A Brief Review: The *Z*-curve Theory and its Application in Genome Analysis

Ren Zhang[1],* and Chun-Ting Zhang[2],*

[1]*Center for Molecular Medicine and Genetics, Wayne State University Medical School, Detroit, MI 48201, USA;* [2]*Department of Physics, Tianjin University, Tianjin 300072, China*

Abstract: In theoretical physics, there exist two basic mathematical approaches, algebraic and geometrical methods, which, in most cases, are complementary. In the area of genome sequence analysis, however, algebraic approaches have been widely used, while geometrical approaches have been less explored for a long time. The *Z*-curve theory is a geometrical approach to genome analysis. The *Z*-curve is a three-dimensional curve that represents a given DNA sequence in the sense that each can be uniquely reconstructed given the other. The *Z*-curve, therefore, contains all the information that the corresponding DNA sequence carries. The analysis of a DNA sequence can then be performed through studying the corresponding *Z*-curve. The *Z*-curve method has found applications in a wide range of areas in the past two decades, including the identifications of protein-coding genes, replication origins, horizontally-transferred genomic islands, promoters, translational start sides and isochores, as well as studies on phylogenetics, genome visualization and comparative genomics. Here, we review the progress of *Z*-curve studies from aspects of both theory and applications in genome analysis.

Received on: September 08, 2013- Revised on: October 16, 2013- Accepted on: October 16, 2013

Keywords: GC profile, Gene finding, Genomic island, Replication origin, *Z*-curve.

1. INTRODUCTION

In theoretical physics, there exist two basic mathematical approaches, algebraic and geometrical methods, which, in most cases, are complementary. In the area of genome studies, however, algebraic approaches, such as Markov chain models and hidden Markov chain models, have been widely used, while geometrical approaches have been less explored for a long time. The *Z*-curve theory is a geometric approach to genome analysis.

The *Z*-curve is a 3-dimensional curve that represents a given DNA sequence in the sense that each can be *uniquely* reconstructed given the other [1-3]. The *Z*-curve, therefore, contains all the information that the corresponding DNA sequence carries. The analysis of a DNA sequence can then be performed through studying the corresponding *Z*-curve.

Historically, various methods for the graphical representation of DNA sequences were proposed, such as the H curve [4] and the 2-dimensional DNA walk [5]. It has been shown that most of these methods are, in fact, special cases of the *Z*-curve, and an extensive comparison between the *Z*-curve and other representations was detailed in reference [2]. One of the advantages of the *Z*-curve is its intuitiveness, enabling global and local compositional features of genomes to be grasped quickly in a perceivable form. The methodology of the *Z*-curve is a suitable platform on which other methods, such as statistics, can be integrated to address bioinformatics questions. The *Z*-curve method [1, 2] has found many applications in genome analysis since its initiation two decades ago. Here, we review the progress of the *Z*-curve studies from aspects of both theory and applications in genome research.

2. PART-1: THEORY OF THE *Z*-CURVE

2.1. Symmetry of Four DNA Bases and its Geometric Representation

The DNA sequence is composed of 4 kinds of nucleotides, adenine, cytosine, guanine and thymine, denoted by A, C, G and T, respectively. The number of possible combinations when taking 2 bases at a time from 4 bases is 6. The 6 combinations are: R (A/G) and Y (C/T); M (A/C) and K (G/T); W (A/T) and S (G/C), where R, Y, M, K, W and S represent the bases of puRine, pYrimidine, aMino, Keto, Weak hydrogen bonds and Strong hydrogen bonds, respectively, according to the NC-IUB recommendation [6]. The chemical structures of the four bases are shown in (Fig. **1**), illustrating the symmetry among the four bases. According to different criteria, the four bases can be classified into two categories.

(i) Criterion 1, according to the chemical structure of having single or double rings

$$\text{Bases}\begin{cases}\text{Purine,} & \text{R} = \text{A},\ \ \text{G},\\ \text{Pyrimidine,} & \text{Y} = \text{C},\ \ \text{T}.\end{cases}$$

(ii) Criterion 2, according to the chemical structure of having an amino or keto group

*Address correspondence to these authors at the Center for Molecular Medicine and Genetics, Wayne State University Medical School, Detroit, MI 48201, USA; Tel: +1 313 577 0027; Fax: +1 313 577 5218; E-mail: rzhang@med.wayne.edu and Department of Physics, Tianjin University, Tianjin 300072, China; Tel: +86 22 2740 2987; Fax: +86 22 2740 2697; E-mail: ctzhang@tju.edu.cn

1875-5488/14 $58.00+.00

$$\text{Bases}\begin{cases}\text{Amino,} & M = A,\ C,\\ \text{Keto,} & K = G,\ T.\end{cases}$$

(iii) Criterion 3, according to the structure of the double helix forming two or three hydrogen bonds in the Watson-Crick pair

$$\text{Bases}\begin{cases}\text{Weak,} & W = A,\ T,\\ \text{Strong,} & S = G,\ C.\end{cases}$$

We seek to find some geometrical representation for the above symmetry. If a 2-dimensional (plane) graph is adopted, we find that the symmetry can be represented by (Fig. **1A-C**). If a 3-dimensional graph is adopted, the regular cube, as shown in (Fig. **2A**), seems to be the unique choice to represent the symmetry. Each face of the cube is assigned to one, and only one, of the six characters: R, Y, M, K, W and S, thereby keeping the rule that R and Y, M and K, as well as W and S are on opposite sides. To prepare (Fig. **2A**), readers may cut (Fig. **2B**) and fold it along the dashed lines. Diagonals of the regular cube form a regular tetrahedron ACGT, as shown in (Fig. **2C**). Assigning one of A, C, G and T to each vertex of the tetrahedron as shown in (Fig. **2C**) is not arbitrary. Note that the vertex A of the tetrahedron is also the vertex of the cube, at which three faces of the cube, R, M and W, are crossed. The intersection base of R (A/G), M (A/C) and W(A/T) is A. Similar assignments can be applied to the vertices C, G and T, as shown in (Fig. **2C** and **D**).

To further the study, a coordinate system needs to be established (Fig. **2D**). The line connecting the middle point of an edge and that of the opposite edge of the tetrahedron is called the middle line. There are a total of three middle lines in a tetrahedron, crossing at the center O, and they are perpendicular to each other. A Cartesian coordinate system OXYZ can be set up by using the three middle lines, as shown in (Fig. **2D**).

Thus, the cube-tetrahedron geometric entity established here correctly reflects the symmetry of the four DNA bases.

2.2. The DNA Group

A regular tetrahedron is a geometric entity of high symmetry. All possible rotational motions which keep the tetrahedron fixed in the space form a group, called a tetrahedron group or T-group. As shown in (Fig. **2D**), a tetrahedron group consists of 12 operational elements, which are described below.

I, i.e., the identity operation;

R_x, R_y and R_z, i.e., the 180° rotation along x, y and z axes, respectively;

R_A, R_C, R_G and R_T, i.e., the 120° rotation along AO, CO, GO and TO axes, respectively;

$R^2{}_A$, $R^2{}_C$, $R^2{}_G$ and $R^2{}_T$, i.e., the 240° rotation along AO, CO, GO and TO axes, respectively.

A point with coordinates x, y and z will be transformed accordingly under the operational elements of the T-group. For example,

$$\text{I: } x \Leftrightarrow x,\ y \Leftrightarrow y,\ z \Leftrightarrow z;$$

$$R_x: x \Leftrightarrow x,\ y \Leftrightarrow -y,\ z \Leftrightarrow -z;$$

$$R_y: x \Leftrightarrow -x,\ y \Leftrightarrow y,\ z \Leftrightarrow -z;$$

$$R_z: x \Leftrightarrow -x,\ y \Leftrightarrow -y,\ z \Leftrightarrow z.$$

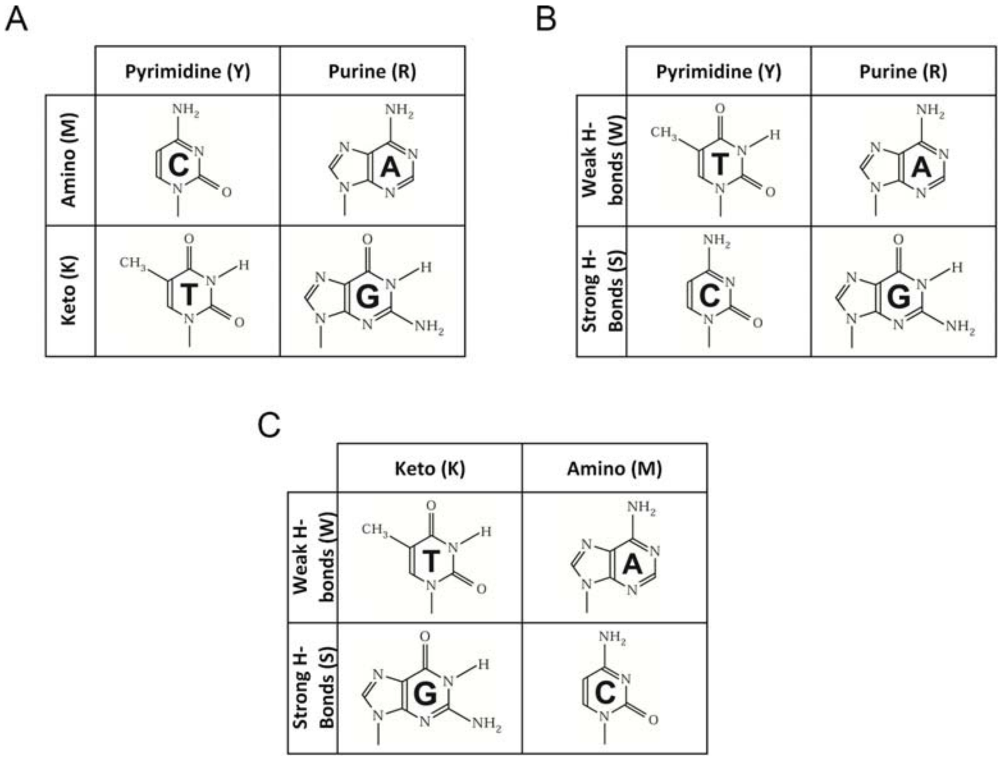

Fig. (1). Chemical structures of four DNA bases, displaying the basic symmetry.

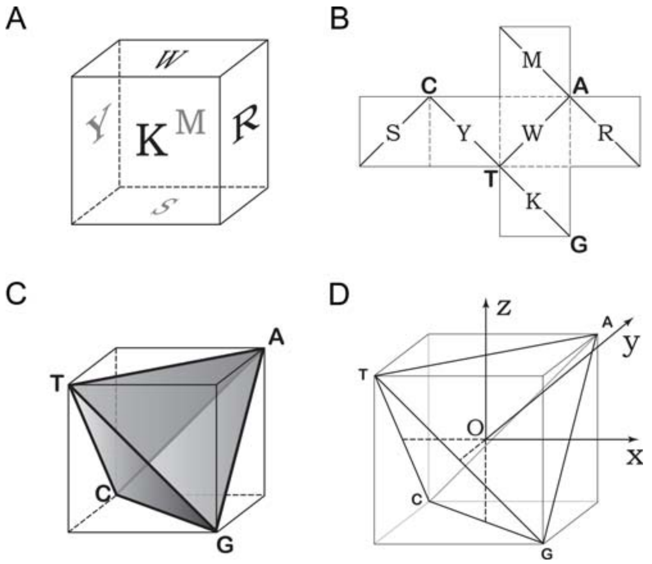

Fig. (2). The coordinate system based on the regular tetrahedron. **A**) a cube displaying the basic symmetry: R/Y, M/K and S/W symmetry; **B**) an extended plot for the cube. **C**) a cube and its inscribed tetrahedron; **D**) a coordinate system is set up to establish the *Z*-curve theory.

Table 1. Twelve Elements of the DNA Group (A_4 Group or the Tetrahedron Group).

Element	A_4 Group	Tetrahedron Group
I	A C G T	x y z
R_x	G T A C	x -y -z
R_y	C A T G	-x y -z
R_z	T G C A	-x -y z
R_A	A T C G	z x y
R_C	G C T A	z -x -y
R_G	T A G C	-z -x y
R_T	C G A T	-z x -y
R^2_A	A G T C	y z x
R^2_C	T C A G	-y -z x
R^2_G	C T G A	-y z -x
R^2_T	G A C T	y -z -x

The transforms of *x*, *y* and *z* under the 12 operations of the T-group are listed in (Table **1**). The four elements I, R_x, R_y and R_z form an invariant subgroup of the T-group, which is isomorphic to the Klein-4 group, or K4 group. The K4 group and its cosets exhaust the T-group. Therefore, all the 12 elements of the T-group can be divided into four classes, which are (I), (R_x, R_y, R_z), (R_A, R_C, R_G, R_T) and (R^2_A, R^2_C, R^2_G, R^2_T).

On the other hand, the set of all possible permutations of four objects forms a symmetric group, denoted by S_4. Among the 24 elements of the symmetric group S_4, the set of all 12 even permutations forms an invariant subgroup of S_4, referred to as the alternative group of order 4, denoted by A_4. The DNA group is defined as a particular A_4 group, in which the permuted objects are the four DNA bases A, C, G and T. According to the group theory, the T-group and A_4 group are isomorphic with each other. From the perspective of the abstract group, the T-group and the A_4 group are the same group, because they have the same group structure and matrix representation. The four bases A, C, G and T are assigned to the four vertices of the tetrahedron, as shown in (Fig. **2D**).

The four characters A, C, G and T will be transformed accordingly under the 12 operational elements of the DNA group or the A_4 group. For example,

$$\mathrm{I}:\ \mathrm{A} \Leftrightarrow \mathrm{A},\ \mathrm{C} \Leftrightarrow \mathrm{C},\ \mathrm{G} \Leftrightarrow \mathrm{G},\ \mathrm{T} \Leftrightarrow \mathrm{T};$$

$$R_x:\ \mathrm{A} \Leftrightarrow \mathrm{G},\ \mathrm{C} \Leftrightarrow \mathrm{T};$$

$$R_y:\ \mathrm{A} \Leftrightarrow \mathrm{C},\ \mathrm{G} \Leftrightarrow T;$$

$$R_z:\ \mathrm{A} \Leftrightarrow \mathrm{T},\ \mathrm{G} \Leftrightarrow \mathrm{C}.$$

Biologically, the transform R_x is called transition, whereas the transform R_y and R_z are called transversion. Here R_z is termed as the complementary transform.

We have previously established that the T-group and the A_4 group are the same group [3], and thus their elements should have one-to-one corresponding relations, as shown in (Table **1**). Both the A_4 group and the T-group are called the DNA group, which forms the basis of the *Z*-curve theory.

2.3. The *Z*-transform Formulas

Let the occurrence frequencies of the four bases, A, C, G and T in a DNA sequence be denoted by *a, c, g* and *t*, respectively. The normalized condition reads

$$a + c + g + t = 1, \tag{1}$$

indicating that among the four real numbers *a, c, g* and *t*, only three of them are independent.

Suppose that X, Y and Z are the coordinates of a point P in the coordinate system shown in (Fig. **2D**), which can be expressed by a linear combination of the four frequencies *a, c, g* and *t*, as follows

$$\begin{cases} X = a_{11}a + a_{12}c + a_{13}g + a_{14}t, \\ Y = b_{11}a + b_{12}c + b_{13}g + b_{14}t, \\ Z = c_{11}a + c_{12}c + c_{13}g + c_{14}t, \end{cases} \tag{2}$$

where a_{11}, a_{12}, …, c_{13}, c_{14} are real coefficients. Eqs. (**2**) can be re-written as a matrix form

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ b_{11} & b_{12} & b_{13} & b_{14} \\ c_{11} & c_{12} & c_{13} & c_{14} \end{pmatrix} \times \begin{pmatrix} a \\ c \\ g \\ t \end{pmatrix}. \tag{3}$$

The coordinates of the four vertices of the regular tetrahedron A, C, G and T are already known, and shown in (Table **2**). Based on the 12 numbers in (Table **2**), the 12 coefficients can be uniquely determined, and eqs. (**3**) becomes

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = \frac{\sqrt{3}}{4}\begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \times \begin{pmatrix} a \\ c \\ g \\ t \end{pmatrix}. \qquad \textbf{(4a)}$$

Equivalently, eqs. (**4**) may be re-written as

$$\begin{cases} X = \frac{\sqrt{3}}{4}[(a+g)-(c+t)] \\ Y = \frac{\sqrt{3}}{4}[(a+c)-(g+t)] \quad X,\ Y,\ Z \in [-\frac{\sqrt{3}}{4},\ \frac{\sqrt{3}}{4}] \\ Z = \frac{\sqrt{3}}{4}[(a+t)-(g+c)] \end{cases} \qquad \textbf{(4b)}$$

Table 2. Coordinates of the 4 Vertices of the Regular Tetrahedron ACGT[a]**.**

Coordinates	**Vertices**			
	A	**C**	**G**	**T**
X	$\sqrt{3}/4$	$-\sqrt{3}/4$	$\sqrt{3}/4$	$-\sqrt{3}/4$
Y	$\sqrt{3}/4$	$\sqrt{3}/4$	$-\sqrt{3}/4$	$-\sqrt{3}/4$
Z	$\sqrt{3}/4$	$-\sqrt{3}/4$	$-\sqrt{3}/4$	$\sqrt{3}/4$

[a] Refer to Fig. 2 (d) for the original coordinate system, where the height of the tetrahedron is 1. Consequently, the edge length of the tetrahedron is $\sqrt{6}/2$, and the edge length of the cube is $\sqrt{3}/2$.

Eqs. (**4**) are called the *Z*-transform formulas, which were first derived in 1991 by a totally different way [1]. The *Z*-transform formulas transform the four base frequencies into three coordinates of a point (called a mapping point) in a three-dimensional space. As previously indicated [1], for convenience, we introduced the reduced coordinate system *x*, *y* and *z*

$$\begin{cases} X = \frac{\sqrt{3}}{4}x, \\ Y = \frac{\sqrt{3}}{4}y, \\ Z = \frac{\sqrt{3}}{4}z, \end{cases} \qquad \textbf{(5)}$$

such that

$$\begin{cases} x = (a+g)-(c+t), \\ y = (a+c)-(g+t), \quad x,\ y,\ z \in [-1,\ 1]. \\ z = (a+t)-(g+c). \end{cases} \qquad \textbf{(6a)}$$

In what follows, we always use the *Z*-transform formulas based on the reduced coordinate system eqs. (**6**), unless otherwise indicated. Equivalently, eqs. (**6a**) can be also re-written as a matrix form

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \times \begin{pmatrix} a \\ c \\ g \\ t \end{pmatrix}. \qquad \textbf{(6b)}$$

Letting

$$U = \begin{pmatrix} x \\ y \\ z \end{pmatrix},\ V = \begin{pmatrix} a \\ c \\ g \\ t \end{pmatrix}, \qquad \textbf{(7)}$$

and

$$Z = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}, \qquad \textbf{(8a)}$$

Equivalently, eqs. (**6b**) may be re-written with a simplified form

$$U = Z \times V. \qquad \textbf{(9)}$$

The reverse equation of eqs. (**6**) is

$$\begin{pmatrix} a \\ c \\ g \\ t \end{pmatrix} = \frac{1}{4} \times \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \frac{1}{4} \times \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \\ -1 & -1 & 1 \end{pmatrix} \times \begin{pmatrix} x \\ y \\ z \end{pmatrix}. \qquad \textbf{(10)}$$

It is shown that regardless of the values of *x*, *y* and *z*, $a + c + g + t \equiv 1$. In fact, it was shown in 1991 that the mapping point P (*x*, *y*, *z*), corresponding to *a*, *c*, *g* and *t*, is always situated within the tetrahedron ACGT shown in (Fig. **2D**) [1].

To provide a clear visualization, the tetrahedron and the mapping points within it are projected onto some coordinate planes. Referring to (Fig. **2C** and **D**), note that the tetrahedron ACGT has 4 vertices A, C, G and T, and six edges AC, AG, AT, CG, CT and GT. Interestingly, the projection of six edges onto any coordinate plane forms a regular square and two diagonal lines within the square, where the projection of four vertices of the tetrahedron forms four vertices of the square, as shown in (Fig. **3A**, **B** and **C**), for the x-y, x-z and y-z planes, respectively. Note that (Fig. **3A**, **B** and **C**) are in accordance with (Fig. **1A, B** and **C**), respectively. It should be noted that A, C, G and T are sometimes used to denote DNA bases, while the same symbols can represent vertices of the tetrahedron or squares. Refer to (Fig. **3A**) first. Projections of four edges AG, AC, CT and GT form the four sides of the square, whereas those of AT and GC form the two diagonal lines of the square. Based on the *Z*-transform formulas eqs. (**6**), the base composition of a DNA sequence, i.e., the values of *a, c, g* and *t*, can be visualized by observing the position of the mapping point in the square. For example, if the DNA sequence has only one kind of base, say, A, then *a* = 1, *c* = *g* = *t* = 0. The corresponding mapping point is situated at the vertex A in (Fig. **2A**). Similar results for (Fig. **3A**) are summarized as follows.

Vertex A: $a = 1,\ \ c = g = t = 0;$ Vertex C: $c = 1,\ \ a = g = t = 0;$

Vertex G: $g = 1,\ \ a = c = t = 0;$ Vertex T: $t = 1,\ \ a = c = g = 0;$

Side AG: $a+g=1,\ \ c=t=0;$ Side GT: $g+t=1,\ \ a=c=0;$

Side TC: $c+t=1,\ \ a=g=0;$ Side CA: $a+c=1,\ \ g=t=0;$

$x>0,\ \ a+g>1/2;\ \ x=0,\ \ a+g=1/2;\ \ x<0,\ \ a+g<1/2;$

$y>0,\ \ a+c>1/2;\ \ y=0,\ \ a+c=1/2;\ \ y<0,\ \ a+c<1/2;$

First quadrant: $a+g>1/2$ and $a+c>1/2;$

Second quadrant: $c+t>1/2$ and $a+c>1/2;$

Third quadrant: $c+t>1/2$ and $g+t>1/2;$

Fourth quadrant: $a+g>1/2$ and $g+t>1/2;$

Diagonal AOT: $g=c;\ \ \Delta\mathrm{AGT}:\ g>c;\ \ \Delta\mathrm{ATC}:\ c>g;$

Diagonal COG: $a=t;\ \ \Delta\mathrm{AGC}:\ a>t;\ \ \Delta\mathrm{CTG}:\ t>a;$

$\Delta\mathrm{AOG}:\ a>t$ and $g>c;\ \ \Delta\mathrm{AOC}:\ a>t$ and $c>g;$

$\Delta\mathrm{COT}:\ t>a$ and $c>g;\ \ \Delta\mathrm{GOT}:\ t>a$ and $g>c;$

Origin O: $a=c=g=t=1/4.$

Similar deductions for the annotation of (Fig. **3B** and **C**) are left out for readers who might be interested in doing so.

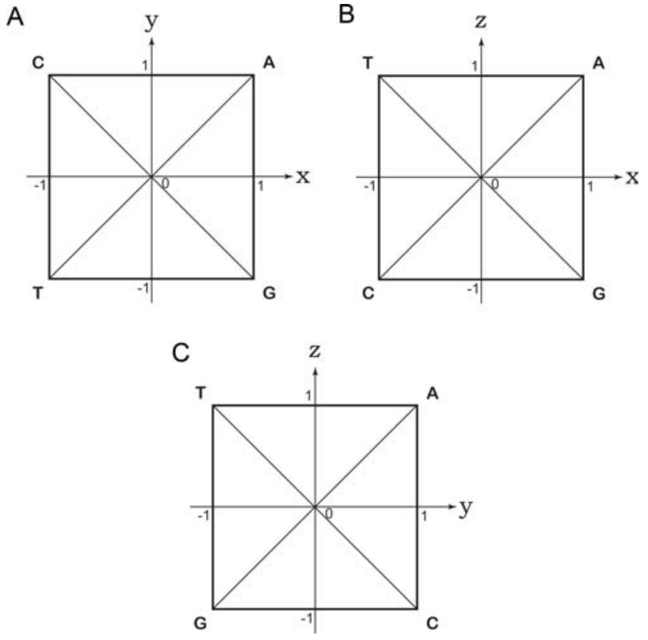

Fig. (3). Projection of the 3-D coordinates onto planes. The projection of the 3-D coordinate system onto the **A**) x-y, **B**) x-z and **C**) y-z planes.

2.4. Linear Representation of the DNA Group

Based on the reduced coordinate system, the coordinates of the four vertices A, C, G and T can be represented by

$$(A)=\begin{pmatrix}1\\1\\1\end{pmatrix},\ C=\begin{pmatrix}-1\\1\\-1\end{pmatrix},\ (G)=\begin{pmatrix}1\\-1\\-1\end{pmatrix},\ T=\begin{pmatrix}-1\\-1\\1\end{pmatrix}. \qquad \textbf{(11)}$$

We previously established the linear representation of the DNA group, i.e., the tetrahedron group or the alternative group A$_4$ in 1997 [3]. For readers' convenience, here we re-write the result (see eqs. (**6**) in [3] as follows:

$$I=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix},\ R_x=\begin{pmatrix}1&0&0\\0&-1&0\\0&0&-1\end{pmatrix},\ R_y=\begin{pmatrix}-1&0&0\\0&1&0\\0&0&-1\end{pmatrix},\ R_z=\begin{pmatrix}-1&0&0\\0&-1&0\\0&0&1\end{pmatrix},$$

$$R_A=\begin{pmatrix}0&0&1\\1&0&0\\0&1&0\end{pmatrix},\ R_C=\begin{pmatrix}0&0&1\\-1&0&0\\0&-1&0\end{pmatrix},\ R_G=\begin{pmatrix}0&0&-1\\-1&0&0\\0&1&0\end{pmatrix},\ R_T=\begin{pmatrix}0&0&-1\\1&0&0\\0&-1&0\end{pmatrix},$$

$$R_A^2=\begin{pmatrix}0&1&0\\0&0&1\\1&0&0\end{pmatrix},\ R_C^2=\begin{pmatrix}0&-1&0\\0&0&-1\\1&0&0\end{pmatrix},\ R_G^2=\begin{pmatrix}0&-1&0\\0&0&1\\-1&0&0\end{pmatrix},\ R_T^2=\begin{pmatrix}0&1&0\\0&0&-1\\-1&0&0\end{pmatrix}. \qquad \textbf{(12)}$$

This matrix representation depicts correct relationships among the 12 elements of the DNA group. For example, a rotation of 180° along the *x*-axis in (Fig. **2D**), followed by another similar rotation, leads to the original state, i.e.,

$$R_x\times R_x=I\ \Leftrightarrow\ \begin{pmatrix}1&0&0\\0&-1&0\\0&0&-1\end{pmatrix}\times\begin{pmatrix}1&0&0\\0&-1&0\\0&0&-1\end{pmatrix}=\begin{pmatrix}1&0&0\\0&1&0\\0&0&1\end{pmatrix}=I\ . \qquad \textbf{(13)}$$

That is to say, the matrix representation not only results in a one-to-one correspondence among elements of the DNA group, but also correctly reflects their relations based on the multiplication of matrices.

In the following, we show that the transform matrix (3x4) eq. (**8a**) and its variants also constitute a one-to-one representation to each element of the DNA group. For this purpose, eq. (**8a**) can be re-written as

$$Z=\begin{pmatrix}1&-1&1&-1\\1&1&-1&-1\\1&-1&-1&1\end{pmatrix}=\begin{pmatrix}(A)&(C)&(G)&(T)\end{pmatrix}, \qquad \textbf{(8b)}$$

where (A), (C), (G) and (T) are denoted by eqs. (**11**). Referring to (Table **2**), we find that the order of the four nucleotides above correspond to the element I of the A$_4$ group. Similarly, its 11 variants can be derived, and are listed as follows

$$Z_I=\begin{pmatrix}1&-1&1&-1\\1&1&-1&-1\\1&-1&-1&1\end{pmatrix},\ Z_x=\begin{pmatrix}1&-1&1&-1\\-1&-1&1&1\\-1&1&1&-1\end{pmatrix},\ Z_y=\begin{pmatrix}-1&1&-1&1\\1&1&-1&-1\\-1&1&1&-1\end{pmatrix},$$

$$Z_z=\begin{pmatrix}-1&1&-1&1\\-1&-1&1&1\\1&-1&-1&1\end{pmatrix},\ Z_A=\begin{pmatrix}1&-1&-1&1\\1&-1&1&-1\\1&1&-1&-1\end{pmatrix},\ Z_C=\begin{pmatrix}1&-1&-1&1\\-1&1&-1&1\\-1&-1&1&1\end{pmatrix},$$

$$Z_G=\begin{pmatrix}-1&1&1&-1\\-1&1&-1&1\\1&1&-1&-1\end{pmatrix},\ Z_T=\begin{pmatrix}-1&1&1&-1\\1&-1&1&-1\\-1&-1&1&1\end{pmatrix},\ Z_A^2=\begin{pmatrix}1&1&-1&-1\\1&-1&-1&1\\1&-1&1&-1\end{pmatrix},$$

$$Z_C^2=\begin{pmatrix}-1&-1&1&1\\-1&1&1&-1\\1&-1&1&-1\end{pmatrix},\ Z_G^2=\begin{pmatrix}-1&-1&1&1\\1&-1&-1&1\\-1&1&-1&1\end{pmatrix},\ Z_T^2=\begin{pmatrix}1&1&-1&-1\\-1&1&1&-1\\-1&1&-1&1\end{pmatrix}, \qquad \textbf{(14)}$$

where $Z_I=Z$. Based on the 12 Z matrices, we have

$$Z_i\times V=R_i\times U,\qquad i=I,\ x,\ y,\ z,\ A,\ C,\ G,\ T,\ A^2,\ C^2,\ G^2,\ T^2. \qquad \textbf{(15)}$$

Note that each Z_i corresponds to each R_i by a way of one-to-one correspondence. Therefore, the Z matrix also constitutes a representation of the DNA group in this sense. However, it is not an ordinary representation of the DNA group, because the similar multiplication relation such as eq. (**13**) does not exist among the Z matrices.

It should also be noted that the *Z*-transform formulas eqs. (**6**), which transform the nucleotide frequencies into the coordinates of a point in a three-dimensional space, are unique and invariant under the operations of the DNA group. The *Z*-transform formulas shown in eqs. (**6**) represent the unique set of equations which reflect the inherent features of the DNA group. The *Z*-transform formulas are the core of the *Z*-curve theory.

2.5. The *Z*-transform Formulas for Studying Correlations of Multiple Nucleotides

To extract features of a given DNA sequence, in addition to considering occurrence frequencies of a single nucleotide, correlations of multiple nucleotides should also be considered. Therefore, the *Z*-transform formulas should be extended to consider the correlations of multiple nucleotides.

1). The case of a single nucleotide

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = Z \times \begin{pmatrix} p(A) \\ p(C) \\ p(G) \\ p(T) \end{pmatrix}, \quad a = p(A), \quad c = p(C), \quad g = p(G), \quad t = p(T). \qquad (\mathbf{16})$$

Eqs. (**16**) are equivalent to eqs. (**9**). Here we have 3 (= 3x4°) parameters.

2). The case of di-nucleotides

$$\begin{pmatrix} x_H \\ y_H \\ z_H \end{pmatrix} = Z \times \begin{pmatrix} p(HA) \\ p(HC) \\ p(HG) \\ p(HT) \end{pmatrix}, \quad H = A, \ C, \ G, \ T, \qquad (\mathbf{17})$$

where $p(HA)$ represents occurrence frequencies of the di-nucleotide *HA*, and so forth. Here we have 12 (=3x4^1) parameters.

3). The case of tri-nucleotides

$$\begin{pmatrix} x_{HI} \\ y_{HI} \\ z_{HI} \end{pmatrix} = Z \times \begin{pmatrix} p(HIA) \\ p(HIC) \\ p(HIG) \\ p(HIT) \end{pmatrix}, \quad H,I = A, \ C, \ G, \ T, \qquad (\mathbf{18})$$

where $p(HIA)$ represents occurrence frequencies of the tri-nucleotide *HIA*, and so forth. Here we have 48 (=3x4^2) parameters.

4). The case of tetra-nucleotides

$$\begin{pmatrix} x_{HIJ} \\ y_{HIJ} \\ z_{HIJ} \end{pmatrix} = Z \times \begin{pmatrix} p(HIJA) \\ p(HIJC) \\ p(HIJG) \\ p(HIJT) \end{pmatrix}, \quad H,I,J = A, \ C, \ G, \ T, \qquad (\mathbf{19})$$

where $p(HIJA)$ represents occurrence frequencies of the four-nucleotide *HIJA*, and so forth. Here we have 192 (=3x4^3) parameters.

5). The case of penta-nucleotides

$$\begin{pmatrix} x_{HIJK} \\ y_{HIJK} \\ z_{HIJK} \end{pmatrix} = Z \times \begin{pmatrix} p(HIJKA) \\ p(HIJKC) \\ p(HIJKG) \\ p(HIJKT) \end{pmatrix}, \quad H,I,J,K = A, \ C, \ G, \ T, \qquad (\mathbf{20})$$

where $p(HIJKA)$ represents occurrence frequencies of the five-nucleotide *HIJKA*, and so forth. Here we have 768 (=3x4^4) parameters.

6). The case of hexa-nucleotides

$$\begin{pmatrix} x_{HIJKL} \\ y_{HIJKL} \\ z_{HIJKL} \end{pmatrix} = Z \times \begin{pmatrix} p(HIJKLA) \\ p(HIJKLC) \\ p(HIJKLG) \\ p(HIJKLT) \end{pmatrix}, \quad H,I,J,K,L = A, \ C, \ G, \ T, \qquad (\mathbf{21})$$

where $p(HIJKLA)$ represents occurrence frequencies of the six-nucleotide *HIJKLA*, and so forth. Here we have 3072 (=3x4^5) parameters.

To calculate the occurrence frequencies of multiple nucleotides for a given DNA sequence, we use a moving window with size = 1, 2, 3, 4, 5 and 6. Starting from the first nucleotide or base, move the sliding window rightward one base at a time, and then the frequencies can be calculated. Substitute the frequencies into eqs. (**16**) to (**21**), and then the *Z*-curve parameters can be obtained. For some applications, eq. (**16**), i.e., 3 parameters are sufficient. However, in some cases, more parameters are needed. The space spanned by the 3 parameters is denoted by V_1, and similarly we have V_2, V_3, V_4, V_5 and V_6, respectively, corresponding to eqs. (**16**) to (**21**). Usually, the direct sum among different spaces is needed. For most applications, there are six possible choices

$$\mathrm{V} = \begin{pmatrix} \mathrm{V}_1, \\ \mathrm{V}_1 \oplus \mathrm{V}_2, \\ \mathrm{V}_1 \oplus \mathrm{V}_2 \oplus \mathrm{V}_3, \\ \mathrm{V}_1 \oplus \mathrm{V}_2 \oplus \mathrm{V}_3 \oplus \mathrm{V}_4, \\ \mathrm{V}_1 \oplus \mathrm{V}_2 \oplus \mathrm{V}_3 \oplus \mathrm{V}_4 \oplus \mathrm{V}_5, \\ \mathrm{V}_1 \oplus \mathrm{V}_2 \oplus \mathrm{V}_3 \oplus \mathrm{V}_4 \oplus \mathrm{V}_5 \oplus \mathrm{V}_6, \end{pmatrix}, \qquad (\mathbf{22})$$

where the symbol $\oplus$ represents the direct sum of two spaces. The dimensions of the spaces V_1 to V_6 are 3, 15 (= 3+12), 63 (= 15+48), 255 (= 63+192), 1023 (= 255+768) and 4095 (= 1023+3072), respectively. Generally, for the space $\mathrm{V} = \mathrm{V}_1 \oplus \mathrm{V}_2 \oplus ... \oplus \mathrm{V}_\mathrm{m}$, the dimension is $4^m - 1$.

2.6. Quadratic Form of *x*, *y* and *z*

Starting from eq. (**9**)

$$U = Z \times V, \qquad (\mathbf{9})$$

We have

$$U^T = V^T \times Z^T, \qquad (\mathbf{23})$$

where "T" means the transpose operation of a matrix. Furthermore, we find

$$U^T \times U = V^T \times Z^T \times Z \times V. \qquad (\mathbf{24})$$

Simple derivation shows that

$$x^2 + y^2 + z^2 = 4S - 1, \qquad (\mathbf{25})$$

where S is defined by

$$S = a^2 + c^2 + g^2 + t^2. \quad \textbf{(26)}$$

S, named as "genome order index" [7], is useful for designing a fast genome segmentation algorithm [8, 9]. We also observed that for most genomes

$$S < 1/3. \quad \textbf{(27)}$$

Eq. (**27**) has a clear geometrical explanation. The surface of the inscribed sphere is described by the equation

$$x^2 + y^2 + z^2 = \left(\frac{1}{\sqrt{3}}\right)^2 = \frac{1}{3}. \quad \textbf{(28)}$$

Therefore, S<1/3 implies that the mapping point is within the inscribed sphere [7].

2.7. The *Z*-curve

One of the most important applications of the *Z*-transform formulas is to derive the equation of the *Z*-curve. Consider a DNA sequence with *N* bases that are inspected one base at a time. From the first base to the n^{th} base, compute accumulative numbers of the bases A, C, G and T, denoted by A_n, C_n, G_n and T_n, respectively. Based on the *Z*-transform formulas eqs (**6**), we find

$$\begin{pmatrix} x(n) \\ y(n) \\ z(n) \end{pmatrix} = Z \times \begin{pmatrix} A_n/n \\ C_n/n \\ G_n/n \\ T_n/n \end{pmatrix}. \quad \textbf{(29)}$$

Multiplied by n to both hands of eq. (**29**), and letting

$$x_n = n \times x(n), \quad y_n = n \times y(n), \quad z_n = n \times z(n), \quad \textbf{(30)}$$

we have

$$\begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix} = Z \times \begin{pmatrix} A_n \\ C_n \\ G_n \\ T_n \end{pmatrix}, \quad \textbf{(31a)}$$

or equivalently

$$\begin{cases} x_n = (A_n + G_n) - (C_n + T_n), \\ y_n = (A_n + C_n) - (G_n + T_n), \\ z_n = (A_n + T_n) - (G_n + C_n), \end{cases} n = 0,\ 1,\ 2,\ 3,....,\ N,\ \ x_x, y_n, z_n \in [-N, N], \quad \textbf{(31b)}$$

which was first derived in 1994 by an entirely different method [2]. It should be noted that A_n, C_n, G_n and T_n are the cumulative occurrence numbers of A, C, G and T, respectively, in the sub-sequence from the 1^{st} base to the *n*th base in the sequence with length *N*. We define A_0=C_0=G_0=T_0=0, therefore, $x_0 = y_0 = z_0 = 0$. The *Z*-curve is defined as the connection of the nodes P_0 (x_0, y_0, z_0), , P_2 (x_2, y_2, z_2), …, P_N(x_N, y_N, z_N) one by one sequentially with straight lines. The connection results in a curve with a zigzag shape, hence the name *Z*-curve. Note that the *Z*-curve always starts from the origin of the three-dimensional coordinate system. Once the coordinates x_n, y_n and z_n (n = 1, 2, …, *N*) of a *Z*-curve are given, the corresponding DNA sequence can be reconstructed uniquely from the so-called inverse *Z*-transform formulas

$$\begin{pmatrix} A_n \\ C_n \\ G_n \\ T_n \end{pmatrix} = \frac{n}{4} \times \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} + \frac{1}{4} \times \begin{pmatrix} 1 & 1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & -1 \\ -1 & -1 & 1 \end{pmatrix} \times \begin{pmatrix} x_n \\ y_n \\ z_n \end{pmatrix}, \quad n = 1,\ 2,\ ...,\ N, \quad \textbf{(32)}$$

where the normalized relation of $A_n + C_n + G_n + T_n = n$ is used.

The three components of the *Z*-curve, x_n, y_n and z_n, represent three independent distributions, that is, those of purine/pyrimidine (R/Y), amino/keto (M/K) and strong-H bond/weak-H bond (S/W) bases, respectively, and they completely describe the DNA sequence being studied. In the sub-sequence constituted from the 1^{st} base to the *n*th bases of the sequence, when purine bases (A/G) are in excess of pyrimidine bases (C/T), $x_n > 0$, otherwise, $x_n < 0$, and when the numbers of purine (A/G) and pyrimidine bases (C/T) are identical, $x_n = 0$. Similarly, when amino bases (A/C) are in excess of keto bases (G/T), $y_n > 0$, otherwise, $y_n < 0$, and when the numbers of amino (A/C) and keto bases (G/T) are identical, $y_n = 0$. Finally, when weak H-bond bases (A/T) are in excess of strong H-bond bases (G/C), $z_n > 0$, otherwise, $z_n < 0$, and when the numbers of (A/T) and (G/C) bases are identical, $z_n = 0$. The x_n and y_n components are termed RY and MK disparity curves, respectively. Similarly, the AT and GC disparity curves are defined by $(x_n + y_n)/2$ and $(x_n-y_n)/2$, which shows the excess of A over T and G over C, along the genome. The RY and MK disparity curves, as well as AT and GC disparity curves, can be used to predict replication origins of various genomes.

2.8. The GC Profile

For most genome sequences, Chargaff Parity Rule II holds, i.e., $A_N \cong T_N$ and $C_N \cong G_N$, where *N* is the length of a genome or a chromosome. According to eqs (**31**), we find

$$x_n \cong 0,\ y_n \cong 0,\ z_n >> 1, \quad \text{for} \quad n >> 1 \quad \textbf{(33)}$$

Therefore, the curves of $z_n \sim n$ are roughly straight lines in this case. To amplify the variations of the straight-line-like curve, the curve of $z_n \sim n$ is firstly fitted by a straight line using the least square technique,

$$z = kn, \quad \textbf{(34)}$$

where (z, n) is the coordinate of a point on the fitted straight line and k is its slope. We define the *z'* curve, where

$$z'_n = z_n - kn. \quad \textbf{(35)}$$

Therefore, the variations of $z_n \sim n$ curve deviated from the straight line, which corresponds to a constant G+C content (see eq. (**36**) below), are protruded by the *z'* curve. One may also use the average slope of the $z_n \sim n$ curve to compute k, $k = z_N / N$, where z_N is the terminal coordinate of the $z_n \sim n$ curve and N is the sequence length. The essence of the z' curve is to display the variations of the G+C content along a genome or chromosome based on the cumulative count of G and C bases. Let $\overline{G+C}$ denote the average G+C content within a region Δn in a sequence, it was shown that [10].

$$\overline{\mathrm{G+C}} = \frac{1}{2}(1-k-\frac{\Delta z'_n}{\Delta n}) \equiv \frac{1}{2}(1-k-k'), \qquad \mathbf{(36)}$$

where $k' = \Delta z'_n / \Delta n$ is the average slope of the *z*' curve within the region Δn. Both quantities of $\Delta z'_n$ and Δn can be calculated by using the *z*' curve. It is clear to see from eq. (**36**) that a jump in the *z*' curve, i.e., $k' > 0$, indicates a decrease of G+C content or an increase of A+T content, whereas a drop in the *z*' curve, i.e., $k' < 0$, indicates an increase of G+C content or a decrease of A+T content. The region Δn is usually chosen to be a fragment of a DNA sequence. The above method to calculate G+C content is called the windowless technique [10].

The GC profile is defined as $-z'$, because it is more intuitive in the sense that a jump denotes an increase in GC content. We emphasize the importance of the GC profile for genome studies, because it represents a windowless technique to calculate the G+C content along genome sequences.

2.9. A Segmentation Algorithm Based on the *Z*-transform

Let *n* be a point within a DNA sequence of length $N (2 \le n \le N-1)$, which divides the whole sequence into two parts: the right and left sub-sequences, and then denote frequencies of bases in the right sub-sequence and left sub-sequence by (a_r, c_r, g_r, t_r) and (a_l, c_l, g_l, t_l), respectively. The frequencies are mapped onto two points, P_R (x_r, y_r, z_r) and P_L(x_l, y_l, z_l), in a 3-D space, where

$$\begin{pmatrix} x_i \\ y_i \\ z_i \end{pmatrix} = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \times \begin{pmatrix} a_i \\ c_i \\ g_i \\ t_i \end{pmatrix}, \qquad i = r, l. \qquad \mathbf{(37)}$$

The square of Euclidean distance between the two points is denoted by D, where

$$D(n) = (x_r - x_l)^2 + (y_r - y_l)^2 + (z_r - z_l)^2 . \qquad \mathbf{(38)}$$

Substituting eqs. (**37**) into eq. (**38**), we have

$$D(n) = C \times [(a_r - a_l)^2 + (c_r - c_l)^2 + (g_r - g_l)^2 + (t_r - t_l)^2], \qquad \mathbf{(39)}$$

where C is a constant. Note that D is a function of *n*. Suppose that when $n = n^* \ (2 \le n^* \le N-1), D(n^*) = \text{Maximun}.$ Then the point n^* is called a compositional segmentation point [8]. The segmentation algorithm is recursive, i.e., after *n** is determined, the same procedure is applied to both the left and right sub-sequences recursively, until *D*(*n*) is less than a given threshold. For more details refer to [8].

Eq. (**39**) can be extended to the case of a binary sequence. For example, by replacing the bases G and C with S, and bases A and T with W, a DNA sequence can be transformed into a binary sequence of S and W. In this case, the algorithm results in compositional segmentation points according to GC content. A software, called GC-Profile, was developed to implement the algorithm for genome segmentation [9].

3. PART-2: APPLICATIONS IN GENOME ANALYSIS

The *Z*-curve theory has been successfully applied in many different research areas in analyzing genomes of bacteria, archaea, eukaryotes and viruses. The applications include, to name a few, the identifications of protein-coding genes, replication origins, horizontally-transferred genomic islands, isochore structures, genome segmentation points, promoters and translational start sites, as well as studies on nucleosome positioning, DNA curvature profiles, phylogenetics and comparative genomics, in various organisms (Table **3**). It is not practical to cover all these areas in detail in a single review, and thus we will only highlight some studies.

3.1. Identification of Protein-coding Genes

One of the most important applications of the *Z*-curve theory is gene-finding in various genomes. The principle in using the *Z*-curve theory to identify protein-coding genes is straightforward. Based on the *Z*-transform formulas, the occurrence frequencies of 4 bases in a DNA sequence are mapped onto a point in a 3-dimensional (3-D) or 15-, 63-, 255-, 1023- and 4095-D space, depending on the number of correlated bases under consideration (eqs. (**16**) to (**22**)). The first application was for gene recognition in the budding yeast genome, where a 3-D space (eqs. (**16**)) was adopted [11]. However, since the protein coding sequence has 3 phases, the 3 *Z*-curve parameters are expanded to 9 (3x3 phases) parameters. Adding the genome order index *S* (eq. (**26**)) into the set of 9 *Z*-curve parameters, a 10-D space is spanned by the 10 parameters. It was observed that the mapping points of protein coding sequences and non-coding sequences are distributed in two distinct regions in the 10-D space, although there is minor overlapping [12]. Therefore, the two kinds of points can be discriminated by the Fisher discriminant method, or other classifiers, such as support vector machines.

The *Z*-curve algorithm was first applied to recognize protein coding genes in the budding yeast (*Saccharomyces cerevisiae*) genome with an accuracy better than 95%, where the accuracy is defined as the average of sensitivity and specificity [11]. The same algorithm achieved an accuracy rate over 98% in the *Vibrio cholerae* genome, based on 9 parameters only [13]. The success of the above studies led to the development of a series of *ab initio* gene-finding software for various species with different numbers of *Z*-curve parameters.

3.1.1. Gene-finding in Bacterial, Archaeal, Phage and Virus Genomes

Based on 33 *Z*-curve parameters, we developed ZCURVE 1.0, which is an *ab initio* gene-finding software for bacterial and archaeal genomes [12]. Based on the 9 *Z*-curve parameters, ZCURVE_V was developed for identifying protein-coding genes in viral and phage genomes [14]. We also developed the software, ZCURVE_CoV, for gene-finding in coronavirus genomes, with special applications for SARS-coronavirus genomes [15, 16].

The above set of gene-finding software has been widely used in various laboratories worldwide. For example, ZCURVE 1.0 has been used for annotating protein-coding genes in many newly sequenced bacterial genomes, such as those of *Acinetobacter baumannii* [17], *Variovorax paradoxus* [18], *Amycolatopsis mediterranei* [19], *Bacillus thuringiensis* [20], *Streptomyces tendae* [21],

Table 3. A Partial List of *Z*-curve Applications in Genome Analysis.

Research areas	Involved *Z*-Curve Components	Algorithm, Software or Database	Life Domains or Virus	Species
Protein-coding gene recognition [a]	*x, y, z, S*	*Z*-curve algorithm [1, 2], ZCURVE [12]	Bacteria	*Acinetobacter baumannii* [17], *Variovorax paradoxus* [18], *Amycolatopsis mediterranei* [19], *Bacillus thuringiensis* [20], *Streptomyces tendae* [21], *Phaeobacter gallaeciensis* [22], *Desulfobacterium autotrophicum* [23], *Mycobacterium tuberculosis* [24], *Magnetospirillum gryphiswaldense* [25], *Beggiatoa* [26]
			Phage, plasmid	Fosmids of marine Planctomycetes [127], plasmids in the human gut [128], phage Rtp [129]
			Archaea	Archaea of the ANME-1 group [27]
			Eukaryotes	*Leptospira interrogans* [130], Yeast [11], Short human protein-coding genes [56, 131], Drosophila [55]
		ZCURVE_V [14], ZCURVE_CoV [15]	Virus, Coronavirus, phages	Prophage [33], Me Tri virus [28], novel human coronaviruses NL63 and HKU1 [34], novel bat coronaviruses [35], bat coronaviruses 1A, 1B and HKU8 [36], novel human coronavirus [37]
			SARS_CoV	Various strains of SARS_CoV [38-53]
Replication origin identification	AT, GC, MK and RY disparity [b]	Ori-finder [78], DoriC [132, 133]	Archaea	*Methanosarcina mazei*[69], *Halobacterium* species NRC-1[63], *Methanocaldococcus jannaschii* [68], *Sulfolobus acidocaldarius* [72], *Haloferax volcanii* [73], *Desulfurococcus kamchatkensis* [74], *Thermococcus sibiricus* [75], *Sulfolobus islandicus* [76]
			Bacteria	*Moraxella catarrhalis* [79], *Sorangium cellulosum* [80], *Microcystis aeruginosa* [80], *Cyanothece* [81], *Cupriavidus metallidurans* [82], *Azolla filiculoides* [83], *Variovorax paradoxus* [18], *Corynebacterium pseudotuberculosis* [84], [85], *Orientia tsutsugamushi* [86], *Propionibacterium freudenreichii* [87], *Laribacter hongkongensis* [88], *Legionella pneumophila* [89], *Ehrlichia canis* [90]
			Phage, plasmid	*Streptococcus pneumoniae* Virulent Phage Dp-1 [134], R-plasmid pPRS3a from *Bacillus cereus* [135]
Genomic island identification	*z'*	GC profile [9, 10]	Bacteria	*Corynebacterium efficiens* [105], *Rhodopseudomonas palustris* [106], *Corynebacterium glutamicum* [104], *Vibrio vulnificus* and *Bacillus cereus* [103], *Agrobacterium tumefaciens*, *Rolstonia solanacearum*, *Xanthomonas axonopodis*, *Xanthomonas campestris*, *Xylella fastidiosa* and *Pseudomonas syringae* [107], *Streptomyces lividans* [108], *Parachlamydiaceae* UWE25 [109], *epsilon proteobacteria Sulfurovum* and *Nitratiruptor* [110], *Acinetobacter oleivorans* [111], *Silicibacter pomeroyi* [112]
			Archaea	*Haloquadratum walsbyi* [136]
GC content variation, isochore, genome segmentation	*z', S*	GC profile [9, 10]	Eukaryotes	Human genome: isochores [94, 98, 137] and replication time zones [138]; Isochores for chicken [97], *Arabidopsis thaliana* [96], mice [95] and pig [99]; DNA curvature profile for *Aspergillus fumigatus* [100]
			Bacteria	*Bifidobacterium longum* [139], *Streptomyces avermitilis* [140], *Erwinia amylovora* [141], *Ralstonia pickettii* [142]
Promoter, translational start sites, nucleosome positioning	*x, y, z*	*Z*-curve algorithm [11, 12], GS-finder [113]	Bacteria	Translational start sites [113] and promoters [115] of *Escherichia coli* and *Bacillus subtilis*
			Eukaryotes	Human Pol II promoter [114], Yeast genome for stable and dynamic nucleosome positioning [116]

(Table 3) contd….

Research areas	Involved *Z*-Curve Components	Algorithm, Software or Database	Life Domains or Virus	Species
Comparative genomics, genome visualization	*x, y, z, z'*	*Z*-curve database [117]	Bacteria, archaea, eukaryotes and viruses	*Bacillus cereus* [103], *Bacillus cereus* ATCC 10987 [119], Coronavirus [118], human immunodeficiency virus [120], human [121, 143], *E. coli* [122], Seven GC-rich bacteria [126], 90 species [1], *Aeropyrum pernix* K1 [124], *Streptomyces coelicolor* [125]

[a] For some genomes, e.g., those of the bacterium *Mycobacterium tuberculosis* H37Ra [24] and Me Tri virus [28], ZCURVE was the only one used for genome annotation; in most cases, however, protein-coding gene recognition was performed by combining results of ZCURVE with those of others [31, 32], such as Glimmer [29] and Genmark [30]. It is noteworthy that ZCURVE is especially suitable for genomes with high GC content, e.g., GC content > 56% [12].

[b] RY, MK, AT and GC disparity curves correspond to *x*, *y*, (*x*+*y*)/2, (*x*-*y*)/2, respectively.

Phaeobacter gallaeciensis [22], *Desulfobacterium autotrophicum* [23], *Mycobacterium tuberculosis* [24], *Magnetospirillum gryphiswaldense* [25] and *Beggiatoa* [26]. ZCURVE 1.0 was also used for annotating archaeal genomes, e.g., archaea of the ANME-1 group [27] (Table **3**).

For some genomes, e.g., those of the bacterium *Mycobacterium tuberculosis* H37Ra [24] and *Me Tri* virus [28], ZCURVE 1.0 was the only software used for genome annotation, more frequently, however, results of ZCURVE 1.0 were combined with those of others, such as Glimmer [29] and Genmark [30]. For instance, ZCURVE 1.0 is integrated into meta-gene-finding tool YACOP [31] and GARSA [32]. It is noteworthy that ZCURVE 1.0 is especially suitable for genomes with high GC contents, e.g., GC content > 56% [12]. Likewise, ZCURVE_V and ZCURVE_CoV have been widely used for annotating protein-coding genes in newly sequenced genomes of viruses, coronaviruses [28, 33-37] and SARS coronaviruses [38-53].

3.1.2. Gene-finding in Eukaryotic Genomes

Algorithms based on the *Z*-curve theory have been used for recognizing protein coding genes in a number of eukaryotic genomes, e.g., the budding yeast genome [11], *Leptospira interrogans* genome [54] and *Drosophila* genomes [55]. The *Z*-curve algorithm has also been used in recognizing short coding sequences of human genes [56]. The algorithm based on the 189 *Z*-curve parameters was shown to be the most accurate among those tested for a given database, with the second one being an algorithm based on the Markov chain of order five [56], and the result was later confirmed by an independent study [57]. Recognition of exons and introns of human genes was also studied by using the *Z*-curve method [58].

3.1.3. Gene-finding Using the Fast Fourier Transform (FFT) Technique

The standard genetic code defines a mapping between a codon and an amino acid. According to this mapping, protein coding regions are divided into a series of tri-nucleotides (codon or triplet), resulting in a period-3 property in coding regions. Therefore, it is possible to find coding regions by exploring the 3-periodicity of DNA sequences. Consequently, the first step is to transform the DNA sequence into a digital sequence or signal, and the *Z*-curve is especially suitable for this purpose.

According to eqs. (**31**), $\Delta x_n = x_{n+1} - x_n = \pm 1$, $\Delta y_n = y_{n+1} - y_n = \pm 1$ and $\Delta z_n = z_{n+1} - z_n = \pm 1$. Applying the FFT to Δx, Δy and Δz, respectively, we are able to detect the 3-periodicity in the FFT power spectrum for each of the three numerical sequences. To increase the sensitivity, a lengthen-shuffle FFT algorithm was proposed for finding protein coding regions [59]. For example, the method was used to detect introns in the *C. elegance* chromosome III [60], and was later improved by using an adaptive filter to predict the exons in DNA sequences [61]. The relationship between the *Z*-curve and the Fourier transform for DNA sequence classification was studied in details [62].

3.2. Prediction of Replication Origins

3.2.1. Prediction of Replication Origins of Archaeal Genomes

Bacterial and eukaryotic genomes contain single and multiple replication origins, respectively. It was once a mystery whether archaea could have multiple *oriC*s.

Using the *Z*-curve method, we firstly predicted three *oriC*s as well as their precise locations for *Sulfolobus solfataricus* [63], and the prediction was consistent with later experimental evidence [64-67].

The archaeon *Methanococcus jannaschii* was the first to have its genome sequenced, however, its *oriCs* were notoriously difficult to locate by both theoretical and experimental methods. The *Z*-curve method predicted 2 *oriC*s [68] that were supported by later experimental evidence [66]. Similarly, we predicted a single *oriC* in the genome of *Methanosarcina mazei* [69] and 2 *oriC*s in the genome of *A. pernix* [70], which were also supported by experimental evidence [71]. The Z-curve method has been commonly used for annotating newly sequenced archaeal genomes, such as those of *Sulfolobus acidocaldarius* [72], *Haloferax volcanii* [73], *Desulfurococcus kamchatkensis* [74], *Thermococcus sibiricus* [75], and *Sulfolobus islandicus* [76].

3.2.2. Prediction of Replication Origins in Bacterial Genomes

The *Z*-curve method is an effective technique that detects the asymmetrical nucleotide distribution around replication origins. The *Z*-curve contains all the information of its corresponding DNA sequence, and therefore the GC-skew [77] is a special case of the *Z*-curve. Thus the *Z*-curve can reveal nucleotide asymmetry that is not detectable by GC skew [70]. For instance, RY, MK and AT disparity curves show an *oriC* in the archaeon *Methanosarcina mazei* Tuc01 (Fig. **4A**), while RY, MK, and GC disparity curves show an *oriC* in the bacterium *Salmonella enterica* tr. CT18 (Fig. **4B**).

Ori-Finder, an integrated *in silico* method to predict *oriC* regions of bacterial genomes, has been developed, based on the *Z*-curve method, along with distributions of DnaA box patterns, indicator genes, and phylogenetic relationships [78]. Ori-finder has become a commonly used annotation tool for identifying *oriC*s in newly sequenced archaeal and bacterial genomes, e.g., those of *Moraxella catarrhalis* [79], *Sorangium cellulosum* [80], *Microcystis aeruginosa* [80], *Cyanothece* [81], *Cupriavidus metallidurans* [82], *Azolla filiculoides* [83], *Variovorax paradoxus* [18], *Corynebacterium pseudotuberculosis* [84, 85], *Orientia tsutsugamushi* [86], *Propionibacterium freudenreichii* [87], *Laribacter hongkongensis* [88], *Legionella pneumophila* [89], and *Ehrlichia canis* [90] (Table **3**).

3.3. Studies of Genome Domain Structures

G+C content is an important characteristic of genome sequences. In the human genome, based on density gradient ultra-centrifugation experiments, it was found that long domains of relatively homogenous G+C content exist, and these domains are referred to as isochores [91, 92]. Traditionally, the G+C content along the genome is calculated using an overlapping or non-overlapping sliding window technique, based on which, however, isochores are hard to identify [93]. We developed a windowless technique in G+C content calculation, the GC profile [9, 10], which was used to study isochore structures in genomes of human [94], mouse [95], *Arabidopsis thaliana* [96] and chicken [97]. Based on the GC profile, the technique of wavelet multi-resolution analysis was used to identify isochore boundaries in the human genome [98]. For instance, a clear domain structure is revealed by the GC profile in chromosome 11 of finch (Fig. **4C**). Other groups also used GC-Profile to study isochores in the pig genome [99] and to assess DNA curvature profiles for *Aspergillus fumigatus* [100].

3.4. Identification of Horizontally-transferred Genomic Islands in Bacterial Genomes

It is generally accepted that horizontal gene transfer (HGT) plays an important role throughout the genome evolution of prokaryotes, because HGT alters the genotype of a bacterium, and could potentially lead to new traits [101]. Genomic islands (GIs) contain clusters of horizontally transferred genes and therefore, identification of horizontally-transferred GIs is an important biological issue. Because the GC profile is sensitive to changes in GC content, it is a powerful tool in identifying GIs [102].

Based on the method of GC profile, GIs in many bacteria have been identified, e.g., *Bacillus cereus* [103], *Corynebacterium glutamicum* [104], *Corynebacterium efficiens* [105], *Vibrio vulnificus* CMCP6 [104], and *Rhodopseudomonas palustris* [106]. For instance, it was once believed that *R.*

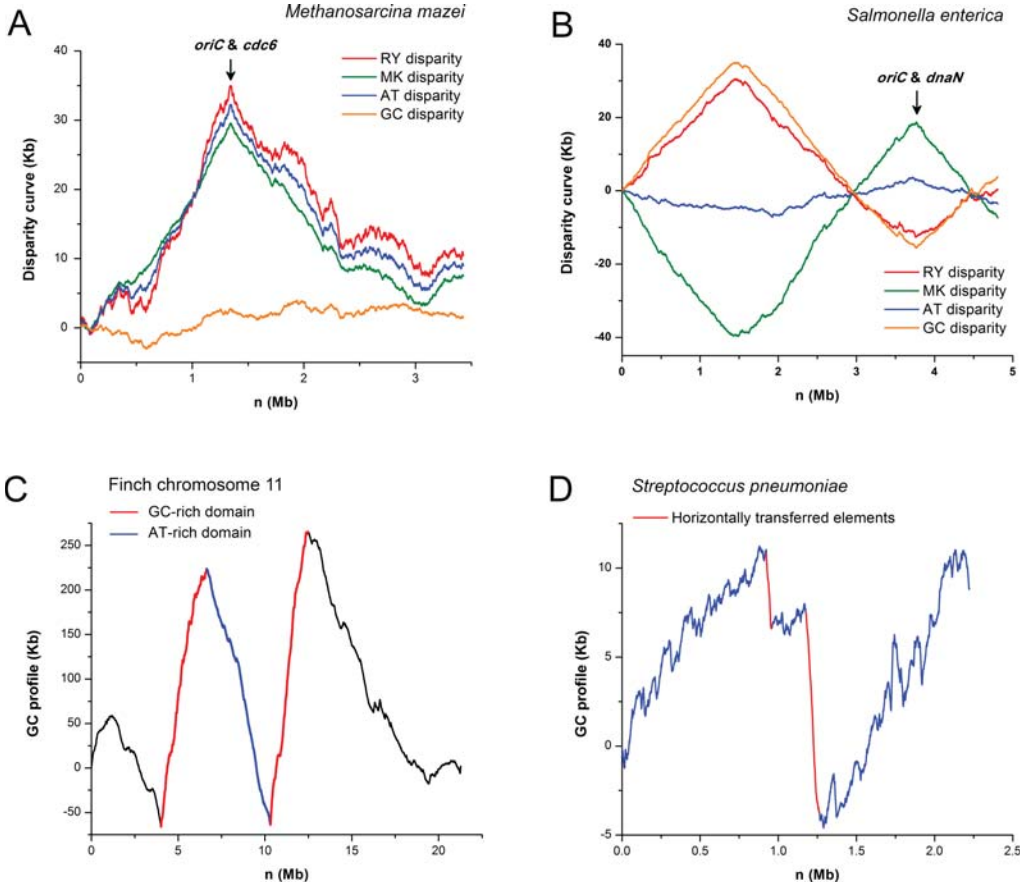

Fig. (4). The *Z*-curve reveals features of archaeal, bacterial and eukaryotic genomes. The *Z*-curve shows replication origins in genomes of **A**) the archaeon *Methanosarcina mazei* Tuc01 and **B**) the bacterium *Salmonella enterica* subsp. Typhi str. CT18. The *Z*-curve shows **C**) the domain structure in chromosome 11 of finch, and **D**) horizontally-transferred genomic elements in the genome of *Streptococcus pneumoniae* ATCC 700669.

palustris does not have GIs, but analysis based on the GC profile identified 3 GIs that help explain how this bacterium survives in a versatile environment [106]. C*orynebacterium efficiens* can grow and produce glutamate at temperature above 40°C; unexpectedly, however, an aspartate kinase is less thermostable. This kinase gene is located in a GI that we identified, and this result suggests an explanation for its being less thermostable, i.e., the adaptive mutations have not occurred extensively due to the recent HGT [105]. For instance, horizontally transferred elements in *Streptococcus pneumoniae* ATCC 700669 can be clearly shown by the GC profile (Fig. **4D**). The GC profile method has also been used for identification of GIs in other genomes, e.g., those of plant pathogens [107], *Streptomyces lividans* [108], *Parachlamydiaceae* UWE25 [109], *epsilon proteobacteria Sulfurovum and Nitratiruptor* [110], *Acinetobacter oleivorans* [111], and *Silicibacter pomeroyi* [112].

3.5. Identification of Promoters, Translation Start Sites and Nucleosome Positioning

Based on the behavior of the *Z*-curve near the bacterial gene translation start sites (TSS), a self-training method was proposed to find TSS with high accuracy [113]. It is likely that methods based on the same principle can also be used to recognize TSS in archaea and eukaryotes as well. Indeed, the *Z*-curve method was used to recognize human Pol II promoters [114] and promoters for bacterial genomes [115]. The positioning of nucleosomes, an elementary structural unit in eukaryotic chromatin, is pivotal in regulating many cellular processes, such as gene transcription. The *Z*-curve algorithm has been used to construct a genome-wide dynamic nucleosome positioning map for the budding yeast [116].

3.6. Visualization of DNA Sequences, Comparative Genomics and the *Z*-curve Database

One of the aims for developing the *Z*-curve theory is to visualize DNA sequences. By using the *Z*-curve, features of related DNA sequences can be grasped quickly in a perceivable form [1, 2]. Therefore, we constructed the *Z*-curve database (www.zcurve.net), which contains *Z*-curves for currently available genomes, online *Z*-curve drawing tools and other *Z*-curve related software [117]. For instance, human chromosome 6 and chimpanzee chromosome 6 are homologous, and they apparently have similar *Z*-curve patterns (Fig. **5A** and **B**). A typical example is the visualization of the ge-

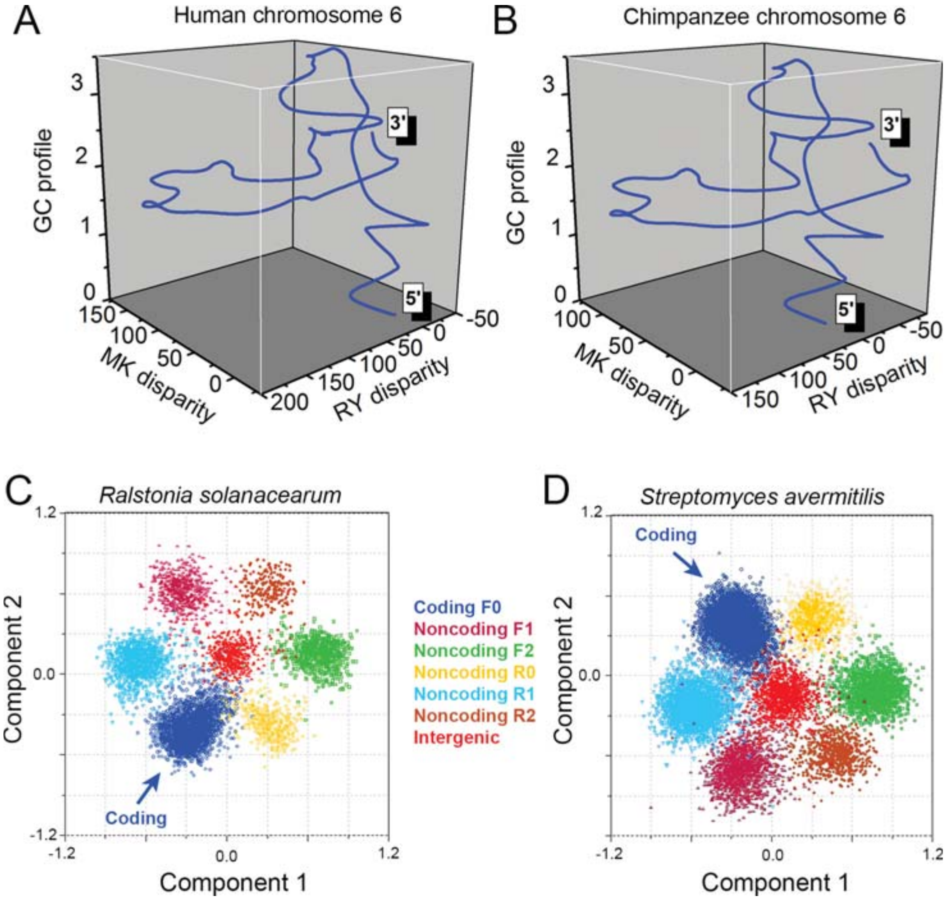

Fig. (5). Genomic nucleotide composition features revealed by the *Z*-curve method. 3-D *Z*-curves for human chromosome 6 (**A**) and chimpanzee chromosome 6 (**B**). The 2 homologous chromosomes show similar *Z*-curves. To show global nucleotide composition patterns, *Z*-curves have been smoothed for 50,000 times by using the B-spline function. An ORF-flower phenomenon is revealed by the *Z*-curve method in genomes with high GC content. All open reading frames are mapped onto a 9-dimensional space using the *Z*-curve method, and protein-coding ORFs are located in a distinct region, compared with non-coding ORFs and intergenic sequences. Shown are principal component analysis for the genomes of *Ralstonia solanacearum* GMI1000 (**C**) and *Streptomyces avermitilis* MA 4680 (**D**). F0, F1, F2, R0, R1, and R2 stand for reading frames of protein-coding, forward 1, forward 2, non-coding reverse 0, reverse 1 and reverse 2, respectively.

nomes of related SARS-coronaviruses. Based on the 3-D coordinates of the corresponding *Z*-curves, the phylogenetic tree was constructed and was found to be in agreement with that based on sequence alignment [118]. Comparative genomics based on the GC profile was used to identify genomic islands [103, 119].

According to eqs. (**6**), the base composition of a DNA sequence can be represented by a point in a 3-D space, thus providing an intuitive method to display base compositions. This method was used to study the codon usage in the genomes of AIDS virus [120], human [121], *E. coli* [122], *Vibrio cholerae* [123], *Aeropyrum pernix* K1 [124], *Streptomyces coelicolor* A3(2) [125] and seven GC-rich bacteria [126]. In prokaryotic genomes with high-GC content, coding ORFs and non-coding ORFs are located in distinct regions in a 9-dimensional space revealed by the *Z*-curve method, forming a flower-like pattern (Fig. **5C** and **D**).

4. SUMMARY

The three components of the *Z*-curve, *x*, *y* and *z*, which display distributions of purine/pyrimidine (R/Y), amino/keto (M/K) and strong-H bond/weak-H bond (S/W) bases, respectively, are independent, and completely describe the DNA sequence. The *x* and *y* components are related to the disparities of RY, MK, AT and GC bases, and can therefore be used to identify *oriC* regions in prokaryotic and eukaryotic genomes. The component *z* is related to G+C content, and can therefore be used to identify domain structures of eukaryotic genomes and genomic islands of prokaryotic genomes. The set of all three components can be used in identifications of protein-coding genes, promoters, translational start sites or in other bioinformatics issues. Generally, further applications are expected to benefit from the use of functions based on the three components, i.e., $f(x, y, z)$, with potential integration of other parameters.

In conclusion, the methodology of the *Z*-curve provides a geometrical approach to analyzing genomic DNA sequences. Considerable progress in applying the *Z*-curve method has been achieved, and the *Z*-curve theory provides a solid basis for future developments.

CONFLICT OF INTEREST

The authors confirm that this article content has no conflicts of interest.

ACKNOWLEDGEMENTS

The authors cordially thank the students and staff in the laboratory for assistance during the course of *Z*-curve studies.

REFERENCES

[1] Zhang, C.T.; Zhang, R. Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic Acids Res.*, **1991**, *19* (22), 6313-6317.

[2] Zhang, R.; Zhang, C.T. *Z*-curves, an intuitive tool for visualizing and analyzing the DNA sequences. *J. Biomol. Struct. Dyn.*, **1994**, *11* (4), 767-782.

[3] Zhang, C.T. A symmetrical theory of DNA sequences and its applications. *J. Theor. Biol.*, **1997**, *187* (3), 297-306.

[4] Hamori, E.; Ruskin, J. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. *J. Biol. Chem.*, **1983**, *258* (2), 1318-1327.

[5] Lobry, J.R. A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie.*, **1996**, *78* (5), 323-326.

[6] Cornish-Bowden, A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res.*, **1985**, *13*(9), 3021-3030.

[7] Zhang, C.T.; Zhang, R. A nucleotide composition constraint of genome sequences. *Comput. Biol. Chem.*, **2004**, *28*(2), 149-153.

[8] Zhang, C.T.; Gao, F.; Zhang, R. Segmentation algorithm for DNA sequences. *Phys. Rev. E. Stat. Nonlin Soft Matter Phys.*, **2005**, *72* (4 Pt 1), 041917.

[9] Gao, F.; Zhang, C.T. GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. *Nucleic Acids Res.*, **2006**, *34*(Web Server issue), W686-691.

[10] Zhang, C.T.; Wang, J.; Zhang, R. A novel method to calculate the G+C content of genomic DNA sequences. *J. Biomol. Struct. Dyn.*, **2001**, *19* (2), 333-341.

[11] Zhang, C.T.; Wang, J. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the *Z*-curve. *Nucleic Acids Res.*, **2000**, *28* (14), 2804-2814.

[12] Guo, F.B.; Ou, H.Y.; Zhang, C.T. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res.*, **2003**, *31* (6), 1780-1789.

[13] Wang, J.; Zhang, C.T. Identification of protein-coding genes in the genome of Vibrio cholerae with more than 98% accuracy using occurrence frequencies of single nucleotides. *Eur. J. Biochem.*, **2001**, *268*(15), 4261-4268.

[14] Guo, F.B.; Zhang, C.T. ZCURVE_V: a new self-training system for recognizing protein-coding genes in viral and phage genomes. *BMC Bioinform.*, **2006**, *7*, 9.

[15] Chen, L.L.; Ou, H.Y.; Zhang, R.; Zhang, C.T. ZCURVE_CoV: a new system to recognize protein coding genes in coronavirus genomes, and its applications in analyzing SARS-CoV genomes. *Biochem. Biophys. Res. Commun.*, **2003**, *307*(2), 382-388.

[16] Gao, F.; Ou, H.Y.; Chen, L.L.; Zheng, W.X.; Zhang, C.T. Prediction of proteinase cleavage sites in polyproteins of coronaviruses and its applications in analyzing SARS-CoV genomes. *FEBS Lett.*, **2003**, *553*(3), 451-456.

[17] Gao, F.; Wang, Y.; Liu, Y.J.; Wu, X.M.; Lv, X.; Gan, Y.R.; Song, S.D.; Huang, H. Genome sequence of Acinetobacter baumannii MDR-TJ. *J. Bacteriol.*, **2011**, *193*(9), 2365-2366.

[18] Han, J.I.; Choi, H.K.; Lee, S.W.; Orwin, P.M.; Kim, J.; Laroe, S.L.; Kim, T.G.; O'Neil, J.; Leadbetter, J.R.; Lee, S.Y.; Hur, C.G.; Spain, J.C.; Ovchinnikova, G.; Goodwin, L.; Han, C. Complete genome sequence of the metabolically versatile plant growth-promoting endophyte Variovorax paradoxus S110. *J. Bacteriol.*, **2011**, *193*(5), 1183-1190.

[19] Zhao, W.; Zhong, Y.; Yuan, H.; Wang, J.; Zheng, H.; Wang, Y.; Cen, X.; Xu, F.; Bai, J.; Han, X.; Lu, G.; Zhu, Y.; Shao, Z.; Yan, H.; Li, C.; Peng, N.; Zhang, Z.; Zhang, Y.; Lin, W.; Fan, Y.; Qin, Z.; Hu, Y.; Zhu, B.; Wang, S.; Ding, X.; Zhao, G.P. Complete genome sequence of the rifamycin SV-producing Amycolatopsis mediterranei U32 revealed its genetic characteristics in phylogeny and metabolism. *Cell Res.*, **2010**, *20*(10), 1096-1108.

[20] He, J.; Shao, X.; Zheng, H.; Li, M.; Wang, J.; Zhang, Q.; Li, L.; Liu, Z.; Sun, M.; Wang, S.; Yu, Z. Complete genome sequence of Bacillus thuringiensis mutant strain BMB171. *J. Bacteriol.*, **2010**, *192*(15), 4074-4075.

[21] Lopez, P.; Hornung, A.; Welzel, K.; Unsin, C.; Wohlleben, W.; Weber, T.; Pelzer, S. Isolation of the lysolipin gene cluster of Streptomyces tendae Tu 4042. *Gene*, **2010**, *461*(1-2), 5-14.

[22] Zech, H.; Thole, S.; Schreiber, K.; Kalhofer, D.; Voget, S.; Brinkhoff, T.; Simon, M.; Schomburg, D.; Rabus, R. Growth phase-dependent global protein and metabolite profiles of Phaeobacter gallaeciensis strain DSM 17395, a member of the marine Roseobacter-clade. *Proteomics*, **2009**, *9*(14), 3677-3697.

[23] Strittmatter, A.W.; Liesegang, H.; Rabus, R.; Decker, I.; Amann, J.; Andres, S.; Henne, A.; Fricke, W.F.; Martinez-Arias, R.; Bartels, D.; Goesmann, A.; Krause, L.; Puhler, A.; Klenk, H.P.; Richter, M.; Schuler, M.; Glockner, F.O.; Meyerdierks, A.; Gottschalk, G.; Amann, R. Genome sequence of Desulfobacterium autotrophicum HRM2, a marine sulfate reducer oxidizing organic carbon completely to carbon dioxide. *Environ. Microbiol.*, **2009**, *11*(5), 1038-1055.

[24] Zheng, H.; Lu, L.; Wang, B.; Pu, S.; Zhang, X.; Zhu, G.; Shi, W.; Zhang, L.; Wang, H.; Wang, S.; Zhao, G.; Zhang, Y. Genetic basis of virulence attenuation revealed by comparative genomic analysis of Mycobacterium tuberculosis strain H37Ra versus H37Rv. *PLoS ONE,* **2008**, *3*(6), e2375.

[25] Richter, M.; Kube, M.; Bazylinski, D.A.; Lombardot, T.; Glockner, F.O.; Reinhardt, R.; Schuler, D. Comparative genome analysis of four magnetotactic bacteria reveals a complex set of group-specific genes implicated in magnetosome biomineralization and function. *J. Bacteriol.,* **2007**, *189*(13), 4899-4910.

[26] Mussmann, M.; Hu, F.Z.; Richter, M.; de Beer, D.; Preisler, A.; Jorgensen, B.B.; Huntemann, M.; Glockner, F.O.; Amann, R.; Koopman, W.J.; Lasken, R.S.; Janto, B.; Hogg, J.; Stoodley, P.; Boissy, R.; Ehrlich, G.D. Insights into the genome of large sulfur bacteria revealed by analysis of single filaments. *PLoS Biol.,* **2007**, *5*(9), e230.

[27] Meyerdierks, A.; Kube, M.; Kostadinov, I.; Teeling, H.; Glockner, F.O.; Reinhardt, R.; Amann, R. Metagenome and mRNA expression analyses of anaerobic methanotrophic archaea of the ANME-1 group. *Environ. Microbiol.,* **2010**, *12*(2), 422-439.

[28] Tan le, V.; Ha do, Q.; Hien, V.M.; van der Hoek, L.; Farrar, J.; de Jong, M.D. Me Tri virus: a Semliki Forest virus strain from Vietnam? *J. Gen. Virol.,* **2008**, *89*(Pt 9), 2132-2135.

[29] Delcher, A.L.; Harmon, D.; Kasif, S.; White, O.; Salzberg, S.L. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.,* **1999**, *27*(23), 4636-4641.

[30] Besemer, J.; Lomsadze, A.; Borodovsky, M. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.,* **2001**, *29*(12), 2607-2618.

[31] Tech, M.; Merkl, R. YACOP: Enhanced gene prediction obtained by a combination of existing methods. *In Silico Biol.,* **2003**, *3*(4), 441-451.

[32] Dávila, A.M.R.; Lorenzini, D.M.; Mendes, P.N.; Satake, T.S.; Sousa, G.R.; Campos, L.M.; Mazzoni, C.J.; Wagner, G.; Pires, P.F.; Grisard, E.C.; Cavalcanti, M.C.R.; Campos, M.L.M. GARSA: genomic analysis resources for sequence annotation. *Bioinform.,* **2005**, *21*(23), 4302-4303.

[33] Lan, S.F.; Huang, C.H.; Chang, C.H.; Liao, W.C.; Lin, I.H.; Jian, W.N.; Wu, Y.G.; Chen, S.Y.; Wong, H.C. Characterization of a new plasmid-like prophage in a pandemic Vibrio parahaemolyticus O3:K6 strain. *Appl. Environ. Microbiol.,* **2009**, *75*(9), 2659-2667.

[34] Pyrc, K.; Berkhout, B.; van der Hoek, L. The novel human coronaviruses NL63 and HKU1. *J. Virol.,* **2007**, *81*(7), 3051-3057.

[35] Tang, X.C.; Zhang, J.X.; Zhang, S.Y.; Wang, P.; Fan, X.H.; Li, L.F.; Li, G.; Dong, B.Q.; Liu, W.; Cheung, C.L.; Xu, K.M.; Song, W.J.; Vijaykrishna, D.; Poon, L.L.; Peiris, J.S.; Smith, G.J.; Chen, H.; Guan, Y. Prevalence and genetic diversity of coronaviruses in bats from China. *J. Virol.,* **2006**, *80*(15), 7481-7490.

[36] Chu, D.K.; Peiris, J.S.; Chen, H.; Guan, Y.; Poon, L.L. Genomic characterizations of bat coronaviruses (1A, 1B and HKU8) and evidence for co-infections in Miniopterus bats. *J. Gen. Virol.,* **2008**, *89*(Pt 5), 1282-1287.

[37] van der Hoek, L.; Pyrc, K.; Jebbink, M.F.; Vermeulen-Oost, W.; Berkhout, R.J.; Wolthers, K.C.; Wertheim-van Dillen, P.M.; Kaandorp, J.; Spaargaren, J.; Berkhout, B. Identification of a new human coronavirus. *Nat. Med.,* **2004**, *10*(4), 368-373.

[38] Tan, Y.J.; Goh, P.Y.; Fielding, B.C.; Shen, S.; Chou, C.F.; Fu, J.L.; Leong, H.N.; Leo, Y.S.; Ooi, E.E.; Ling, A.E.; Lim, S.G.; Hong, W. Profiles of antibody responses against severe acute respiratory syndrome coronavirus recombinant proteins and their potential use as diagnostic markers. *Clin. Diagn. Lab. Immunol.,* **2004**, *11*(2), 362-371.

[39] Zhang, H.Z.; Zhang, H.; Kemnitzer, W.; Tseng, B.; Cinatl, J. Jr.; Michaelis, M.; Doerr, H.W.; Cai, S.X. Design and synthesis of dipeptidyl glutaminyl fluoromethyl ketones as potent severe acute respiratory syndrome coronovirus (SARS-CoV) inhibitors. *J. Med. Chem.,* **2006**, *49*(3), 1198-1201.

[40] Chen, L.; Gui, C.; Luo, X.; Yang, Q.; Günther, S.; Scandella, E.; Drosten, C.; Bai, D.; He, X.; Ludewig, B.; Chen, J.; Luo, H.; Yang, Y.; Yang, Y.; Zou, J.; Thiel, V.; Chen, K.; Shen, J.; Shen, X.; Jiang, H. Cinanserin Is an Inhibitor of the 3C-Like Proteinase of Severe Acute Respiratory Syndrome Coronavirus and Strongly Reduces Virus Replication *In Vitro*. *J. Virol.,* **2005**, *79*(11), 7095-7103.

[41] Chen, S.; Hu, T.; Zhang, J.; Chen, J.; Chen, K.; Ding, J.; Jiang, H.; Shen, X. Mutation of Gly-11 on the Dimer Interface Results in the Complete Crystallographic Dimer Dissociation of Severe Acute Respiratory Syndrome Coronavirus 3C-like Protease. *J. Biol. Chem.,* **2008**, *283*(1), 554-564.

[42] Chen, S.; Luo, H.; Chen, L.; Chen, J.; Shen, J.; Zhu, W.; Chen, K.; Shen, X.; Jiang, H. An overall picture of SARS coronavirus (SARS-CoV) genome-encoded major proteins: structures, functions and drug development. *Curr. Pharm. Des.,* **2006**, *12*(35), 4539-4553.

[43] Fan, K.; Ma, L.; Han, X.; Liang, H.; Wei, P.; Liu, Y.; Lai, L. The substrate specificity of SARS coronavirus 3C-like proteinase. *Biochem. Biophy. Res. Commun.,* **2005**, *329*(3), 934-940.

[44] Fang, S.; Shen, H.; Wang, J.; Tay, F.P.L.; Liu, D.X. Functional and Genetic Studies of the Substrate Specificity of Coronavirus Infectious Bronchitis Virus 3C-Like Proteinase. *J. Virology.,* **2010**, *84*(14), 7325-7336.

[45] Fang, S.G.; Shen, H.; Wang, J.; Tay, F.P.L.; Liu, D.X. Proteolytic processing of polyproteins 1a and 1ab between non-structural proteins 10 and 11/12 of Coronavirus infectious bronchitis virus is dispensable for viral replication in cultured cells. *Virol.,* **2008**, *379*(2), 175-180.

[46] Goetz, D.H.; Choe, Y.; Hansell, E.; Chen, Y.T.; McDowell, M.; Jonsson, C.B.; Roush, W.R.; McKerrow, J.; Craik, C.S. Substrate Specificity Profiling and Identification of a New Class of Inhibitor for the Major Protease of the SARS Coronavirus†,‡. *Biochem.,* **2007**, *46*(30), 8744-8752.

[47] Han, Y.S.; Chang, G.G.; Juo, C.G.; Lee, H.J.; Yeh, S.H.; Hsu, J.T.A.; Chen, X. Papain-like protease 2 (PLP2) from severe acute respiratory syndrome coronavirus (SARS-CoV): Expression, purification, characterization, and inhibition. *Biochem.,* **2005**, *44*(30), 10349-10359.

[48] Joseph, J.S.; Saikatendu, K.S.; Subramanian, V.; Neuman, B.W.; Brooun, A.; Griffith, M.; Moy, K.; Yadav, M.K.; Velasquez, J.; Buchmeier, M.J.; Stevens, R.C.; Kuhn, P. Crystal Structure of Nonstructural Protein 10 from the Severe Acute Respiratory Syndrome Coronavirus Reveals a Novel Fold with Two Zinc-Binding Motifs. *J. Virol.,* **2006**, *80*(16), 7894-7901.

[49] Joseph, J.S.; Saikatendu, K.S.; Subramanian, V.; Neuman, B.W.; Buchmeier, M.J.; Stevens, R.C.; Kuhn, P. Crystal Structure of a Monomeric Form of Severe Acute Respiratory Syndrome Coronavirus Endonuclease nsp15 Suggests a Role for Hexamerization as an Allosteric Switch. *J. Virol.,* **2007**, *81*(12), 6700-6708.

[50] Kiemer, L.; Lund, O.; Brunak, S.; Blom, N. Coronavirus 3CLpro proteinase cleavage sites: possible relevance to SARS virus pathology. *BMC Bioinform.,* **2004**, *5*, 72-72.

[51] Lin, C.-W.; Tsai, C.-H.; Tsai, F.-J.; Chen, P.-J.; Lai, C.-C.; Wan, L.; Chiu, H.-H.; Lin, K.-H. Characterization of trans- and cis-cleavage activity of the SARS coronavirus 3CLpro protease: basis for the *in vitro* screening of anti-SARS drugs. *FEBS Lett.,* **2004**, *574*(1-3), 131-137.

[52] Sydnes, M.O.; Hayashi, Y.; Sharma, V.K.; Hamada, T.; Bacha, U.; Barrila, J.; Freire, E.; Kiso, Y. Synthesis of glutamic acid and glutamine peptides possessing a trifluoromethyl ketone group as SARS-CoV 3CL protease inhibitors. *Tetrahedron.,* **2006**, *62*(36), 8601-8609.

[53] Tian, X.; Lu, G.; Gao, F.; Peng, H.; Feng, Y.; Ma, G.; Bartlam, M.; Tian, K.; Yan, J.; Hilgenfeld, R.; Gao, G.F. Structure and Cleavage Specificity of the Chymotrypsin-Like Serine Protease (3CLSP/nsp4) of Porcine Reproductive and Respiratory Syndrome Virus (PRRSV). *J. Mol. Biol.,* **2009**, *392*(4), 977-993.

[54] Ren, S.-X.; Fu, G.; Jiang, X.-G.; Zeng, R.; Miao, Y.-G.; Xu, H.; Zhang, Y.-X.; Xiong, H.; Lu, G.; Lu, L.-F.; Jiang, H.-Q.; Jia, J.; Tu, Y.-F.; Jiang, J.-X.; Gu, W.-Y.; Zhang, Y.-Q.; Cai, Z.; Sheng, H.-H.; Yin, H.-F.; Zhang, Y.; Zhu, G.-F.; Wan, M.; Huang, H.-L.; Qian, Z.; Wang, S.-Y.; Ma, W.; Yao, Z.-J.; Shen, Y.; Qiang, B.-Q.; Xia, Q.-C.; Guo, X.-K.; Danchin, A.; Saint Girons, I.; Somerville, R.L.; Wen, Y.-M.; Shi, M.-H.; Chen, Z.; Xu, J.-G.; Zhao, G.-P. Unique physiological and pathogenic features of Leptospira interrogans revealed by whole-genome sequencing. *Nature,* **2003**, *422*(6934), 888-893.

[55] Lin, M.F.; Deoras, A.N.; Rasmussen, M.D.; Kellis, M. Performance and scalability of discriminative metrics for comparative gene identification in 12 Drosophila genomes. *PLoS Comput. Biol.,* **2008**, *4*(4), e1000067.

[56] Gao, F.; Zhang, C.T. Comparison of various algorithms for

recognizing short coding sequences of human genes. *Bioinfor- m.,* **2004**, *20*(5), 673-681.

[57] Saeys, Y.; Rouzé, P.; Van de Peer, Y. In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi and protists. *Bioinform.,* **2007**, *23*(4), 414-420.

[58] Wu, Y.; Liew, A.W.-C.; Yan, H.; Yang, M. Classification of short human exons and introns based on statistical features. *Phys. Rev. E,* **2003**, *67*(6), 061916.

[59] Yan, M.; Lin, Z.S.; Zhang, C.T. A new fourier transform approach for protein coding measure based on the format of the *Z*-curve. *Bioinform.,* **1998**, *14*(8), 685-690.

[60] Rushdi, A.; Tuqan, J. In Acoustics, Speech and Signal Processing, 2006. ICASSP **2006** Proceedings. **2006** IEEE International Conference on, **2006**; Vol. 2, pp II-II.

[61] Ma, B.; Zhu, Y.-S. In Bioinformatics and Biomedical Engineering, 2007. ICBBE 2007. The 1st International Conference on, **2007**, pp 188-191.

[62] Law, N.F.; Cheng, K.O.; Siu, W.C. On relationship of *Z*-curve and Fourier approaches for DNA coding sequence classification. *Bioinform.,* **2006**, *1*(7), 242-246.

[63] Zhang, R.; Zhang, C.T. Multiple replication origins of the archaeon Halobacterium species NRC-1. *Biochem. Biophys. Res. Commun.,* **2003**, *302*(4), 728-734.

[64] Robinson, N.P.; Dionne, I.; Lundgren, M.; Marsh, V.L.; Bernander, R.; Bell, S.D. Identification of Two Origins of Replication in the Single Chromosome of the Archaeon Sulfolobus solfataricus. *Cell,* **2004**, *116*(1), 25-38.

[65] Robinson, N.P.; Bell, S.D. Origins of DNA replication in the three domains of life. *The FEBS J.,* **2005**, *272*(15), 3757-3766.

[66] Lundgren, M.; Bernander, R. Archaeal cell cycle progress. *Curr. Opin. Microbiol.,* **2005**, *8*(6), 662-668.

[67] Soppa, J. From genomes to function: haloarchaea as model organisms. *Microbiol.,* **2006**, *152*(3), 585-590.

[68] Zhang, R.; Zhang, C.T. Identification of replication origins in the genome of the methanogenic archaeon, Methanocaldococcus jannaschii. *Extremophiles,* **2004**, *8*(3), 253-258.

[69] Zhang, R.; Zhang, C.T. Single replication origin of the archaeon Methanosarcina mazei revealed by the *Z*-curve method. *Biochem. Biophys. Res. Commun.,* **2002**, *297*(2), 396-400.

[70] Zhang, R.; Zhang, C.T. Identification of replication origins in archaeal genomes based on the *Z*-curve method. *Archaea.,* **2005**, *1*(5), 335-346.

[71] Robinson, N.P.; Bell, S.D. Extrachromosomal element capture and the evolution of multiple replication origins in archaeal chromosomes. *Proc. Natl. Acad. Sci.,* **2007**, *104*(14), 5806-5811.

[72] Chen, L.; Brügger, K.; Skovgaard, M.; Redder, P.; She, Q.; Torarinsson, E.; Greve, B.; Awayez, M.; Zibat, A.; Klenk, H.-P.; Garrett, R.A. The Genome of Sulfolobus acidocaldarius, a Model Organism of the Crenarchaeota. *J. Bacteriol.,* **2005**, *187*(14), 4992-4999.

[73] Norais, C.; Hawkins, M.; Hartman, A.L.; Eisen, J.A.; Myllykallio, H.; Allers, T. Genetic and Physical Mapping of DNA Replication Origins in Haloferax volcanii. *PLoS Genet,* **2007**, *3*(5), e77.

[74] Ravin, N.V.; Mardanov, A.V.; Beletsky, A.V.; Kublanov, I.V.; Kolganova, T.V.; Lebedinsky, A.V.; Chernyh, N.A.; Bonch-Osmolovskaya, E.A.; Skryabin, K.G. Complete Genome Sequence of the Anaerobic, Protein-Degrading Hyperthermophilic Crenarchaeon Desulfurococcus kamchatkensis. *J. Bacteriol.,* **2009**, *191*(7), 2371-2379.

[75] Mardanov, A.V.; Ravin, N.V.; Svetlitchnyi, V.A.; Beletsky, A.V.; Miroshnichenko, M.L.; Bonch-Osmolovskaya, E.A.; Skryabin, K.G. Metabolic Versatility and Indigenous Origin of the Archaeon Thermococcus sibiricus, Isolated from a Siberian Oil Reservoir, as Revealed by Genome Analysis. *App. Environ. Microbiol.,* **2009**, *75*(13), 4580-4588.

[76] Flynn, K.M.; Vohr, S.H.; Hatcher, P.J.; Cooper, V.S. Evolutionary Rates and Gene Dispensability Associate with Replication Timing in the Archaeon Sulfolobus islandicus. *Genom. Biol. Evol.,* **2010**, *2*, 859-869.

[77] Lobry, J.R. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol. Biol. Evol.,* **1996**, *13*(5), 660-665.

[78] Gao, F.; Zhang, C.T. Ori-Finder: a web-based system for finding oriCs in unannotated bacterial genomes. *BMC Bioinform.,* **2008**, *9*, 79.

[79] de Vries, S.P.; van Hijum, S.A.; Schueler, W.; Riesbeck, K.; Hays, J.P.; Hermans, P.W.; Bootsma, H.J. Genome analysis of Moraxella catarrhalis strain BBH18, [corrected] a human respiratory tract pathogen. *J. Bacteriol.,* **2010**, *192*(14), 3574-3583.

[80] Gao, F.; Zhang, C.T. Origins of replication in Sorangium cellulosum and Microcystis aeruginosa. *DNA Res.,* **2008**, *15*(3), 169-171.

[81] Gao, F.; Zhang, C.T. Origins of replication in Cyanothece 51142. *Proc. Natl. Acad. Sci. U. S. A.,* **2008**, *105*(52), E125; author reply E126-127.

[82] Janssen, P.J.; Van Houdt, R.; Moors, H.; Monsieurs, P.; Morin, N.; Michaux, A.; Benotmane, M.A.; Leys, N.; Vallaeys, T.; Lapidus, A.; Monchy, S.; Medigue, C.; Taghavi, S.; McCorkle, S.; Dunn, J.; van der Lelie, D.; Mergeay, M. The complete genome sequence of Cupriavidus metallidurans strain CH34, a master survivalist in harsh and anthropogenic environments. *PLoS ONE,* **2010**, *5*(5), e10433.

[83] Ran, L.; Larsson, J.; Vigil-Stenman, T.; Nylander, J.A.; Ininbergs, K.; Zheng, W.W.; Lapidus, A.; Lowry, S.; Haselkorn, R.; Bergman, B. Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. *PLoS ONE,* **2010**, *5*(7), e11486.

[84] Trost, E.; Ott, L.; Schneider, J.; Schroder, J.; Jaenicke, S.; Goesmann, A.; Husemann, P.; Stoye, J.; Dorella, F.A.; Rocha, F.S.; Soares Sde, C.; D'Afonseca, V.; Miyoshi, A.; Ruiz, J.; Silva, A.; Azevedo, V.; Burkovski, A.; Guiso, N.; Join-Lambert, O.F.; Kayal, S.; Tauch, A. The complete genome sequence of Corynebacterium pseudotuberculosis FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. *BMC Genom.,* **2010**, *11*, 728.

[85] Paul, D.; Bridges, S.M.; Burgess, S.C.; Dandass, Y.S.; Lawrence, M.L. Complete genome and comparative analysis of the chemolithoautotrophic bacterium Oligotropha carboxidovorans OM5. *BMC Genom.,* **2010**, *11*, 511.

[86] Nakayama, K.; Kurokawa, K.; Fukuhara, M.; Urakami, H.; Yamamoto, S.; Yamazaki, K.; Ogura, Y.; Ooka, T.; Hayashi, T. Genome comparison and phylogenetic analysis of Orientia tsutsugamushi strains. *DNA Res.,* **2010**, *17*(5), 281-291.

[87] Falentin, H.; Deutsch, S.M.; Jan, G.; Loux, V.; Thierry, A.; Parayre, S.; Maillard, M.B.; Dherbecourt, J.; Cousin, F.J.; Jardin, J.; Siguier, P.; Couloux, A.; Barbe, V.; Vacherie, B.; Wincker, P.; Gibrat, J.F.; Gaillardin, C.; Lortal, S. The complete genome of Propionibacterium freudenreichii CIRM-BIA1, a hardy actinobacterium with food and probiotic applications. *PLoS ONE,* **2010**, *5*(7), e11748.

[88] Lau, S.K.; Fan, R.Y.; Ho, T.C.; Wong, G.K.; Tsang, A.K.; Teng, J.L.; Chen, W.; Watt, R.M.; Curreem, S.O.; Tse, H.; Yuen, K.Y.; Woo, P.C. Environmental adaptability and stress tolerance of Laribacter hongkongensis: a genome-wide analysis. *Cell Biosci.,* **2011**, *1*(1), 22.

[89] Bryan, A.; Swanson, M.S. Oligonucleotides stimulate genomic alterations of Legionella pneumophila. *Mol. Microbiol.,* **2011**, *80*(1), 231-247.

[90] Wei, W.; Guo, F.B. Strong Strand Composition Bias in the Genome of Ehrlichia canis Revealed by Multiple Methods. *Open Microbiol. J.,* **2010**, *4*, 98-102.

[91] Macaya, G.; Thiery, J.-P.; Bernardi, G. An approach to the organization of eukaryotic genomes at a macromolecular level. *J. Mol. Biol.,* **1976**, *108*(1), 237-254.

[92] Cuny, G.; Soriano, P.; Macaya, G.; Bernardi, G. The major components of the mouse and human genomes. 1. Preparation, basic properties and compositional heterogeneity. *Eur. J. Biochem. FEBS.,* **1981**, *115*(2), 227-233.

[93] Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J.P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, N.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J.C.; Mungall, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R.H.; Wilson, R.K.; Hillier, L.W.; McPherson, J.D.; Marra, M.A.; Mardis, E.R.; Fulton, L.A.;

Chinwalla, A.T.; Pepin, K.H.; Gish, W.R.; Chissoe, S.L.; Wendl, M.C.; Delehaunty, K.D.; Miner, T.L.; Delehaunty, A.; Kramer, J.B.; Cook, L.L.; Fulton, R.S.; Johnson, D.L.; Minx, P.J.; Clifton, S.W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J.F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R.A.; Muzny, D.M.; Scherer, S.E.; Bouck, J.B.; Sodergren, E.J.; Worley, K.C.; Rives, C.M.; Gorrell, J.H.; Metzker, M.L.; Naylor, S.L.; Kucherlapati, R.S.; Nelson, D.L.; Weinstock, G.M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Smith, D.R.; Doucette-Stamm, L.; Rubenfield, M.; Weinstock, K.; Lee, H.M.; Dubois, J.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R.W.; Federspiel, N.A.; Abola, A.P.; Proctor, M.J.; Myers, R.M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D.R.; Olson, M.V.; Kaul, R.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G.A.; Athanasiou, M.; Schultz, R.; Roe, B.A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W.R.; de la Bastide, M.; Dedhia, N.; Blocker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J.A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D.G.; Burge, C.B.; Cerutti, L.; Chen, H.C.; Church, D.; Clamp, M.; Copley, R.R.; Doerks, T.; Eddy, S.R.; Eichler, E.E.; Furey, T.S.; Galagan, J.; Gilbert, J.G.; Harmon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L.S.; Jones, T.A.; Kasif, S.; Kaspryzk, A.; Kennedy, S.; Kent, W.J.; Kitts, P.; Koonin, E.V.; Korf, I.; Kulp, D.; Lancet, D.; Lowe, T.M.; McLysaght, A.; Mikkelsen, T.; Moran, J.V.; Mulder, N.; Pollara, V.J.; Ponting, C.P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A.F.; Stupka, E.; Szustakowski, J.; Thierry-Mieg, D.; Thierry-Mieg, J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y.I.; Wolfe, K.H.; Yang, S.P.; Yeh, R.F.; Collins, F.; Guyer, M.S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K.A.; Patrinos, A.; Morgan, M.J.; de Jong, P.; Catanese, J.J.; Osoegawa, K.; Shizuya, H.; Choi, S.; Chen, Y.J. Initial sequencing and analysis of the human genome. *Nature,* **2001**, *409*(6822), 860-921.

[94] Zhang, C.T.; Zhang, R. An isochore map of the human genome based on the *Z*-curve method. *Gene,* **2003**, *317*(1-2), 127-135.

[95] Zhang, C.T.; Zhang, R. Isochore structures in the mouse genome. *Genom.,* **2004**, *83*(3), 384-394.

[96] Zhang, R.; Zhang, C.T. Isochore structures in the genome of the plant Arabidopsis thaliana. *J. Mol. Evol.,* **2004**, *59*(2), 227-238.

[97] Gao, F.; Zhang, C.T. Isochore structures in the chicken genome. *FEBS J.,* **2006**, *273*(8), 1637-1648.

[98] Wen, S.Y.; Zhang, C.T. Identification of isochore boundaries in the human genome using the technique of wavelet multiresolution analysis. *Biochem. Biophys. Res. Commun.,* **2003**, *311*(1), 215-222.

[99] Zhang, W.; Wu, W.; Lin, W.; Zhou, P.; Dai, L.; Zhang, Y.; Huang, J.; Zhang, D. Deciphering heterogeneity in pig genome assembly Sscrofa9 by isochore and isochore-like region analyses. *PLoS ONE,* **2010**, *5*(10), e13303.

[100] Do, J.H.; Miyano, S. The GC and window-averaged DNA curvature profile of secondary metabolite gene cluster in Aspergillus fumigatus genome. *Appl. Microbiol. Biotechnol.,* **2008**, *80*(5), 841-847.

[101] Ochman, H.; Lawrence, J.G.; Groisman, E.A. Lateral gene transfer and the nature of bacterial innovation. *Nature,* **2000**, *405*(6784), 299-304.

[102] Charkowski, A.O. Making sense of an alphabet soup: the use of a new bioinformatics tool for identification of novel gene islands. Focus on "Identification of genomic islands in the genome of Bacillus cereus by comparative analysis with Bacillus anthracis". *Physiol. Genom.,* **2004**, *16*(2), 180-181.

[103] Zhang, R.; Zhang, C.T. Identification of genomic islands in the genome of Bacillus cereus by comparative analysis with Bacillus anthracis. *Physiol. Genom.,* **2003**, *16*(1), 19-23.

[104] Zhang, R.; Zhang, C.T. A systematic method to identify genomic islands and its applications in analyzing the genomes of Corynebacterium glutamicum and Vibrio vulnificus CMCP6 chromosome I. *Bioinform.,* **2004**, *20*(5), 612-622.

[105] Zhang, R.; Zhang, C.T. Genomic islands in the Corynebacterium efficiens genome. *Appl. Environ. Microbiol.,* **2005**, *71*(6), 3126-3130.

[106] Zhang, C.T.; Zhang, R. Genomic islands in Rhodopseudomonas palustris. *Nat. Biotechnol.,* **2004**, *22*(9), 1078-1079.

[107] Chen, L.L. Identification of genomic islands in six plant pathogens. *Gene,* **2006**, *374*, 134-141.

[108] Jayapal, K.P.; Lian, W.; Glod, F.; Sherman, D.H.; Hu, W.S. Comparative genomic hybridizations reveal absence of large Streptomyces coelicolor genomic islands in Streptomyces lividans. *BMC Genom.,* **2007**, *8*, 229.

[109] Greub, G.; Collyn, F.; Guy, L.; Roten, C.A. A genomic island present along the bacterial chromosome of the Parachlamydiaceae UWE25, an obligate amoebal endosymbiont, encodes a potentially functional F-like conjugative DNA transfer system. *BMC Microbiol.,* **2004**, *4*, 48.

[110] Nakagawa, S.; Takaki, Y.; Shimamura, S.; Reysenbach, A.L.; Takai, K.; Horikoshi, K. Deep-sea vent epsilon-proteobacterial genomes provide insights into emergence of pathogens. *Proc. Natl. Acad. Sci. U. S. A.,* **2007**, *104*(29), 12146-12150.

[111] Jung, J.; Madsen, E.L.; Jeon, C.O.; Park, W. Comparative genomic analysis of Acinetobacter oleivorans DR1 to determine strain-specific genomic regions and gentisate biodegradation. *Appl Environ. Microbiol.,* **2011**, *77*(20), 7418-7424.

[112] Yan, D.Z.; Kang, J.X.; Liu, D.Q. Genomic analysis of the aromatic catabolic pathways from Silicibacter pomeroyi DSS-3. *Ann. Microbiol.,* **2009**, *59*(4), 789-800.

[113] Ou, H.Y.; Guo, F.B.; Zhang, C.T. GS-Finder: a program to find bacterial gene start sites with a self-training method. *Int. J. Biochem. Cell. Biol.,* **2004**, *36*(3), 535-544.

[114] Yang, J.Y.; Zhou, Y.; Yu, Z.G.; Anh, V.; Zhou, L.Q. Human Pol II promoter recognition based on primary sequences and free energy of dinucleotides. *BMC Bioinform.,* **2008**, *9*, 113.

[115] Song, K. Recognition of prokaryotic promoters based on a novel variable-window *Z*-curve method. *Nucleic Acids Res.,* **2012**, *40*(3), 963-971.

[116] Wu, X.; Liu, H.; Su, J.; Lv, J.; Cui, Y.; Wang, F.; Zhang, Y. *Z*-curve theory-based analysis of the dynamic nature of nucleosome positioning in Saccharomyces cerevisiae. *Gene,* **2013**, *530*(1), 8-18.

[117] Zhang, C.T.; Zhang, R.; Ou, H.Y. The *Z*-curve database: a graphic representation of genome sequences. *Bioinform.,* **2003**, *19*(5), 593-599.

[118] Zheng, W.X.; Chen, L.L.; Ou, H.Y.; Gao, F.; Zhang, C.T. Coronavirus phylogeny based on a geometric approach. *Mol. Phylogenet. Evol.,* **2005**, *36*(2), 224-232.

[119] Zhang, R.; Zhang, C.T. Accurate localization of the integration sites of two genomic islands at single-nucleotide resolution in the genome of Bacillus cereus ATCC 10987. *Comp. Funct. Genom.,* **2008**, *1*, 451930.

[120] Chou, K.C.; Zhang, C.T. Diagrammatization of codon usage in 339 human immunodeficiency virus proteins and its biological implication. *AIDS Res Hum Retroviruses,* **1992**, *8*(12), 1967-1976.

[121] Zhang, C.T.; Zhan, Y. Analysis on the distribution of bases in 1487 human protein coding sequences. *J. Theor. Biol.,* **1994**, *167*(2), 161-166.

[122] Zhang, C.T.; Chou, K.C. A graphic approach to analyzing codon usage in 1562 Escherichia coli protein coding sequences. *J. Mol. Biol.,* **1994**, *238*(1), 1-8.

[123] Wang, J.; Zhang, C.T. Analysis of the codon usage pattern in the Vibrio cholerae genome. *J. Biomol. Struct. Dyn.,* **2001**, *18*(6), 872-880.

[124] Guo, F.B.; Wang, J.; Zhang, C.T. Gene recognition based on nucleotide distribution of ORFs in a hyper-thermophilic crenarchaeon, Aeropyrum pernix K1. *DNA Res.,* **2004**, *11*(6), 361-370.

[125] Ou, H.Y.; Guo, F.B.; Zhang, C.T. Analysis of nucleotide distribution in the genome of Streptomyces coelicolor A3(2) using the *Z*-curve method. *FEBS Lett.,* **2003**, *540*(1-3), 188-194.

[126] Chen, L.L.; Zhang, C.T. Seven GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages. *Biochem. Biophys. Res. Commun.,* **2003**, *306*(1), 310-317.

[127] Woebken, D.; Teeling, H.; Wecker, P.; Dumitriu, A.; Kostadinov, I.; Delong, E.F.; Amann, R.; Glockner, F.O. Fosmids of novel marine Planctomycetes from the Namibian and Oregon coast upwelling systems and their cross-comparison with planctomycete genomes. *ISME J.,* **2007**, *1*(5), 419-435.

[128] Jones, B.V.; Marchesi, J.R. Transposon-aided capture (TRACA) of plasmids resident in the human gut mobile metagenome. *Nat. Methods,* **2007**, *4*(1), 55-61.

[129] Wietzorrek, A.; Schwarz, H.; Herrmann, C.; Braun, V. The genome of the novel phage Rtp, with a rosette-like tail tip, is homologous to the genome of phage T1. *J. Bacteriol.,* **2006**, *188*(4), 1419-1436.

[130] Ren, S.X.; Fu, G.; Jiang, X.G.; Zeng, R.; Miao, Y.G.; Xu, H.; Zhang, Y.X.; Xiong, H.; Lu, G.; Lu, L.F.; Jiang, H.Q.; Jia, J.; Tu, Y.F.; Jiang, J.X.; Gu, W.Y.; Zhang, Y.Q.; Cai, Z.; Sheng, H.H.; Yin, H.F.; Zhang, Y.; Zhu, G.F.; Wan, M.; Huang, H.L.; Qian, Z.; Wang, S.Y.; Ma, W.; Yao, Z.J.; Shen, Y.; Qiang, B.Q.; Xia, Q.C.; Guo, X.K.; Danchin, A.; Saint Girons, I.; Somerville, R.L.; Wen, Y.M.; Shi, M.H.; Chen, Z.; Xu, J.G.; Zhao, G.P. Unique physiological and pathogenic features of Leptospira interrogans revealed by whole-genome sequencing. *Nature,* **2003**, *422*(6934), 888-893.

[131] Song, K.; Zhang, Z.; Tong, T.P.; Wu, F. Classifier assessment and feature selection for recognizing short coding sequences of human genes. *J. Comput. Biol.,* **2012**, *19*(3), 251-260.

[132] Gao, F.; Luo, H.; Zhang, C.T. DoriC 5.0: an updated database of oriC regions in both bacterial and archaeal genomes. *Nucleic Acids Res.,* **2013**, *41*(Database issue), D90-93.

[133] Gao, F.; Zhang, C.T. DoriC: a database of oriC regions in bacterial genomes. *Bioinform.,* **2007**, *23*(14), 1866-1867.

[134] Sabri, M.; Hauser, R.; Ouellette, M.; Liu, J.; Dehbi, M.; Moeck, G.; Garcia, E.; Titz, B.; Uetz, P.; Moineau, S. Genome annotation and intraviral interactome for the Streptococcus pneumoniae virulent phage Dp-1. *J. Bacteriol.,* **2011**, *193*(2), 551-562.

[135] Jain, P.K.; Kush, D.; Ramachandran, S.; Verma, S.K. Isolation and characterization of R-plasmid pPRS3a from Bacillus cereus GC subgroup a PRS3. *Int. J. Integr. Biol. Int. J. Integ. Biol.,* **2011**, *11*(1), 1-7.

[136] Dyall-Smith, M.L.; Pfeiffer, F.; Klee, K.; Palm, P.; Gross, K.; Schuster, S.C.; Rampp, M.; Oesterhelt, D. Haloquadratum walsbyi: limited diversity in a global pond. *PLoS ONE,* **2011**, *6*(6), e20968.

[137] Zheng, W.X.; Zhang, C.T. Biological implications of isochore boundaries in the human genome. *J. Biomol. Struct. Dyn.,* **2008**, *25*(4), 327-336.

[138] Gao, F.; Zhang, C.T. Prediction of replication time zones at single nucleotide resolution in the human genome. *FEBS Lett.,* **2008**, *582*(16), 2441-2444.

[139] Sela, D.A.; Chapman, J.; Adeuya, A.; Kim, J.H.; Chen, F.; Whitehead, T.R.; Lapidus, A.; Rokhsar, D.S.; Lebrilla, C.B.; German, J.B.; Price, N.P.; Richardson, P.M.; Mills, D.A. The genome sequence of Bifidobacterium longum subsp. infantis reveals adaptations for milk utilization within the infant microbiome. *Proc. Natl. Acad. Sci. U. S. A.,* **2008**, *105*(48), 18964-18969.

[140] Hernandez, A.; Lopez, J.C.; Santamaria, R.; Diaz, M.; Fernandez-Abalos, J.M.; Copa-Patino, J.L.; Soliveri, J. Xylan-binding xylanase Xyl30 from Streptomyces avermitilis: cloning, characterization, and overproduction in solid-state fermentation. *Int. Microbiol.,* **2008**, *11*(2), 133-141.

[141] McNally, R.R.; Toth, I.K.; Cock, P.J.; Pritchard, L.; Hedley, P.E.; Morris, J.A.; Zhao, Y.; Sundin, G.W. Genetic characterization of the HrpL regulon of the fire blight pathogen Erwinia amylovora reveals novel virulence factors. *Mol. Plant. Pathol.,* **2012**, *13*(2), 160-173.

[142] Ryan, M.P.; Pembroke, J.T.; Adley, C.C. Novel Tn4371-ICE like element in Ralstonia pickettii and genome mining for comparative elements. *BMC Microbiol.,* **2009**, *9*, 242.

[143] Zhang, C.T. In *Visualizing biological information*. Pickover, C.A. Ed.; World Scientific: Singapore; River Edge, N.J. **1995**, pp 84-95.