Eleventh International Workshop on Algebraic and Combinatorial Coding Theory
June 16-22, 2008, Pamporovo, Bulgaria *pp. 86-91*

DNA codes based on stem Hamming similarity

A.G. D'YACHKOV dyachkov@mech.math.msu.su
A.N. VORONINA vorronina@gmail.com
Department of Probability Theory, Faculty of Mechanics and Mathematics,
Moscow State University, Moscow 119992, RUSSIA

Abstract. For q-ary n-sequences, we continue the development [1, 2] of similarity functions that can be used (for $q = 4$) to model a thermodynamic similarity of DNA sequences. Codes based on similarity functions are called DNA codes [1]. In this paper, we discuss a biologically motivated [2] additive similarity function called a *stem Hamming similarity* and defined as the total number of common 2-blocks containing adjacent symbols in the longest common Hamming subsequence between two q-ary n-sequences. Conventional lower and upper bounds called the Gilbert-Varshamov, Plotkin and Elias bounds [3] on the rate of corresponding DNA codes are obtained.

1 Notations and definitions

Symbol $\triangleq$ denotes definitional equalities and 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, $\mathcal{A} \triangleq \{0, 1, \ldots, q-1\}$ is the standard alphabet of size $|\mathcal{A}| = q$ and $\lfloor u \rfloor$ ($\lceil u \rceil$) denotes the largest (smallest) integer $\leq u$ ($\geq u$).

For any letter $x \in \mathcal{A}$, we define $\bar{x} \triangleq (q-1)-x \in \mathcal{A}$, which is called a *complement* of the letter x. For any q-ary n-sequence $\boldsymbol{x} = (x_1, x_2, \ldots, x_{n-1}, x_n) \in \mathcal{A}^n$, we define its *reverse complement* $\widetilde{\overline{\boldsymbol{x}}} \triangleq (\bar{x}_n, \bar{x}_{n-1}, \ldots, \bar{x}_2, \bar{x}_1) \in \mathcal{A}^n$. If $\boldsymbol{y} \triangleq \widetilde{\overline{\boldsymbol{x}}}$, then $\boldsymbol{x} = \widetilde{\overline{\boldsymbol{y}}}$ for any $\boldsymbol{x} \in \mathcal{A}^n$.

Consider two arbitrary q-ary n-sequences

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{A}^n \quad \text{and} \quad \boldsymbol{y} = (y_1, y_2, \ldots, y_n) \in \mathcal{A}^n.$$

The number

$$H_{st}(\boldsymbol{x}, \boldsymbol{y}) \triangleq \sum_{i=1}^{n-1} s_i(\boldsymbol{x}, \boldsymbol{y}), \qquad \text{where}$$

$$s_i(\boldsymbol{x}, \boldsymbol{y}) \triangleq \begin{cases} 1 & \text{if } x_i = y_i,\ x_{i+1} = y_{i+1}, \\ 0 & \text{otherwise,} \end{cases} \quad i = 1, 2 \ldots, n-1, \tag{1}$$

is called a *stem Hamming similarity* between $\boldsymbol{x}$ and $\boldsymbol{y}$. Evidently, $H_{st}(\boldsymbol{x}, \boldsymbol{y})$ can be defined as the total number of common 2-blocks containing adjacent symbols in the longest common Hamming subsequence between sequences $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{A}^n$.

In addition, $0 \le H_{st}(\boldsymbol{x},\boldsymbol{y}) \le n-1$ and $H_{st}(\boldsymbol{x},\boldsymbol{y}) = n-1$ if and only if $\boldsymbol{x} = \boldsymbol{y}$. Therefore, the difference $\mathcal{D}_{st}(\boldsymbol{x},\boldsymbol{y}) \triangleq (n-1) - H_{st}(\boldsymbol{x},\boldsymbol{y}) \ge 0$, $\boldsymbol{x},\boldsymbol{y} \in \mathcal{A}^n$, can be called a *stem Hamming distance* between $\boldsymbol{x}$ and $\boldsymbol{y}$.

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

A code X is called a *DNA* (n,D)*-code* [1] based on the stem Hamming similarity if the following two conditions are fulfilled: (i). For any $j \in [N]$, there exists $j' \in [N]$, $j' \ne j$, such that $\boldsymbol{x}(j') = \widetilde{\boldsymbol{x}(j)} \ne \boldsymbol{x}(j)$. In other words, X is a collection of $N/2$ pairs of mutually reverse complementary sequences. (ii). For any $j \ne j'$, distance $\mathcal{D}_{st}\left(\boldsymbol{x}(j),\boldsymbol{x}(j')\right) \ge D$, i.e., similarity

$$H_{st}\left(\boldsymbol{x}(j),\boldsymbol{x}(j')\right) \le (n-1) - D, \quad j \ne j', \qquad 0 < D < n-1. \tag{2}$$

Let $N_{st}(n,D)$ be the maximal size of DNA (n,D)-codes. If d, $0 < d < 1$, is a fixed number, then

$$R_{st}(d) \triangleq \varlimsup_{n\to\infty} \frac{\log_q N_{st}(n,dn)}{n}, \qquad 0 < d < 1, \tag{3}$$

is called a *rate* of DNA codes based on the stem Hamming similarity.

2 Lower bound on $R_{st}(d)$

Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be *independent identically distributed* random sequences having the *uniform* distribution on $\mathcal{A}^n$. Introduce binary random variables

$$\eta_i \triangleq \begin{cases} 0 & \text{if } x_i = y_i,\ x_{i+1} = y_{i+1}, \\ 1 & \text{otherwise,}\ \ i = 1,2\ldots,n-1 \end{cases} \tag{4}$$

and their sum

$$S_n \triangleq \sum_{i=1}^{n-1} \eta_i \ = \ (n-1) \ - \ H_{st}(\boldsymbol{x},\boldsymbol{y}) \ = \ \mathcal{D}_{st}(\boldsymbol{x},\boldsymbol{y}). \tag{5}$$

Denote by $\overline{\xi}$, the average value of random variable ξ. From definition (4) it follows

$$\overline{\eta_i} = \frac{q^2-1}{q^2}, \qquad \overline{S_n} = \overline{\sum_{i=1}^{n-1} \eta_i} = (n-1)\frac{q^2-1}{q^2}.$$

Let d, $0 < d < \frac{q^2-1}{q^2}$, be a fixed parameter. Introduce function

$$\underline{R}_{st}(d) \triangleq \varliminf_{n\to\infty} \frac{-\log_q \Pr\left\{H_{st}(\boldsymbol{x},\boldsymbol{y}) \ge (1-d)n\right\}}{n} =$$

$$= \overline{\lim_{n\to\infty}} \frac{-\log_q \Pr\{S_n \le d\,n\}}{n}. \tag{6}$$

The random coding method for DNA codes [1] leads to

Proposition 1 *If* $0 < d \le \frac{q^2-1}{q^2}$, *then the rate* $R_{st}(d)$ *of DNA codes based on stem Hamming similarity satisfies inequality* $R_{st}(d) \ge \underline{R}_{st}(d)$, *where* $\underline{R}_{st}(d)$ *is defined by* (6).

Function $\underline{R}_{st}(d)$ is called a *random coding bound* (or the Gilbert-Varshamov bound) on the rate (3) of DNA codes identified by inequality (2).

2.1 Calculation of random coding bound

For the sum (5), introduce the *generating function*

$$G_n(u) \triangleq \sum_{a=0}^{n-1} \Pr\{S_n = a\}\, q^{ua} = \overline{q^{u\,S_n}}, \qquad -\infty < u < \infty, \tag{7}$$

and the *semi-invariant generating function*

$$\mu_n(u) \triangleq \log_q G_n(u), \qquad -\infty < u < \infty. \tag{8}$$

Define independent identically distributed random variables

$$\xi_i \triangleq \begin{cases} 1 & \text{if } x_i = y_i, \\ 0 & \text{otherwise,} \end{cases} \qquad \Pr\{\xi_i = a\} = \begin{cases} \frac{1}{q} & \text{if } a = 1, \\ \frac{q-1}{q} & \text{if } a = 0\ . \end{cases} \tag{9}$$

One can easily see that the vector sequence $\underline{\xi}_i \triangleq (\xi_i, \xi_{i+1})$, $i = 1, \ldots, n-1$, is a stationary Markov chain with transition probabilities:

$$\Pr\left\{ \underline{\xi}_i = (a_1, a_2) | \ \underline{\xi}_{i-1} = (a_3, a_4) \right\} = \begin{cases} \Pr\{\xi_{i+1} = a_2\} & \text{if } a_1 = a_4, \\ 0 & \text{if } a_1 \ne a_4, \end{cases} =$$

$$= \begin{cases} \frac{q-1}{q} & \text{if } a_1 = a_4,\ a_2 = 0, \\ \frac{1}{q} & \text{if } a_1 = a_4,\ a_2 = 1, \\ 0 & \text{if } a_1 \ne a_4. \end{cases} \tag{10}$$

In addition, η_i, $i = 1, \ldots, n-1$, defined by (4) can be written in the form: $\eta_i = f(\underline{\xi}_i) \triangleq 1 - \xi_i \xi_{i+1}$, i.e., the given sequence is a *deterministic* function of Markov chain (10)[1]. Hence, using the standard Markov arguments [4],

[1]Note that η_i, $i = 1, \ldots, n-1$, is not a Markov chain because for any i, $3 \le i \le n-1$, the conditional probability

$$\Pr\{\eta_i = 0 | \eta_{i-1} = 1, \eta_{i-2} = 0\} = 0 \quad \text{and} \quad \Pr\{\eta_i = 0 | \eta_{i-1} = 1\} = \frac{1}{q(q+1)}.$$

pp. 230-232, we can calculate the generating functions (7)-(8) and obtain the following asymptotic ($n \to \infty$) formula:

$$\mu_n(u) \,=\, \log_q G_n(u) = n\,\mu(u) + O(1), \qquad \mu(u) \triangleq \log_q \lambda(u)$$

$$\lambda(u) \triangleq \frac{1}{2q}\left[1 + (q-1)q^u + \sqrt{[1+(q-1)q^u]^2 - 4(q-1)q^u(1-q^u)}\right]. \quad (11)$$

Finally, applying the Large Deviations Principle [5] to S_n, we get

Theorem 1 *Random coding bound* $\underline{R}_{st}(d)$ *defined by* (6) *has the form*

$$\underline{R}_{st}(d) \,=\, L_\mu(d) \,\triangleq\, \max_{u\le 0}\{ud - \mu(u)\}, \quad 0 < d < \frac{q^2-1}{q^2}, \qquad (12)$$

where $L_\mu(d)$, $0 \le d \le \frac{q^2-1}{q^2}$, *is a decreasing* $\bigcup$*-convex function and*

$$L_\mu(0) = 1, \quad L_\mu\left(\frac{q^2-1}{q^2}\right) = 0, \qquad L_\mu(d) > 0, \quad 0 < d < \frac{q^2-1}{q^2}. \qquad (13)$$

3 Upper bounds on $N_{st}(n, D)$ and $R_{st}(d)$

3.1 The Plotkin upper bound on $R_{st}(d)$

A standard upper bound on the rate $R_{st}(d)$ is given by

Proposition 2 . *If* $\frac{q^2-1}{q^2} \le d < 1$, *then* $R_{st}(d) = 0$ *and*

$$R_{st}(d) \,\le\, 1 - \frac{q^2}{q^2-1}\,d \quad \textit{if} \quad 0 < d < \frac{q^2-1}{q^2}. \qquad (14)$$

3.2 On sphere size for stem Hamming similarity

For $q \ge 2$, introduce three recurrent Fibonacci-type sequences [6] of numbers $F_q^1(t)$, $F_q^2(t)$, $F_q^3(t)$, $t = 1, 2, \ldots$, where

$$F_q^i(t) \,\triangleq\, (q-1)F_q^i(t-1) + (q-1)F_q^i(t-2), \quad i = 1,2,3, \quad t \ge 3, \qquad (15)$$

and $F_q^1(1) \triangleq q$, $F_q^1(2) \triangleq q^2 - 1$; $F_q^2(1) \triangleq q-1$, $F_q^2(2) \triangleq (q-1)^2$; $F_q^3(1) \triangleq q-1$, $F_q^3(2) \triangleq q(q-1)$. One can prove, that $F_q^1(t)$ ($F_q^2(t)/F_q^3(t)$) is the number of q-ary sequences $\boldsymbol{x} \in \mathcal{A}^n$ which do not contain 2-stems of the form $(0,0)$ (and do not start and end/do not start or do not end with 0, correspondingly).

Let $\mathbf{t}^{(k)} \triangleq (t_1, t_2, \ldots, t_k)$, $k = 1, 2, \ldots$, denote an ordered collection of k integers. For fixed integers s, $1 \le s \le n-1$, and k, $1 \le k \le \min\left\{s\,;\,\lceil \frac{n-s}{2} \rceil\right\}$, define set

$$T(s,k) \triangleq \Big\{ \mathbf{t}^{(k+1)} \,:\, t_1 \ge 0,\ t_{k+1} \ge 0, \quad t_i \ge 1,\ i = 2, 3, \ldots k, \\ \sum_{i=1}^{k+1} t_i = n - (s+k) \Big\}. \tag{16}$$

Proposition 3 *For any s, $0 \le s \le n-1$, the sphere size $\mathcal{S}_{st}(n,s) \triangleq |\{\boldsymbol{y} : H_{st}(\boldsymbol{x},\boldsymbol{y}) = s\}|$ does not depend on its center $\boldsymbol{x} \in \mathcal{A}^n$. If $s = 0$, then*

$$\mathcal{S}_{st}(n,0) \triangleq |\{\boldsymbol{y} :\ H_{st}(\boldsymbol{x},\boldsymbol{y}) = 0\}| = F_q^1(n). \tag{17}$$

If $1 \le s \le n-1$, then

$$\mathcal{S}_{st}(n,s) = \sum_{k=1}^{\min\left\{s\,;\,\lceil \frac{n-s}{2} \rceil\right\}} \binom{s-1}{k-1} \sum_{T(s,k)} \left\{ F_q^3(t_1) \prod_{i=2}^{k} F_q^2(t_i)\, F_q^3(t_{k+1}) \right\}. \tag{18}$$

For the case $q \ge 2$, Proposition 3 means that the random coding bound $\underline{R}_{st}(d) = L_\mu(d)$, $0 < d < \frac{q^2-1}{q^2}$, (defined by (6) and calculated in Theorem 1) can be also written as

$$\underline{R}_{st}(d) = L_\mu(d) = 1 - \varlimsup_{n\to\infty} \frac{\log_q \mathcal{S}(n, (1-d)n)}{n}, \qquad 0 < d < \frac{q^2-1}{q^2}. \tag{19}$$

3.3 The Elias upper bound on $R_{st}(d)$

The standard Elias arguments [3] and asymptotic formula (19) yield

Theorem 2 *For any d, $0 < d < \frac{q^2-1}{q^2}$, the rate $R_{st}(d) \le U_\mu(d)$, and upper bound $U_\mu(d)$ is presented by parametric equations*

$$U_\mu(d) = u\mu'(u) - \mu(u), \qquad d = \mu'(u)\left[2 - \mu'(u)\frac{q^2}{q^2-1}\right], \quad u \le 0, \tag{20}$$

where function $\mu(u)$, $u \le 0$, is defined in Theorem 1.

Upper bound $U_\mu(d)$ can be called the Elias bound [3]. The given bound improves the Plotkin bound (14) for small values of d, $0 < d < d_q$. We calculated $d_2 \approx 0.60$ and $d_4 \approx 0.13$.

Acknowledgement. The authors are grateful to L. A. Bassalygo, V. M. Blinovsky, B. M. Gurevich and S. A. Pirogov for discussions and valuable info about Large Deviations Principle [5].

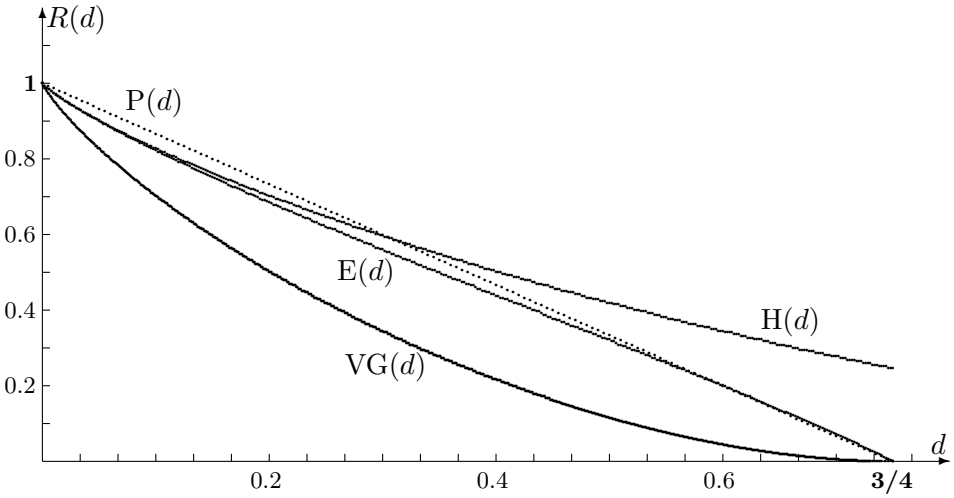

Figure 1: Upper and lower bounds for the rate of DNA-codes for $q = 2$:
H(d) – Hamming bound, P(d) – Plotkin bound,
E(d) – Elias bound, VG(d) – Varshamov-Gilbert bound.

References

[1] 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, *Probl. Pered. Inform.* 41, 2005, 57-77 (in Russian). English transl.: *Probl. Inform. Transm.* 41, 2005, 349-367.

[2] 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 *J. Comput. Biol.* 14, 2007, 1088-1104.

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

[4] V. N. Tutubalin, *The Theory of Probability and Random Processes*, Moscow: Publishing House of Moscow State University, 1992 (in Russian).

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

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