediate application in determining the similarities between genes, expressed as DNA sequences, in any existing genome. Codes built on spaces of DNA sequences can be implemented in Biomolecular Computing and could have other important applications.

M.H. Garzon and H. Yan (Eds.): DNA 13, LNCS 4848, pp. 146–151, 2008.

2 Notations, 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 $q = 2, 4, \ldots$ be an arbitrary even integer, $A \triangleq \{0, 1, \ldots, q-1\}$ be the standard q-nary alphabet. Consider two arbitrary q-nary n-sequences $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in A^n$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n) \in A^n$. By symbol $\mathbf{z} = (z_1, z_2, \ldots, z_\ell) \in A^\ell$, $\ell \in [n]$, we will denote a *common subsequence* [5] 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 $1 \le b \le r \le n$ be arbitrary integers. A fixed r-sequence $\mathbf{a} = (a_1, a_2, \ldots, a_r)$, $a_i \in A = \{0, 1, \ldots, q-1\}$, $i \in [r]$, 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 $r-(b-1)$ common b-stems $a_i, a_{i+1}, \ldots, a_{i+(b-1)}$, $i \in [r-(b-1)]$, containing b adjacent symbols of the given common $(\mathbf{x}, \mathbf{y})$-block.*

Definition 2. *Let $1 \le t \le \ell \le n$ be integers. A sequence $\mathbf{z} = (z_1, z_2, \ldots, z_\ell)$, $z_i \in A$, $i \in [\ell]$, is called a common t-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 $\ge t$.*

Let $\mathcal{Z}_t(\mathbf{x}, \mathbf{y})$ be the set of all common t-block subsequences between $\mathbf{x}$ and $\mathbf{y}$. For any $\mathbf{z} \in \mathcal{Z}_t(\mathbf{x}, \mathbf{y})$, we denote by $k(\mathbf{z}, \mathbf{x}, \mathbf{y})$, $1 \le k(\mathbf{z}, \mathbf{x}, \mathbf{y}) \le |\mathbf{z}|/t$, the *minimal number* of common $(\mathbf{x}, \mathbf{y})$–blocks which *constitute* the given subsequence $\mathbf{z}$.

Note that for any integer b, $2 \le b \le t$, the difference $|\mathbf{z}| - (b-1)\, k(\mathbf{z}, \mathbf{x}, \mathbf{y})$, $\mathbf{z} \in \mathcal{Z}_t(\mathbf{x}, \mathbf{y})$, is a total number of common b-stems containing adjacent symbols in common $(\mathbf{x}, \mathbf{y})$-blocks constituting $\mathbf{z} \in \mathcal{Z}_t(\mathbf{x}, \mathbf{y})$.

Definition 3. *For any fixed integer b, $2 \le b \le n$,we define*

$$S_b(\mathbf{x}, \mathbf{y}) \triangleq \max_{b \le t \le n} \max_{\mathbf{z} \in \mathcal{Z}_t(\mathbf{x}, \mathbf{y})} \{|\mathbf{z}| - (b-1)\, k(\mathbf{z}, \mathbf{x}, \mathbf{y})\}, \qquad S_b(\mathbf{x}, \mathbf{y}) \ge 0.$$

If $\mathcal{Z}_b(\mathbf{x}, \mathbf{y}) = \varnothing$, then we will say that $S_b(\mathbf{x}, \mathbf{y}) \triangleq 0$. The number

$$S_b(\mathbf{x}, \mathbf{y}) = S_b(\mathbf{y}, \mathbf{x}) \le S_b(\mathbf{x}, \mathbf{x}) = n - (b-1), \quad \mathbf{x} \in A^n, \quad \mathbf{y} \in A^n,$$

is called an b-stem similarity between $\mathbf{x}$ and $\mathbf{y}$. For $b = 2$, the concept of 2-stem similarity and its biological motivation were suggested in [1].

Definition 4. *[1,2]. If $q = 2, 4, \ldots$, then $\bar{x} \triangleq (q-1) - x$, $x \in A = \{0, 1, \ldots, q-1\}$, is called a complement of a letter x. For $\mathbf{x} = (x_1, x_2, \ldots, x_{n-1}, x_n) \in A^n$, we define its reverse complement $\widetilde{\overline{\mathbf{x}}} \triangleq (\bar{x}_n, \bar{x}_{n-1}, \ldots, \bar{x}_2, \bar{x}_1) \in A^n$. If $\mathbf{y} \triangleq \widetilde{\overline{\mathbf{x}}}$, then $\mathbf{x} = \widetilde{\overline{\mathbf{y}}}$ for any $\mathbf{x} \in A^n$. If $\mathbf{x} = \widetilde{\overline{\mathbf{x}}}$, then $\mathbf{x}$ is called a self reverse complementary sequence. If $\mathbf{x} \ne \widetilde{\overline{\mathbf{x}}}$, then a pair $(\mathbf{x}, \widetilde{\overline{\mathbf{x}}})$ is called a pair of mutually reverse complementary sequences.*

Let $\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N)$, where $\mathbf{x}(j) \triangleq (x_1(j), x_2(j), \ldots, x_n(j)) \in A^n$, $j \in [N]$, be *codewords* of a q-ary *code* $X = \{\mathbf{x}(1), \mathbf{x}(2), \ldots, \mathbf{x}(N)\}$ of *length* n and *size* N, where $N = 2, 4, \ldots$ be an *even* number. Let b, $2 \le b \le n$, and D, $b \le D \le n-1$, be arbitrary integers.

Definition 5. *A code X is called a DNA (n, D)-code based on b-stem similarity $S_b(\mathbf{x}, \mathbf{y})$ (briefly, (n, D)-code) if the following two conditions are fulfilled.*

(i). For any number $j \in [N]$ there exists $j' \in [N]$, $j' \ne k$, such that $\mathbf{x}(j') = \widetilde{\mathbf{x}(j)} \ne \mathbf{x}(j)$. In other words, X is a collection of $N/2$ pairs of mutually reverse complementary sequences.

(ii). For any $j, j' \in [N]$, where $j \ne j'$, the similarity

$$S_b(\mathbf{x}(j), \mathbf{x}(j')) \le n - D - 1, \qquad b \le D \le n-1. \tag{1}$$

Definition 6. *Let $N_b(n, D)$ be the maximum size for DNA (n, D)-codes based on b-stem similarity. If d, $0 < d < 1$, is a fixed number, then*

$$R_b(d) \triangleq \varlimsup_{n \to \infty} \frac{\log_q N_b(n, \lfloor nd \rfloor)}{n} \tag{2}$$

is called a rate of DNA $(n, \lfloor nd \rfloor)$-codes based on b-stem similarity.

3 Random Coding Bounds

Let b, $2 \le b \le n$, and s, $0 \le s \le n - (b-1)$, be arbitrary integers and

$$\mathcal{P}_b(n, s) \triangleq \{(\mathbf{x}, \mathbf{y}) \in A^n \times A^n \,:\, S_b(\mathbf{x}, \mathbf{y}) = s\},$$

$$\overline{\mathcal{P}_b}(n, s) \triangleq \{\mathbf{x} \in A^n \,:\, S_b(\mathbf{x}, \tilde{\mathbf{x}}) = s\},$$

be sets of pairs $(\mathbf{x}, \mathbf{y}) \in A^n \times A^n$ (sequences $\mathbf{x} \in A^n$) for which the given similarities be equal to s. Applying combinatorial arguments which are similar to the corresponding arguments of paper [2] for the block similarity function, one can check that the following statement is true.

Lemma 1. *The size*

$$|\mathcal{P}_b(n, s)| \;\le\; q^{2n-s} \cdot \sum_{k=1}^{\min\{s\,,\,(n-s)/(b-1)\}} q^{-(b-1)k} \binom{s-1}{k-1} \binom{n-s-(b-2)k}{k}^2. \tag{3}$$

The set $\overline{\mathcal{P}_b}(n, s)$ is empty if $s \ge 3$ is odd. If $s \ge 2$ is even, then the size

$$|\overline{\mathcal{P}_b}(n, s)| \le q^{n-s/2} \cdot \sum_{k=1}^{\min\{s\,,\,(n-s)/(b-1)\}} q^{-(b-1)k/2} \binom{s/2-1}{k/2-1} \binom{n-s-(b-2)k}{k}. \tag{4}$$

Lemma 1 and the standard random coding method [2] lead to Theorems 1 and 3 which give lower bounds on the size $N_b(n, D)$ and rate $R_b(d)$ of DNA codes based on b-stem similarity.

Theorem 1. *If $D \geq b \geq 2$ are fixed integers and $n \to \infty$, then*

$$N_b(n, D) \geq \frac{1}{4} \cdot \frac{(D_b - 1)! \cdot q^{(b-1)D_b}}{\binom{D-(b-2)\,D_b}{D_b}^2 \cdot q^D} \cdot \frac{q^n}{n^{D_b - 1}} \cdot (1+o(1)), \quad D_b \triangleq \left\lfloor \frac{D}{b-1} \right\rfloor. \tag{5}$$

For the case $b = 2$, number $D_2 = D \geq 2$ and bound (5) has the form

$$N_2(n, D) \geq \frac{(D-1)!}{4} \cdot \frac{q^n}{n^{D-1}} \cdot (1 + o(1)), \qquad D \geq 2. \tag{6}$$

For the case $D = b \geq 3$, number $D_b = 1$ and bound (5) has the form

$$N_b(n, b) \geq \frac{q^{n-1}}{16} \cdot (1 + o(1)), \qquad b \geq 3. \tag{7}$$

An improvement of asymptotic lower bounds (6)-(7) follows from formula (8) for $N_b(n, b)$ presented in the theorem.

Theorem 2. *[2] If $n = qm$, $m = 1, 3, 5, \ldots$, then*

$$N_b(n, b) = \frac{q^{n-1} + q}{2}, \qquad 2 \leq b \leq n - 1. \tag{8}$$

Introduce the standard symbol

$$h_q(u) \triangleq -u \log_q u - (1-u) \log_q (1-u), \qquad 0 < u < 1, \tag{9}$$

for the binary entropy function.

Theorem 3. *(i). The rate*

$$R_b(d) \geq \underline{R}_b(d) \triangleq \min_{0 \leq u \leq d} \{(1-u) - E_b(u)\}, \tag{10}$$

where

$$E_b(u) \triangleq \max_{0 \leq v \leq \min\left\{\frac{u}{b-1},\, 1-u\right\}} F_b(v, u), \tag{11}$$

$$F_b(v, u) \triangleq -(b-1)v + (1-u)\, h_q\left(\frac{v}{1-u}\right) + \\ + 2\,[u - (b-2)v]\, h_q\left(\frac{v}{u-(b-2)v}\right). \tag{12}$$

(ii). Let d_b, $0 < d_b < 1$, be the unique root of equation $1 - d = E_b(d)$. If $0 < d < d_b$, then the rate $R_b(d) > 0$ and the following lower bound

$$R_b(d) \geq \underline{R}_b(d) \triangleq (1-d) - E_b(d), \qquad 0 < d < d_b, \tag{13}$$

holds.

We will say that the number d_b, $0 < d_b < 1$, is a *critical distance fraction* for the random coding bound $\underline{R}_b(d)$.

Maximization (11)-(12). The derivative of binary entropy function (9) is

$$h'_q(v) = \log_q \frac{1-v}{v}, \qquad 0 < v < 1.$$

Thus, the partial derivative of function $F_b(v,u)$ is

$$\frac{\partial F_b(v,u)}{\partial v} \;=\; -(b-1) \;+\; \log_q \frac{(1-u)-v}{v} \;+$$

$$+2\left[-(b-2)\,h_q\left(\frac{v}{u-(b-2)v}\right) \;+\; \frac{u}{u-(b-2)v}\,\log_q\,\frac{u-(b-1)v}{v}\right]. \tag{14}$$

Taking into account that $h_q\left(\frac{v}{u-(b-2)v}\right) \;=$

$$= \frac{v}{u-(b-2)v}\,\log_q\,\frac{u-(b-2)v}{v} \;+\; \frac{u-(b-1)v}{u-(b-2)v}\,\log_q\,\frac{u-(b-2)v}{u-(b-1)v},$$

one can easily check that (14) can be rewritten in the form

$$\frac{\partial F_b(v,u)}{\partial v} \;=\; -(b-1) \;+\; 3\log_q\frac{1}{v} + \log_q[(1-u)-v] +$$

$$+2(b-1)\log_q[u-(b-1)v] - 2(b-2)\log_q[u-(b-2)v].$$

Therefore, for any fixed u, $0 < u < 1$, equation $\frac{\partial F_b(v,u)}{\partial v} \;=\; 0$ is equivalent to equation

$$\left(\frac{1-u}{v}-1\right)\,\left[\frac{u}{v}-(b-1)\right]^{2(b-1)}\,\left[\frac{u}{v}-(b-2)\right]^{-2(b-2)} = q^{b-1}, \quad \frac{u}{v} \geq b-1. \tag{15}$$

Let $v = v(u)$ be the unique root of (15). This means that function

$$E_b(u) \;=\; F_b(v(u),u) \;=\; -(b-1)v(u) \;+\; (1-u)\,h_q\left(\frac{v(u)}{1-u}\right) \;+$$

$$+\,2\,[u-(b-2)v(u)]\,h_q\left(\frac{v(u)}{u-(b-2)v(u)}\right).$$

If we substitute parameter v for $w \triangleq u/v > b-1$, then equation (15) has the form

$$\left(\frac{1-u}{u}\,w \;-\; 1\right)[w-(b-1)]^{2(b-1)}\,[w-(b-2)]^{-2(b-2)} = q^{b-1}, \qquad w > b-1.$$

Hence, the root $v = v(u)$ can be calculated using the following recurrent method:

$$w_1 \triangleq b, \quad w_{m+1} = (b-1) + \sqrt{q} \left\{ \frac{[w_m - (b-2)]^{2(b-2)}}{\frac{1-u}{u} w_m - 1} \right\}^{\frac{1}{2(b-1)}}, \quad m = 1, 2, \ldots,$$

$$v = v(u) = \frac{u}{\lim\limits_{m \to \infty} w_m}. \tag{16}$$

If $q = 4$, then numerical values of critical distance fractions d_b, $b = 2, 3, \ldots 9$, along with the corresponding optimal parameters

$$v(d_b), \quad 0 \le v(d_b) \le \min \left\{ \frac{d_b}{b-1}, 1 - d_b \right\}, \quad b = 2, 3, \ldots 9,$$

for maximization (11)-(12) are given below:

b	2	3	4	5	6	7	8	9
d_b	0.4792	0.6676	0.7931	0.8768	0.9299	0.9618	0.9798	0.9896
$v(d_b)$	0.1903	0.1166	0.0744	0.0461	0.0272	0.0153	0.0082	0.0043

References

1. D'yachkov, A.G., Macula, A.J., Pogozelski, W.K., Renz, T.E., Rykov, V.V., Torney D.C.: A Weighted Insertion—Deletion Stacked Pair Thermodynamic Metric for DNA Codes. In: Proc. of 10th Int. Workshop on DNA Computing. Milan, Italy, pp. 90–103 (2004)
2. 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 (in Russian) 41(4), 57–77 (2005). English translation: Problems of Information Transmission 41(4), 349–367 (2005)
3. D'yachkov, A.G., Erdos, P.L., Macula, A.J., Rykov, V.V., Torney, D.C., Tung, C.S., Vilenkin, P.A., White, P.S.: Exordium for DNA Codes. J. Comb. Optimization 7(4), 369–379 (2003)
4. Levenshtein, V.I.: Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. Dokl. Akad. Nauk USSR (in Russian) 163, 845–848 (1965), English translation: J. Soviet Phys.–Doklady 10, 707–710 (1966)
5. Levenshtein, V.I.: Efficient Reconstruction of Sequences from Their Subsequences and Supersequences. J. Comb. Th., Ser. A 93, 310–332 (2001)