Some properties for DNA curve

2004 J. Phys. A: Math. Gen. 37 7135

(http://iopscience.iop.org/0305-4470/37/28/005)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS A: MATHEMATICAL AND GENERAL

J. Phys. A: Math. Gen. **37** (2004) 7135–7142 PII: S0305-4470(04)78974-X

Some properties for DNA curve

Ping-an He[1] **and Jun Wang**[2,3]

[1] T-Life Research Center, Fudan University, Shanghai 200433, People's Republic of China
[2] Department of Applied Mathematics, Dalian University of Technology, Dalian 116024, People's Republic of China
[3] College of Advanced Science and Technology, Dalian University of Technology, Dalian 116024, People's Republic of China

E-mail: pinganhe@yahoo.com.cn

Received 8 April 2004, in final form 1 June 2004
Published 30 June 2004
Online at stacks.iop.org/JPhysA/37/7135
doi:10.1088/0305-4470/37/28/005

Abstract

A DNA sequence can be identified with a word over the alphabet $\Sigma = \{A, C, G, T\}$. An algebraic method is used to analyse DNA sequences and their three-dimensional vector representation, and, via their geometric representation, an equivalence relation on DNA sequences is introduced, and the number of equivalence classes of sequences is counted. Finally, a kind of inequality involving equivalent sequences' entropy is proved.

PACS numbers: 87.14.Gg, 02.20.−a

1. Introduction

We are living in the era of the explosion of biological information. With the development of the sequencing technique and the genome projects, the known sequences are accumulated at an exponential rate with respect to time. Nucleic acids and proteins are all linear macromolecules and thus can be expressed as linear sequences. A nucleic acid sequence is regarded as a string over four bases (letters), adenine (A), cytosine (C), guanine (G) and thymine (T). Thus, a nucleic acid sequence can be considered as a word w over the alphabet $\Sigma = \{A, C, G, T$ (or $U)\}$. This expression is called a letter sequence representation (LSR) or a DNA primary sequence. A DNA primary sequence conserves its basic hereditary information. Therefore, the statistics and analysis for DNA primary sequences are important in bioinformatics.

Mathematical analysis of large volume genomic DNA sequence data is one of the challenges for bio-scientists. In order to give a visual characterization of DNA sequences, many attempts have been made [3, 4, 6–8, 12]. Graphical representation of a DNA sequence provides a simple way of viewing, sorting and comparing various gene structures. For example, Zhang [9] proposed a three-dimensional vector representation (3DVR) for a DNA primary sequence, which is called a Z-curve. This representation is very useful in studies of nucleotide

distribution and composition, especially in comparative studies of similarity/dissimilarity of DNA sequences [4, 6–8, 12]. In a series of works of Zhang *et al* [9–12], some properties of the Z-curve are investigated.

Using algebraic methods to analyse DNA sequences has been a subject of research in the past [1, 2, 5]. In this paper, we attempt to treat the Z-curves from algebraic and combinatorial points of view based on Zhang's works.

The distribution of this paper is as follows. In section 2, we introduce some preliminary knowledge relating to the rest of this paper. Then, in section 3, we give some operations on Z-curves, and obtain some properties of the DNA curve using group $\mathcal{S}_4$ acting on the DNA curve. In section 4, we define an equivalence relation to the Z-curves, and count the number of the equivalence classes of DNA sequences. In addition, we prove a kind of inequality which relates to the equivalent sequences' entropy.

2. Preliminary

In DNA sequences, the four bases A, C, G, T can be divided into two classes according to their chemical structures, i.e.,

$$\text{Bases}\begin{cases}\text{purine } R = A, G\\ \text{pyrimidine } Y = C, T.\end{cases}$$

The bases can be also divided into another two classes,

$$\text{Bases}\begin{cases}\text{amino group } M = A, C\\ \text{keto group } \ K = G, T.\end{cases}$$

In addition, the division can be made according to the strength of the hydrogen bond, i.e.,

$$\text{Bases}\begin{cases}\text{strong H-bonds } S = G, C\\ \text{weak H-bonds } W = A, T.\end{cases}$$

By the above classifications, a DNA primary sequence can be embedded into the three-dimensional space as follows: assigning the six classes (purine, pyrimidine, amino group, keto group, weak H-bond and strong H-bond) to the six directions in the coordinate system *O-XYZ* associated with the positive and the negative x, y and z axes, such that purine, amino group and weak H-bond correspond to the positive x, y and z axes, respectively, and pyrimidine, keto group and strong H-bond correspond to the negative x, y and z axes, respectively.

For any positive integer n, let $\mathcal{D}_n$ denote the set of all DNA sequences of length n and write $\mathcal{D} = \bigcup_{n\geqslant 1}\mathcal{D}_n$.

Given a $w \in \mathcal{D}$, by $|w|$ we denote its length, that is the number of the letters in w. Suppose $|w| = n$, i.e., $w \in \mathcal{D}_n$. For $1 \leqslant i \leqslant n$ and $L \in \Sigma$, let $L_i(w)$ (or simply L_i, for short) denote the number of L occurring in the first i letters of w. By definition we have $A_i + C_i + G_i + T_i = i$ for $i = 1, 2, \ldots, |w|$.

Now, we inspect w from $5'$ to $3'$ by stepping one base at a time. We start from the original point O. Then, at every step, say the ith step, we obtain a point $P_i(w)$ (or P_i for short) in the coordinate system according to the number of six classes of bases. We thus obtain n points $P_0 = O, P_1, P_2, \ldots, P_n$ in the three-dimensional real space. The Z-curve is an appropriate connection of them one by one. In [10], the coordinates of P_i $(i = 1, 2, \ldots, n)$ are expressed by A_i, C_i, G_i and T_i as follows:

$$\begin{cases}x_i = 2(A_i + G_i) - i\\ y_i = 2(A_i + C_i) - i\\ z_i = 2(A_i + T_i) - i\end{cases} \tag{1}$$

or equivalently,

$$\begin{pmatrix} A_i \\ C_i \\ G_i \\ T_i \end{pmatrix} = \frac{i}{4}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \frac{1}{4}\begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \\ -1 & -1 & 1 \end{pmatrix}\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} \tag{2}$$

where $x_i, y_i, z_i \in [-i, i]$ and $i \in \{0, 1, \ldots, n\}$. Recall that $A_i + C_i + G_i + T_i = i$. So $x_i = 2(A_i + G_i) - i = (A_i + G_i) - (C_i + T_i)$, which is the difference of numbers of purine and pyrimidine after the ith step inspection. y_i and z_i have similar interpretations.

Thus, from a DNA sequence $w \in \mathcal{D}_n$ we obtain a $3 \times n$ matrix of the form

$$\Pi(w) = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \\ z_1 & z_2 & \cdots & z_n \end{pmatrix}$$

which is called a DNA sequence matrix.

Let $\mathcal{M}_n$ denote the set of all $3 \times n$ DNA sequence matrices and write $\mathcal{M} = \bigcup_{n \geqslant 1} \mathcal{M}_n$. It is easy to see that Π is a one-to-one correspondence between $\mathcal{D}$ and $\mathcal{M}$.

For any $v \in \mathcal{D}_n$ and $w \in \mathcal{D}_m$ it is easy to see that $vw \in \mathcal{D}_{n+m}$. Similarly, for $A \in \mathcal{M}_n$ and $B \in \mathcal{M}_m$ we define

$$A * B = (A, B')$$

where (A, B') is obtained by concatenating matrix A and matrix B' with B' obtained by applying formula (3) for matrix B, as follows

$$\begin{cases} x_i' = x_i + x_A \\ y_i' = y_i + y_A \\ z_i' = z_i + z_A \end{cases} \tag{3}$$

where (x_i', y_i', z_i') and (x_i, y_i, z_i) are the ith columns of B' and B, respectively, and (x_A, y_A, z_A) is the coordinate of termination of the Z-curve of A.

Proposition 2.1. *Let v and w be in $\mathcal{D}$. Then*

$$\Pi(vw) = \Pi(v) * \Pi(w).$$

Proof. Let $|v| = n_1$ and $|w| = n_2$. Then $|vw| = n_1 + n_2$.

$$\text{Set } \Pi(v) = \begin{pmatrix} x_i' \\ y_i' \\ z_i' \end{pmatrix}_{1 \leqslant i \leqslant n_1} \qquad \Pi(w) = \begin{pmatrix} x_i'' \\ y_i'' \\ z_i'' \end{pmatrix}_{1 \leqslant i \leqslant n_2} \qquad \Pi(vw) = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix}_{1 \leqslant i \leqslant n_1 + n_2}.$$

For $i \in [1, n_1]$, $x_i = 2(A_i(v) + G_i(v)) - i$, so $x_i = x_i'$. For $i \in [n_1 + 1, n_1 + n_2]$, set $i = n_1 + k$. Then $x_i = 2(A_i(vw) + G_i(vw)) - i = 2\big(A_{n_1}(v) + G_{n_1}(v)\big) + 2(A_k(w) + G_k(w)) - n_1 - k = x_{n_1} + x_k''$. y_i, z_i have similar explanations. From (3), we have $\Pi(vw) = \Pi(v) * \Pi(w)$. □

From (1) and (2) we also see an interesting property: the end point of the Z-curve of a palindromic sequence must be on the Z-axis. In fact, since the number of base A (or C) is equal to the number of base T (or G) in a palindromic sequence, we have $A_n = T_n$ and $C_n = G_n$. By (1) and the equation $A_n + C_n + G_n + T_n = n$, we obtain $x_n = y_n = 0$.

3. Action of the symmetric group $\mathcal{S}_4$ on the Z-curves and the DNA matrices

Let S_4 be the symmetric group on the letters A, C, G and T. Denoting the elements by products of disjoint cycles, we have $S_4 = \{I, (AC)(GT), (AG)(CT), (AT)(CG), (ACG), (GCA), (ACT), (TCA), (AGT), (TGA), (CGT), (TGC), (AC), (AG), (AT), (CG), (CT), (GT), (ACGT), (ACTG), (AGCT), (AGTC), (ATCG), (ATGC)\}$. Then, S_4 acts on DNA primary sequences in a normal way: given an $\alpha \in S_4$ and a DNA sequence $v = v_1v_2v_3\ldots$, define $\alpha(v) = \alpha(v_1)\alpha(v_2)\alpha(v_3)\cdots$.

In order to have a synchronous action on the DNA matrices, we define an isomorphic embedding ϕ of S_4 into $O(3)$, the group of the orthogonal matrices of order three.

Since S_4 is generated by (AC), (AG) and (AT), we define

$$\phi(AC) = \begin{pmatrix} 0 & 0 & -1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{pmatrix} \qquad \phi(AG) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} \qquad \phi(AT) = \begin{pmatrix} 0 & -1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

It is easy to see that $(\phi(AC))^2 = (\phi(AG))^2 = (\phi(AT))^2 = I_3$, the identity matrix, and their determinants are all equal to -1; therefore, they represent reflection transformations in the three-dimensional real space. So, they generate a group homomorphism from S_4 into $O(3)$. It is easy to verify that the kernel of ϕ consists of I, that is, ϕ is an injective homomorphism, i.e., an isomorphic embedding.

For an $\alpha \in S_4$ and a DNA matrix $A \in \mathcal{M}$ define $\alpha(A) = \phi(\alpha)A$. Then S_4 acts on the set $\mathcal{M}$ satisfying the following properties.

Proposition 3.1. *Let $\alpha \in \mathcal{S}_4$ and $P, Q \in \mathcal{M}$. Then*

(1) $\alpha\Pi = \Pi\alpha$, and
(2) $\alpha(P * Q) = \alpha(P) * \alpha(Q)$.

Proof. (1) Clearly, it suffices to verify the equation for $\alpha = (AC), (AG)$ and (AT). Let $v \in \mathcal{D}_n$. Write

$$\Pi(v) = \begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix}_{1\leqslant i\leqslant n} = \begin{pmatrix} 2(A_i + G_i) - i \\ 2(A_i + C_i) - i \\ 2(A_i + T_i) - i \end{pmatrix}_{1\leqslant i\leqslant n}.$$

Recall that $A_i + C_i + G_i + T_i = i$. We have

$$\phi(AC)\Pi(v) = \begin{pmatrix} -z_i \\ y_i \\ -x_i \end{pmatrix}_{1\leqslant i\leqslant n} = \begin{pmatrix} i - 2(A_i + T_i) \\ 2(A_i + C_i) - i \\ i - 2(A_i + G_i) \end{pmatrix}_{1\leqslant i\leqslant n} = \begin{pmatrix} 2(C_i + G_i) - i \\ 2(C_i + A_i) - i \\ 2(C_i + T_i) - i \end{pmatrix}_{1\leqslant i\leqslant n}$$

$$\phi(AG)\Pi(v) = \begin{pmatrix} x_i \\ -z_i \\ -y_i \end{pmatrix}_{1\leqslant i\leqslant n} = \begin{pmatrix} 2(A_i + G_i) - i \\ i - 2(A_i + T_i) \\ i - 2(A_i + C_i) \end{pmatrix}_{1\leqslant i\leqslant n} = \begin{pmatrix} 2(G_i + A_i) - i \\ 2(G_i + C_i) - i \\ 2(G_i + T_i) - i \end{pmatrix}_{1\leqslant i\leqslant n}$$

and

$$\phi(AT)\Pi(v) = \begin{pmatrix} -y_i \\ -x_i \\ z_i \end{pmatrix}_{1\leqslant i\leqslant n} = \begin{pmatrix} i - 2(A_i + C_i) \\ i - 2(A_i + G_i) \\ 2(A_i + T_i) - i \end{pmatrix}_{1\leqslant i\leqslant n} = \begin{pmatrix} 2(T_i + G_i) - i \\ 2(T_i + C_i) - i \\ 2(T_i + A_i) - i \end{pmatrix}_{1\leqslant i\leqslant n}.$$

We thus obtain that

$$\phi(\alpha)\Pi(v) = \begin{pmatrix} 2(\alpha(A)_i + \alpha(G)_i) - i \\ 2(\alpha(A)_i + \alpha(C)_i) - i \\ 2(\alpha(A)_i + \alpha(T)_i) - i \end{pmatrix}_{1 \leqslant i \leqslant n} = \Pi(\alpha(v))$$

holds for every $\alpha \in S_4$, which yields the result.

(2) By condition we have that there are $v, w \in \mathcal{D}$ such that $\Pi(v) = P$ and $\Pi(w) = Q$, and $\alpha(vw) = \alpha(v)\alpha(w)$, which implies $\Pi(\alpha(vw)) = \Pi(\alpha(v)\alpha(w))$. Then the result follows from proposition 2.1 and (1). □

Given a $v \in \mathcal{D}_n$ let $p_i(v)$ denote the content of the bases A, C, G, T in the sequence, respectively, that is, $p_1(v) = A_n/n$, $p_2(v) = C_n/n$, $p_3(v) = G_n/n$ and $p_4(v) = T_n/n$.

Definition 3.2. *Let $v \in \mathcal{D}$. The information entropy of v is defined to be*

$$H(v) = -\sum_{i=1}^{4} p_i(v) \log_2 p_i(v).$$

It is well known that H achieves its maximum value $\ln 4$ when $p_1 = p_2 = p_3 = p_4$. By definition we see that H is a symmetric real function of A, C, G and T on $\mathcal{D}$. We then immediately have the following result.

Proposition 3.3. *H is an invariant quantity of the DNA primary sequence under the action of $\mathcal{S}_4$.*

The information entropy H of the DNA sequence shows a kind of information in the DNA sequence. From the information theory point of view we know that the information of the DNA sequence is an invariant quantity in $\mathcal{S}_4$ acting on this sequence. Thus, each permutation in $\mathcal{S}_4$ may be interpreted as an operator in which the nucleotide text can be permuted to another nucleotide text without loss of information. We can seek information through the $\mathcal{S}_4$ group acting on this sequence by proposition 3.3.

4. Equivalence classes

We first introduce a relation on $\mathcal{D}$.

Definition 4.1. *Let v and w be two DNA sequences. Write $v \sim w$ if the terminal points of the Z-curves of v and w are identical. Precisely, $v \sim w$ if and only if $P_m(v) = P_n(w)$, where $m = |v|$ and $n = |w|$.*

Obviously, '$\sim$' is an equivalence relation on $\mathcal{D}$. By $[w]$ we denote the set of all $v \in \mathcal{D}$ such that $v \sim w$ and $|v| \leqslant |w|$.

Let us discuss the conditions that two DNA sequences are equivalent.

In the rest of this section, let w be a selected sequence in $\mathcal{D}_n$ and v an arbitrary DNA sequence in $\mathcal{D}_m$ with $m \leqslant n$.

First, suppose $m = n$. From (1) and (2) we have that $v \sim w$ if and only if $A_n(v) = A_n(w)$, $C_n(v) = C_n(w)$, $G_n(v) = G_n(w)$ and $T_n(v) = T_n(w)$. We thus immediately obtain the following enumerative result.

Theorem 4.2. *Suppose $w \in \mathcal{D}_n$. Then the cardinality of the set $[w] \cap \mathcal{D}_n$ equals*

$$\binom{n}{A_n, C_n, G_n, T_n} = \frac{n!}{A_n!C_n!G_n!T_n!}.$$

Next, suppose $m < n$. Set $X_1 = A_n(w) - A_m(v)$, $X_2 = C_n(w) - C_m(v)$, $X_3 = G_n(w) - G_m(v)$ and $X_4 = T_n(w) - T_m(v)$. From (1) it follows that $v \sim w$ if and only if

$$\begin{cases} X_1 - X_2 + X_3 - X_4 = 0 \\ X_1 + X_2 - X_3 - X_4 = 0 \\ X_1 - X_2 - X_3 + X_4 = 0 \end{cases} \tag{4}$$

which implies that $X_1 = X_2 = X_3 = X_4$. Writing it as k, we have that $A_n(w) = A_m(v) + k$, $C_n(w) = C_m(v) + k$, $G_n(w) = G_m(v) + k$, $T_n(w) = T_m(v) + k$, and $n = m + 4k$. We thus obtain

Theorem 4.3. *Suppose $w \in \mathcal{D}_n$. Then the cardinality of the set $[w]$ equals*

$$\sum_{k=0}^{r} \binom{n-4k}{A_n - k, C_n - k, G_n - k, T_n - k}$$

where $r = \min\{A_n, C_n, G_n, T_n\}$.

Definition 4.4. *Let $w \in \mathcal{D}_n$ and $r = \min\{A_n, C_n, G_n, T_n\}$. Let $K(w)$ be the set of all sequences obtained by eliminating r A, r C, r G and r T in the sequence w. The sequences in $K(w)$ are called kernels of w. We call w a complete degeneracy sequence if $K(w)$ consists of the empty sequence, i.e., $A_n(w) = C_n(w) = G_n(w) = T_n(w)$.*

Obviously, $K(w) \subseteq [w]$ and its elements are of the least length in $[w]$. But the converse is not generally true. For example, let $w = AACCGGT$. Then $v = CGA \in [w]$ but $v \notin K(w)$.

We now suppose $v \in [w]$ and compare their entropies. From the above discussion we know that $A_n(w) = A_m(v) + k$, $C_n(w) = C_m(v) + k$, $G_n(w) = G_m(v) + k$, $T_n(w) = T_m(v) + k$ and $n = m + 4k$. From definition 3.2, we have

$$\begin{aligned} H(w) = -\Bigg(& \frac{A_m + k}{m + 4k} \ln \frac{A_m + k}{m + 4k} + \frac{C_m + k}{m + 4k} \ln \frac{C_m + k}{m + 4k} + \frac{G_m + k}{m + 4k} \ln \frac{G_m + k}{m + 4k} \\ & + \frac{T_m + k}{m + 4k} \ln \frac{T_m + k}{m + 4k} \Bigg) \end{aligned}$$

$$H(v) = -\left(\frac{A_m}{m} \ln \frac{A_m}{m} + \frac{C_m}{m} \ln \frac{C_m}{m} + \frac{G_m}{m} \ln \frac{G_m}{m} + \frac{T_m}{m} \ln \frac{T_m}{m} \right).$$

Theorem 4.5. *Suppose $v \in [w]$. Then $H(w) \geqslant H(v)$, and the equality holds if and only if $|v| = |w|$ or w is a complete degeneracy sequence.*

This theorem has a direct corollary as follows.

Corollary 4.6. *(1) All complete degeneracy sequences are of the largest entropy* $\ln 4$.

(2) If w is not a complete degeneracy sequence, then in $[w]$, the kernels of w have the minimal entropy, and, the longer the DNA sequence, the larger the entropy.

To prove theorem 4.5 we provide a more general result, stated as a lemma.

Lemma 4.7. *Let $a_1 \geqslant a_2 \geqslant \cdots \geqslant a_n \geqslant 0$ be arbitrary real numbers and write $a = a_1 + a_2 + \cdots + a_n$. Define*

$$f(x) = -\sum_{i=1}^{n} \frac{a_i + x}{a + nx} \ln \frac{a_i + x}{a + nx}.$$

Then $f(x) \equiv \ln n$ if $a_1 = a_2 = \cdots = a_n$. Otherwise, $f(x)$ is a strictly increasing function in $[-a_n, \infty)$.

Proof. Note that $\lim_{x\to+0} x\ln x = 0$. Therefore, $f(x)$ is continuous in $[-a_n,\infty)$ and differentiable in $(-a_n,\infty)$. We have

$$\begin{aligned}
D[f(x)] &= \sum_{i=1}^{n} \frac{\left(1+\ln\frac{a_i+x}{a+nx}\right)(na_i-a)}{(a+nx)^2} \\
&= \sum_{i=1}^{n} \frac{na_i-a}{(a+nx)^2} + \sum_{i=1}^{n} \frac{(na_i-a)\ln\frac{a_i+x}{a+nx}}{(a+nx)^2} \\
&= \frac{1}{(a+nx)^2}\sum_{i=1}^{n}(na_i-a)\ln\frac{a_i+x}{a+nx} \\
&= \frac{1}{(a+nx)^2}\sum_{k=1}^{n}\left((n-k)\sum_{i=1}^{k}a_i - k\sum_{j=k+1}^{n}a_j\right)\left(\ln\frac{a_k+x}{a+nx} - \ln\frac{a_{k+1}+x}{a+nx}\right).
\end{aligned}$$

Since $n-k \geqslant 0$ and

$$(n-k)\sum_{i=1}^{k}a_i - k\sum_{j=k+1}^{n}a_j \geqslant (n-k)ka_k - k(n-k)a_{k+1} \geqslant k(n-k)(a_k-a_{k+1}) \geqslant 0$$

it follows $D[f(x)] \geqslant 0$. The equality holds iff $a_1 = a_2 = \cdots = a_n$. This completes the proof. □

From this lemma theorem 4.5 follows by taking $n = 4$ and $x = k$.

We end this section with an inequality.

Theorem 4.8. *With the notation in lemma 4.7 we have*

$$\frac{\left(\prod_{i=1}^{n} a_i^{a_i}\right)^{a^{-1}}}{\left(\prod_{i=1}^{n}(a_i+k)^{(a_i+k)}\right)^{(a+nk)^{-1}}} \geqslant \frac{a}{a+nk} \tag{5}$$

and the equality holds iff $a_1 = a_2 = \cdots = a_n$ *or* $k = 0$.

Proof. Take $x = k$ in lemma 4.7. Then

$$-\sum_{i=1}^{n}\frac{a_i+k}{\sum_{i=1}^{n}(a_i+nk)}\ln\frac{a_i+k}{a+nk} \geqslant -\sum_{i=1}^{n}\frac{a_i}{a}\ln\frac{a_i}{a}$$

which is equivalent to (5). □

In particular, taking $k = 1$ and $n = 2, 3$, we obtain the following inequalities:

$$\frac{(a^a b^b)^{\frac{1}{a+b}}}{((a+1)^{(a+1)}(b+1)^{(b+1)})^{\frac{1}{a+b+2}}} \geqslant \frac{a+b}{a+b+2}$$

and

$$\frac{(a^a b^b c^c)^{\frac{1}{a+b+c}}}{((a+1)^{(a+1)}(b+1)^{(b+1)}(c+1)^{(c+1)})^{\frac{1}{a+b+c+3}}} \geqslant \frac{a+b+c}{a+b+c+3}.$$

5. Conclusions

DNA sequences and their 3D graphical representation are analysed from algebraic and combinatorial points of view. By observation of the terminal points of the Z-curves, an equivalence relation on DNA sequences is introduced, and the number of equivalence classes of sequences is counted, and the information entropies of their equivalence classes are compared. All their results are mathematical. We wish to find their applications in biology.

Acknowledgments

The authors would like to thank the anonymous referees for many valuable suggestions that have improved this manuscript. This work is supported in part by the National Natural Science Foundation of China and Shanghai Postdoctoral Science Foundation.

References

[1] Mac Donaill D A and Buttimore N H 1996 *Comput. Math. Appl.* **32** 29
[2] Mac Donaill D A and Buttimore N H 1996 *Comput. Math. Appl.* **32** 39
[3] Guo X F, Randic M and Basak S C 2001 *Chem. Phys. Lett.* **350** 106
[4] Li C and Wang J 2004 *Comb. Chem. High Throughput Screen.* **7** 23
[5] Li Z 1999 *Biosystems* **52** 55
[6] Liu Y, Guo X F, Xu J, Pan L and Wang S 2002 *J. Chem. Inf. Comput. Sci.* **42** 529
[7] Randic M, Vracko M, Nandy A and Basak S C 2000 *J. Chem. Inf. Comput. Sci.* **40** 1235
[8] Yau S S, Wang J, Niknejad A, Lu C, Jin N and Ho Y K 2003 *Nucleic Acids Res.* **31** 3078
[9] Zhang R and Zhang C T 1991 *Nucleic Acids Res.* **19** 6313
[10] Zhang R and Zhang C T 1994 *J. Biomol. Struct. Dyn.* **11** 767
[11] Zhang C T 1997 *J. Theor. Biol.* **187** 297
[12] Zhang C T, Zhang R and Ou H Y 2003 *Bioinformatics* **19** 593