New Algorithm for Determining the Similarities of Nucleotide Sequences Based on Comparisons of Torsion of Discrete Curves

Rong CHEN¹,a.*, Bo XUE² and Yong-an XU²

¹Jiangsu Agri-animal Husbandry Vocational College, Taizhou 225300, China
²College of Information Engineering, Yangzhou University, Yangzhou 225009, China

a rongchen1984@163.com
*Corresponding author

Keywords: Curvature, Torsion, Discrete Curve, Comparison.

Abstract. Discrete curves can be used to characterize sequences and the complete coding information of DNA in three-dimensional space. When the curve possesses properties that, in a mapping relationship, mirror those of DNA sequences, its characteristic values reveal the biological characteristics of a DNA sequence. By describing a DNA sequence using the recent Z curve model, we here propose a new method to judge the similarities of discrete DNA curves based on comparisons of the torsion of the curves. This comparative method was verified by quantitatively analyzing the similarities of cDNA sequences of 10 types of avian influenza viruses. Compared with a traditional dynamic programming algorithm, this new method takes full advantage of computer geometry for genetic comparative analyses by shifting from the traditional approach of a direct comparison of gene sequence characteristics to a comparison of the similarity of spatially discrete curves.

Introduction

DNA sequences, the carriers of biological genetic information, record historical information of biological evolution occurring over billions of years and contain clues for understanding nature. With recent progress in sequencing technologies, the volume of available biological gene sequences is continuing to increase exponentially. However, only a small portion of these data can be obtained using conventional laboratory methods [1]. Modern biological and computational techniques have provided powerful comparative methods to determine the similarities among unknown and known sequences. Indeed, we gained much insight from comparisons and analyses of sequence homology based on computer arithmetic. Sequence similarity analyses have determined the similarities of two or more sequences to provide clue into their evolutionary relationships by comparisons of their similar structures or functions and by judging the level of homology among them. Furthermore, quantitative analyses can be performed to establish the degree of similarity among sequences. In 1970, Gibbs and McIntyre first proposed a sequence similarity analysis method based on an image display-dot matrix [2]. Subsequently, Nussinov identified areas of similarity based on a colorful dot matrix [3]. In 1981, Smith and Waterman proposed an optimal algorithm for sequence comparison known as a dynamic programming algorithm [4]; however, this algorithm is not suitable for use with a database search because it has relatively low time and spatial efficiency. In the last few decades, several types of heuristic algorithms

151
have been proposed and are now widely used for characterizing and comparing DNA sequences, including the most commonly applied algorithms FASTA and BLAST [5].

As an alternative approach, Zhang proposed the concept of applying the Z curve, which can reflect DNA sequence information onto a curve comprised of a sequential, discrete three-dimensional point sets, in three-dimensional space [6]. Indeed, several comparison algorithms have been derived based on the Z curve [7-9]. According to the characteristics of the curves, these comparison algorithms can be divided into three main groups: those based on those based on the curvature of the curve [10], the geometric center of the curve [11], and those based on the torsion of the curve [12].

Here, we propose a new computational algorithm based on the curvature of discrete curves and describe the trend features of the curves in space from the perspective of the discrete-point deviation of \( x = y = z \). We demonstrate the application of this algorithm for determining the similarities of DNA sequences through analysis of the characteristics of the curvature of the Z curve.

**Methods**

**Curves and Their Geometrical Characteristics**

**Definition of the Z curve.** Zhang defined the Z curve of a gene sequence based on a regular tetrahedron-hexahedron system. As shown in Fig. 1, starting from analysis of the symmetry of the four base groups of a DNA sequence, A, C, G, and T correspond to the four vertex angles of a regular tetrahedron, and the six faces of the regular hexahedron are divided into three pairs: (1) the upper and bottom faces, corresponding to strong and weak hydrogen bonds; (2) the left and right faces, corresponding to purines and pyrimidines; and (3) the front and back faces, corresponding to amidogens and ketones. This system accurately illustrates the symmetry of the four types of base groups in a DNA sequence.

![Figure 1. Regular tetrahedron-hexahedron system.](image)

Therefore, taking one strand from a DNA double helix sequence, and setting the total amount of base groups as \( n \), the amounts of A, C, G, and T are respectively represented as \( A_n, C_n, G_n, \) and \( T_n \), which are all positive integers: \( A_n + C_n + G_n + T_n = n \). When \( n \) is fixed, three possibilities of \( A_n + C_n + G_n + T_n = n \) are independent, which can be represented as points in a three-dimensional space coordinate system. The sum of the distances from any point of the regular tetrahedron to the four faces is equal to the largest constant. Assuming the height of the regular tetrahedron to be \( n \), then the above four positive integers (\( A_n, C_n, G_n, \) and \( T_n \)) can be solely represented by the point \( P_n \) in
the regular tetrahedron. Considering a section of a DNA sequence with N base groups, when n increases from 1 to N, the heights of the N concentric regular tetrahedron strung together are separately 1, 2, …, N, and the points that show the relationships (A_n, C_n, G_n, and T_n) in the N concentric regular tetrahedron are P_1, P_2, …, P_N. The ligature of the N points is the Z curve of a DNA sequence.

To perform quantitative research, Zhang et al. established a specific rectangular coordinate system within the tetrahedron-hexahedron system, such that the rectangular coordinate of each crunodes is:

\[
\begin{align*}
x_n &= 2(A_n + G_n) - n \\
y_n &= 2(A_n + C_n) - n \\
z_n &= 2(A_n + T_n) - n
\end{align*}
\]

\(x_n, y_n, z_n \in [-n,n], n = 1, 2, \ldots, N\)  

(1)

Eq. (1) can be used to represent any DNA sequence by a three-dimensional space curve, called the Z curve. Thus, the Z curve is a geometrical expression form in which a DNA sequence and the symbol form are completely equivalent. Therefore, analysis of a DNA sequence can be easily transferred to analysis of the Z curve, which contains all of the DNA sequence information.

As an example, Fig. 2a shows a gene sequence fragment of Newcastle virus, containing 300 base pairs. This sequence was transformed using Eq. (1). Fig. 2b presents the Z curve of Fig. 2a.

\[
\begin{align*}
\text{Cgagctaaagctacaccccagcagagagccgtgctgctgccccccgggtgtctctggctgctgctg} \\
\text{agctgacatgcccacctgatctttcctgttggctgtgatcttgatctgtgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatcttgatct tg
\end{align*}
\]

Figure 2. Gene sequence fragment (a) and Z curve of the gene sequence fragment (b) of Newcastle virus.

**Definition of the Torsion of a Discrete Curve**

In differential geometry of an elementary three-dimensional curve, the torsion of a curve (which is also called the twist ratio) is used to measure its degree of distortion. The measurement method that combines the curvature and torsion of a space curve is similar to that used for analysis of the curvature of a plane curve. The Z curve is comprised of a series of spatial discrete points. In this paper, we do not directly
mechanically apply the curvature of a continuous curve in a discrete curve, but rather provide an average curvature estimation method that can be used to draw the Z curve.

O'Rourke optimized a triangle grid based on the principle that the triangle area is the minimum area [13]. Furthermore, Sun et al. dissected a triangle grid by controlling the maximum value of the side length of the triangle [14]. To ensure reconstitution of the quality of the model, Shi et al. set the maximum interior angle of the triangle to be above 15° [15]. Among all methods available to estimate the average curvature of a triangle grid model, the Voronoi method proposed by Mayer et al. shows the best calculation effects on the average discrete curvature and Gaussian curvature [16]. The basic principle of this method is to regard the smooth hook face as a group of extremities or linear approximation values of the grid and to regard each vertex of the triangle grid as an average measurement value of a small voisinage of the grid space.

Accordingly, the average curvature of a discrete point $x_i$ is:

$$k_{avg}(x_i) = \frac{1}{4A_M} \sum_{j \in N_1(i)} \left( \cos \alpha_{ij} + \cos \beta_{ij} \right) (x_j - x_i) \cdot n$$

Where $N_1(i)$ is the adjacent vertex of $i$, $A_M$ is the mixture area, and $n$ is the normal vector of $x_i$. The definitions of $\alpha_{ij}$ and $\beta_{ij}$ are shown in Fig. 4A. According to the shape of a triangle, $A_M$ is either an oxygon or an obtuse or right triangle. In the case in which $A_M$ is an oxygon, the circumcenter of the triangle and the midpoint of the two related sides of $x_i$ are connected to obtain the shadow area shown in Fig. 4B. In the case of an obtuse or right triangle, if the obtuse or right angle is included in the angle of $x_i$, the midpoint of the corresponding side of the obtuse angle and the midpoint of the two related sides of $x_i$ are connected to obtain the shadow area shown in Fig. 4C; thus, $A_M = A_T/2$, where $A_T$ is the triangle area. If the obtuse or right angle is not in the angle of $x_i$, the midpoints of the two related sides of $x_i$ are directly connected to obtain the shadow area shown in Fig. 4D; thus, $A_M = A_T/4$.

Figure 4. Definitions of $\alpha_{ij}$ and $\beta_{ij}$ and different types of triangle shapes ($A_M$).
**Algorithm for the Similarities of Discrete Curves**

The average curvature or accumulated curvature will show the sole form of the Z curve in space. During the process of reflecting a DNA sequence with the Z curve in three-dimensional space, the initial position of the Z curve is (0, 0, 0). When one metric is increased for A, one vector quantity will be increased to the end of the current curve (1, 1, 1). In a similar way, when one metric is increased for T, C, or G, one vector quantity will be increased to the end of the current curve to obtain (-1, -1, 1), (-1, 1, -1), or (1, -1, -1), respectively.

For the purpose of this study, the key point is to confirm the ligature of the triangle grid. The computational formula of the Z curve is simply the accumulation of vector quantities. Therefore, three adjacent points will constitute a triangle when the continuous symbols in the gene sequence are not the same. According to the appearing sequence of points, the midpoint of three vertices of the triangle can be taken as the vertex of the included angle α, and the curvature at this moment can be represented by the included angle α. The curvature is smaller when the included angle α is larger, and vice versa. When two or more continuous symbols are identical, then no triangle will be constituted by three points; this triangle is replaced by a segment with the length. At this moment, the curvature is 0. The ligature figure of a DNA triangle grid is shown in Fig. 5.

![Figure 5. The triangle grid ligature of a DNA curve.](image)

In Fig. 5, the permutation trend of the hypothetical sequence ABCDE shows a progressive increase in the direction of the Z curve. The average curvature of point C is determined by the two adjacent triangles ΔACB and ΔBCD. Similarly, the average curvature of point B is determined by ΔABC and ΔCBD. Thus, the average curvature of each point in the Z curve of DNA can be calculated according to the average curvature formula of the discrete points. The curvatures of points that do not constitute a triangle should be directly replaced by 0.

Therefore, the curvature can be calculated by adding all of the angles of the triangles. The curve is flatter when the sum of the angles is larger. The cosine formula will not be adopted to facilitate calculation. \( \cos \alpha = \frac{|b + mc + na|}{a} \), because when \( c \) or \( b \) is larger than 1, which is proportional to the adjacent side and inversely proportional to the opposite side, \( (l, m, n) \) is the coefficient, where \( n \) is less than 0. By setting \( (l, m, n) = (1, 2, -2) \), the accumulated curvatures of two gene sequences S and T on the Z curve are then \( K_S \) and \( K_T \), respectively; thus, \( \text{dist}K = ||(K_S, K_T)||_2 \) can be calculated to express the affinity according to the 2-norm of the curvature. The specific algorithm used for calculating \( \text{dist}K \) is as follows:

1: Parse two DNA sequences into two Z curves, S and T;
2: for i ∈ \{0, ..., m(S) \} {
3: K_S = sum I_i + mc + na;
4: }
5: for i ∈ \{0, ..., m(T) \} {
6: K_T = sum I_i + mc + na;
7: }
8: Calculate distK;

Result and Discussion

To test this algorithm, we used actual data of homologous cDNA sequences of a group
of avian influenza viruses obtained from the National Center for Biotechnology
Information database. The details of the sequences and strains are shown in Table 1.

Table 1. Sequences and lengths of cDNA of 10 types of avian influenza viruses.

<table>
<thead>
<tr>
<th>Virus number</th>
<th>Length(bp)</th>
<th>Partial cDNA sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB124607</td>
<td>1470</td>
<td>atgtcgtatgggaattcagaacgctctctcgtctgtaga…</td>
</tr>
<tr>
<td>AB124606</td>
<td>1470</td>
<td>atgtcgtatgtggaattcagaacgctctctcgtctgtaga…</td>
</tr>
<tr>
<td>AB124605</td>
<td>1470</td>
<td>atgtcgtatgtggaattcagaacgctctctcgtctgtaga…</td>
</tr>
<tr>
<td>AB124604</td>
<td>1470</td>
<td>atgtcgtatgtggaattcagaacgctctctcgtctgtaga…</td>
</tr>
<tr>
<td>AB124603</td>
<td>1470</td>
<td>atgtcgtatgtggaattcagaacgctctctcgtctgtaga…</td>
</tr>
<tr>
<td>AB124602</td>
<td>1470</td>
<td>atgtcgtatgtggaattcagaacgctctctcgtctgtaga…</td>
</tr>
<tr>
<td>AB124601</td>
<td>1470</td>
<td>atgtcgtatgtggaattcagaacgctctctcgtctgtaga…</td>
</tr>
<tr>
<td>AB124600</td>
<td>1470</td>
<td>atgtcgtatgtggaattcagaacgctctctcgtctgtaga…</td>
</tr>
<tr>
<td>AB124699</td>
<td>1470</td>
<td>atgtcgtatgtggaattcagaacgctctctcgtctgtaga…</td>
</tr>
<tr>
<td>AB124698</td>
<td>1470</td>
<td>atgtcgtatgtggaattcagaacgctctctcgtctgtaga…</td>
</tr>
</tbody>
</table>

Since the lengths of all sequences were equal to 1470 bp, the alignment condition did
not need to be considered during the sequence comparisons for this group.

The scores obtained after comparing the cDNA sequences of the virus AB124607 to
the 10 sequences (including itself) are shown in Table 2. In this table, the Needle score
refers to that obtained with the Needle-Wunsch dynamic programming algorithm [17].
The rule of penalty scores set in this experiment was as follows: a similar condition
yielded a score of 2, a different condition yielded a score of -1, and a blank insertion
yielded a score of -2. The results showed that the scores in the same row could be easily
compared, whereas scores in the same line were not as easily compared. After
normalizing the data for every line from 0 to 100, the sequence results were arranged
from large to small according to the Needle-Wunsch algorithm, and the results are
shown in Table 3.

Table 2. Comparison scores among avian virus cDNA sequences.

<table>
<thead>
<tr>
<th>Virus number</th>
<th>Needle</th>
<th>distK</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB124607</td>
<td>1470</td>
<td>2756.505</td>
</tr>
<tr>
<td>AB124606</td>
<td>1454</td>
<td>2755.368</td>
</tr>
<tr>
<td>AB124605</td>
<td>1352</td>
<td>2754.357</td>
</tr>
<tr>
<td>AB124604</td>
<td>1346</td>
<td>2653.474</td>
</tr>
<tr>
<td>AB124603</td>
<td>1314</td>
<td>2751.203</td>
</tr>
<tr>
<td>AB124602</td>
<td>1370</td>
<td>2754.989</td>
</tr>
<tr>
<td>AB124601</td>
<td>1186</td>
<td>2749.312</td>
</tr>
<tr>
<td>AB124600</td>
<td>1234</td>
<td>2750.068</td>
</tr>
<tr>
<td>AB124699</td>
<td>1290</td>
<td>2752.717</td>
</tr>
<tr>
<td>AB124698</td>
<td>1344</td>
<td>2751.581</td>
</tr>
</tbody>
</table>
Table 3. Comparison scores among avian virus cDNA sequences after normalization.

<table>
<thead>
<tr>
<th>Virus number</th>
<th>Needle</th>
<th>distK</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB124607</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>AB124606</td>
<td>94.366</td>
<td>84.193</td>
</tr>
<tr>
<td>AB124602</td>
<td>64.789</td>
<td>78.924</td>
</tr>
<tr>
<td>AB124605</td>
<td>56.463</td>
<td>70.138</td>
</tr>
<tr>
<td>AB124604</td>
<td>56.338</td>
<td>57.862</td>
</tr>
<tr>
<td>AB124698</td>
<td>55.633</td>
<td>31.545</td>
</tr>
<tr>
<td>AB124603</td>
<td>45.070</td>
<td>26.289</td>
</tr>
<tr>
<td>AB124699</td>
<td>36.620</td>
<td>47.338</td>
</tr>
<tr>
<td>AB124600</td>
<td>16.901</td>
<td>10.510</td>
</tr>
<tr>
<td>AB124601</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

After normalizing the results in Table 2, 100 indicated a close genetic relationship, and 0 indicated the farthest genetic relationship among the 10 samples. Therefore, the comparison of nucleic acid sequence similarity between different viruses was obvious.

**Conclusion**

The algorithms currently used for sequence comparisons are relatively mature, with continuous progress being made. Currently, many researchers are focused on developing methods for traditional sequence comparison analysis from the perspective of geometry. Indeed, several approaches have been proposed by researchers worldwide on translating gene sequence comparisons from a character sequence to a geometric figure. There are many ways to display a gene character sequence using a three-dimensional point set. Moreover, much progress has been made regarding research on analysis of the similarities of curves. However, few attempts have been made to combine these two aspects, namely, to study the similarities of gene sequences according to the similarities of their curves. Therefore, we have proposed a method to analyze the affinity of DNA sequences based on similarities of curvature among curves, which can be used to compare homologous genes of different species. The results showed that the overall trend obtained with the new algorithm was similar to the results obtained with a traditional dynamic programming algorithm, although there were some notable differences between them. The differences in these results clearly demonstrated that gene sequence comparisons had distinct characteristics.

**References**


