COMPLEXITY METRICS AND EXPLORATORY ANALYSIS OF DNA SEQUENCES

T. Rekašius
Vilnius Gediminas Technical University
Vilnius, LITHUANIA
e-mail: tomas.rekasius@fm.vtu.lt

Abstract

A new measure of the distance between nucleotide sequences and the efficient method for visualisation of nucleotide sequences have been proposed. It facilitates visualisation of a long DNA sequence via a compact and smooth picture of all oligonucleotides of length up to 10 in the sequence.

1 Introduction

Let's define the fixed length sequence x of n symbols as follows:

$$x = x_1 x_2, \ldots, x_n, \quad x_l \in \mathcal{A}, \quad l = \overline{1, n}, \tag{1}$$

where $\mathcal{A}$ is a finite set (alphabet). DNA sequences consists of four nucleotides briefly called A, C, G, T, therefore for DNA sequences $\mathcal{A} = \{\text{A,C,G,T}\}$. Let us identify them with vertices of a square with coordinates $(0,0)$, $(1,1)$, $(0,1)$, and $(1,0)$, respectively. Thus, we have a natural isomorphism $\nu : \{\text{A,C,G,T}\} \to \mathcal{A} \times \mathcal{A}$, $\mathcal{A} = \{0, 1\}$, where

$$\{\nu(\text{A}), \nu(\text{C}), \nu(\text{G}), \nu(\text{T})\} = \{(0,0), (1,1), (0,1), (1,0)\}. \tag{2}$$

Chaos game representation (CGR) of DNA algorithm: a) recode DNA sequence into two binary sequences of the same length (2) and identify each of them with a fractional dyadic number, b) sequentially multiply these two numbers by 2 and take their fractional parts as a new pair of numbers, c) plot the pairs obtained on the graph. Resulting picture is called CGR "genome signature" (fig. 1).

"Genome signature" is an efficient way to picture long nucleotide sequences [2], but it has several undesirable features, and one of them is their fractality. For example the difference between dyadic representation in the interval [0,1) of two sequences 10000000000 and 01111111111 is less than 2^{-10} whereas the difference for "similar" sequences 00000000001 and 11111111110 is greater than $1 - 2^{-9}$. Due to the fractal character of DNA signatures the easily comprehensible Euclidean distance does not represent genetical similarity (dissimilarity) of oligonucleotides appropriately and hence the differences between them are difficult to interpret. To make "genome signature" to be appropriate for DNA visualisation the "natural" identification of nucleotide sequence with a (dyadic) point in the unit square should be replaced by a more subtle mapping. This mapping must be continuous in the sense that (genetically) "close" nucleotide sequences are represented by close points in the square and vice versa.

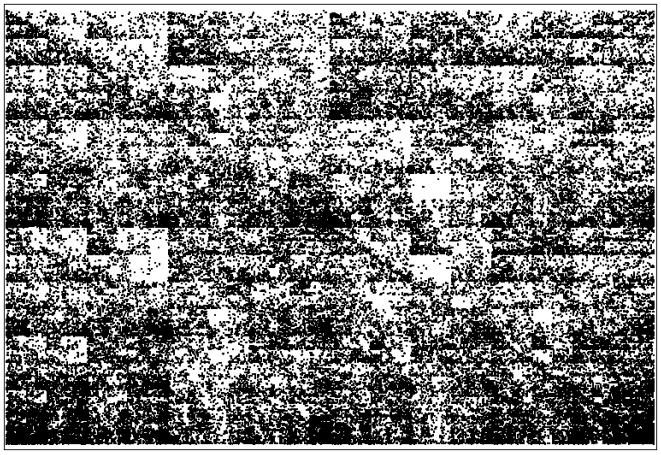

Figure 1: CGR genome signature. Bacteria *Helicobacter pylori J99*

In bioinformatics, different genetic sequences are being compared rather frequently [4]. Relationship of two organisms, when comparing their DNA, also gets to the calculation of the distance between two genomes. With such distances between the species known, phylogenetic trees could be (re)constructed, or the origin of species could be analysed.

Any finite sequence with elements from a finite alphabet can be identified with a rational number from the interval [0,1). In turn, any sequence of real numbers no matter how long it may be and any vector of a very large dimension can be treated as a function, i.e., merely as a point in a functional space. This is the paradigm of the functional data analysis [3]. In this work we attempt to apply this approach to genetic (nucleotide) sequence visualisation and analysis.

2 Distance between binary sequences

The distance proposed below is based on a operator of "differentiation" [1]. In some sense it is a discrete analog of Sobolev norm which is well known in the functional analysis. Assume that the sequence x defined in (1) is binary, i.e. $\mathcal{A} = \{0, 1\}$. In the sequel x is identified with the corresponding vector in the space $\mathbf{R}^n$. The difference ("differentiation") operator $\mathcal{B}$ is defined in the following way. Let:

$$\mathcal{M}_n \overset{\mathcal{B}_1^{(n)}}{\to} \mathcal{M}_{n-1} \overset{\mathcal{B}_1^{(n-1)}}{\to} \mathcal{M}_{n-2} \overset{\mathcal{B}_1^{(n-2)}}{\to} \dots \overset{\mathcal{B}_1^{(3)}}{\to} \mathcal{M}_2 \overset{\mathcal{B}_1^{(2)}}{\to} \mathcal{M}_1,\ \mathcal{M}_i \subset \mathbf{R}^i,\ i = \overline{1, n}. \tag{3}$$

Here the operator $\mathcal{B}_1^{(k)}$, $k \in \{2, \dots, n\}$ is expressed by the formula

$$\mathcal{B}_1^{(k)} x = \{(x_{i+1} - x_i)/2,\ i = \overline{1, k-1}\}, \tag{4}$$

and the operators $\mathcal{B}_l^{(n)} : \mathcal{M}_n \to \mathcal{M}_{n-l}$ are obtained recurrently from the formula

$$\mathcal{B}_l^{(n)} = \mathcal{B}_1^{(n-l+1)} \mathcal{B}_{l-1}^{(n)},\ l = \overline{2, n-1}. \tag{5}$$

The upper index n of the operator $\mathcal{B}_l^{(n)}$ indicates the dimension of the space it acts in, while the lower index l shows the extent to which the dimension of its mapping is

smaller. For a given $x \in \mathcal{M}_n \subset \mathbf{R}^n$, coordinates of the corresponding point $y = \mathcal{B}x$ in the space $\mathbf{R}^{n(n+1)/2}$ are expressed by the formula

$$y = y(x) = \mathcal{B}x = (x,\, \mathcal{B}_1^{(n)}x,\, \mathcal{B}_2^{(n)}x, \ldots, \mathcal{B}_{n-1}^{(n)}x). \tag{6}$$

Let a positive defined diagonal matrix $W = \{w_1, \ldots, w_q\}$ of the order $q = n(n+1)/2$ be given. Define the weighted inner product in $\mathbf{R}^q$ by the equality

$$(u, v)_W := u^\top W v = \sum_{i=1}^{q} w_i u_i v_i \tag{7}$$

and set $|u|_W = \sqrt{(u,u)_W}$. The distance $d = d_W$ between two binary sequences x and z is defined as

$$d(x, z) = |y(x) - y(z)|_W, \quad x, z \in \mathcal{M}_n, \quad y(x), y(z) \in \mathcal{M} \subset \mathbf{R}^q. \tag{8}$$

Define the cyclic shift operator $T:\ \mathcal{M}_n \to \mathcal{M}_n$ by the equality

$$T(x) = x_n x_1 \ldots x_{n-1}, \quad x \in \mathcal{M}_n. \tag{9}$$

For a given positive sequence $\rho_0, \ldots, \rho_{n-1}$ define average (smoothed) distance

$$d_\rho(x, z) = \sum_{j=0}^{n-1} \rho_j \cdot \big(T^j(x), T^j(z)\big). \tag{10}$$

Given a matrix $\mathbf{D} = \{d(x,z)\}$ or matrix $\mathbf{D}_\rho = \{d_\rho(x,z)\}$, $\forall x, z \in \mathcal{M}_n$ consisting of the pairwise dissimilarities of the sequences x (in the space $\mathbf{R}^q$) the goal is to reduce the dimensionality q of the data set to a sufficiently small value so that the distances between the sequences x in the low dimension space would be as close to the original distances as possible. It is a classical problem of multidimensional scaling. To solve it, here we apply *SAS* procedure MDS. Results for $n = 5$ long binary sequences x are presented in table 1.

Table 1: One-dimensional projections $\varphi(x)$ of $n = 5$ long sequences x. Distance $d(x, z)$

No	x	$\varphi(x)$	No	x	$\varphi(x)$	No	x	$\varphi(x)$	No	x	$\varphi(x)$
1	10101	0.0000	9	10110	0.2588	17	00000	0.5083	25	10010	0.7755
2	00101	0.0403	10	11100	0.2969	18	01111	0.5523	26	11000	0.7799
3	10100	0.0623	11	10001	0.3307	19	11110	0.5743	27	00010	0.8224
4	00100	0.1033	12	00110	0.3550	20	11001	0.6193	28	01000	0.8481
5	11101	0.1519	13	01100	0.3807	21	10011	0.6450	29	11011	0.8967
6	10111	0.1776	14	00001	0.4257	22	01110	0.6693	30	01011	0.9377
7	01101	0.1988	15	10000	0.4477	23	01001	0.7155	31	11010	0.9597
8	00111	0.2458	16	11111	0.4917	24	00011	0.7288	32	01010	1.0000

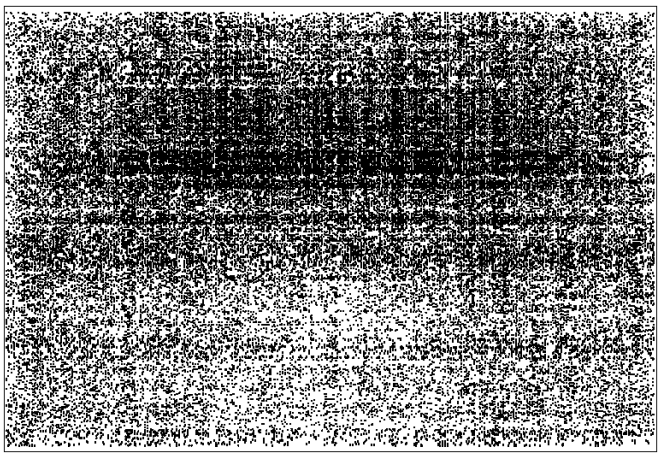

Figure 2: Modified genome signature. Bacteria *Helicobacter pylori J99*

"Genome signature". DNA sequence $S = s_1 s_2 \dots s_m = \{s_i, s_i \in \mathcal{A}, i = \overline{1,m}\}$, $\mathcal{A} = \{\text{A,C,G,T}\}$ can be expressed in an equivalent form as two-dimensional binary sequence (2). Let $\varphi(x)$ be an one-dimensional projection of x (table 1). Attributing the coordinate $\varphi(x(j))$ to the moving binary sequence $x(j) = s_j s_{j+1} \dots s_{j+n-1}$, $j = \overline{1, m-n+1}$ of the length n for the entire DNA we obtain a set of two-dimensional points. This set is called "genome signature".

The signature of the sequence of 10^5 nucleotides is drawn. Differently from traditional "genome signature" obtained by CGR, the patterns obtained are rather smooth, the set of points is not divided into sub-squares and fractality is not so evident (fig. 2).

Using bacterial DNA data from *GenBank*, we have discovered some characteristic patterns. As it could be expected, related bacteria gave similar genome signature. It is clear that such genome signature patterns depend also on one-dimensional code distribution in the sequence. Taking only one genome signature coordinate a DNA sequence can be analysed by one of two characteristics of nucleotides.

Bibliography

[1] Arnold V.I. (2005). Slozhnost konechnyh posledovatelnostei nulei i edinic i geometrija konechnyh funkcionalnyh prostranstv. http://mms.mathnet.ru/meetings/2005/arnold.pdf (in Russian).

[2] Jeffrey H.J. (1990). Chaos Game Representation of Gene Structure. *Nucleic Acids Res.* Vol. **18**, pp. 2163-2170.

[3] Ramsay J.O., Silverman B.W. (1997). *Functional Data Analysis*. Springer-Verlag, New York.

[4] Waterman M.S. (1995). *Introduction to Computational Biology*. Chapman & Hall, London.