Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Coronavirus phylogeny based on a geometric approach

Wen-Xin Zheng a, Ling-Ling Chen a,b, Hong-Yu Ou a, Feng Gao a, Chun-Ting Zhang a,*

a Department of Physics, Tianjin University, Tianjin 300072, China
b Laboratory for Computational Biology, Shandong Provincial Research Center for Bioinformatic Engineering and Technique, Shandong University of Technology, Zibo 255049, China

Received 24 May 2004; revised 12 January 2005
Available online 10 May 2005

Abstract

A novel coronavirus has been identified as the cause of the outbreak of severe acute respiratory syndrome (SARS). Previous phylogenetic analyses based on sequence alignments show that SARS-CoVs form a new group distantly related to the other three groups of previously characterized coronaviruses. In this paper, a geometric approach based on the Z-curve representation of the whole genome sequence is proposed to analyze the phylogenetic relationships of coronaviruses. The evolutionary distances are obtained through measuring the differences among the three-dimensional Z-curves. The Z-curve is approximately described by its geometric center and the associated three eigenvectors, which indicate the center position and the trend of the Z-curve, respectively. Although some information is lost due to the approximate description of the Z-curve, the phylogenetic tree constructed based on these parameters is consistent with those of previous analyses. The present method has the merits of simplicity and intuitiveness, but it is still in its premature stage. Because the phylogenetic relationships are inferred from the whole genome, instead of some individual genes, the present method represents a new direction of phylogeny study in the post-genome era.

© 2005 Elsevier Inc. All rights reserved.

Keywords: Phylogenetic tree; Coronavirus; Severe acute respiratory syndrome; SARS-CoV; Z-curve

1. Introduction

The outbreak of atypical pneumonia, referred to as severe acute respiratory syndrome (SARS) was first identified in Guangdong Province, China, and spread to several countries later (Drosten et al., 2003; Ksiazek et al., 2003; Lee et al., 2003; Peiris et al., 2003; Poutanen et al., 2003; Tsang et al., 2003). A novel coronavirus was isolated and found to be the cause of SARS. Although SARS has been under control, some scattering cases infected by SARS-CoVs were reported. No effective drugs are currently available to cure this disease. Gaining insight into the phylogenetic relationships among coronaviruses would be helpful to discover drugs and develop vaccines against the virus.

The SARS-coronavirus is a new member of the order Nidovirales, family Coronaviridae, and genus Coronavirus. They consist of a diverse group of large, enveloped, positive-stranded RNA viruses that cause respiratory and enteric diseases in humans and other animals (Rota et al., 2003). Excluding SARS-CoVs, coronaviruses can be divided into three groups according to serotypes. Group I and group II contain mammalian viruses, while group II coronaviruses contain a hemagglutinin esterase gene homologous to that of Influenza C virus (Lai and Holmes, 2001). Group III contains only avian viruses. Previous work showed that SARS-CoVs are not closely related to any of the previously characterized coronaviruses and form a distinct group (group IV) within the genus Coronavirus (Marra et al., 2003; Rota et al., 2003).

An intuitive method is proposed to infer the phylogenetic relationships of coronaviruses in this article. Historically, Cork et al. proposed a three-dimensional
representation of genomic sequences, called the \( W \)-curve (Wu et al., 1993). Since then, the \( W \)-curve has been used to analyze genomic sequences and study the phylogeny of bacteria (Cork, 2003; Cork et al., 2002; Cork and Toguem, 2002). Instead of the sequence alignment, we adopt a geometric method based on the \( Z \)-curve of the whole genome. The \( Z \)-curve is a three-dimensional space curve constituting the unique representation of a given DNA sequence in the sense that each can be reconstructed given the other (Zhang and Zhang, 1991, 1994). Based on the \( Z \)-curve method, a coronavirus-specific gene-finding system ZCURVE_CoV has been developed (Chen et al., 2003), and the software is especially suitable for gene recognition in SARS-CoV genomes. The system is further improved by taking the prediction of cleavage sites of viral proteinases in polyproteins into consideration (Gao et al., 2003). Here we use the differences between the three-dimensional space curves as the foundation to derive the phylogeny of coronaviruses. The key problems are what parameters should be used to describe a curve and how to determine evolutionary distances among organisms based on a group of curves. In this paper, we use a series of parameters, such as the geometric center and the covariance matrix to reflect the center position and the distribution pattern of a curve, respectively. The result shows that SARS-CoVs form an independent group, which is consistent with previous analyses.

2. Materials and methods

2.1. Materials

The 24 complete coronavirus genomes used in this paper were downloaded from GenBank, of which 12 are SARS-CoVs and 12 are from other groups of coronaviruses. The name, accession number, abbreviation, and genome length for the 24 genomes are listed in Table 1. According to the existing taxonomic groups, sequences 1–3 belong to group I, and sequences 4–11 are members of group II, while sequence 12 is the only representative of group III. Refer to Table 1 for details.

2.2. The \( Z \)-curve

The \( Z \)-curve is a three-dimensional curve that constitutes a unique representation of a given DNA sequence in the sense that each can be uniquely reconstructed given the other (Zhang and Zhang, 1991, 1994). The resulting curve has a zigzag shape, hence the name \( Z \)-curve. The \( Z \)-curve is briefly presented as follows. Consider a DNA sequence read from the 5\(^\prime\) to the 3\(^\prime\)-end with \( N \) bases. Beginning from the first base, inspect the sequence one base at a time. In the \( n \)th step, where \( n = 1, 2, \ldots, N \), count the cumulative numbers of the bases A, C, G, and T, occurring in the subsequence from the first base to the \( n \)th base in the DNA sequence inspected, and denote them by \( A_n \), \( C_n \), \( G_n \), and \( T_n \), respectively.

### Table 1

| No. | Accession | Group | Abbreviation | Genome               | Length (nt) |
|-----|-----------|-------|--------------|----------------------|-------------|
| 1   | NC_002645 | I     | HCoV-229E    | Human coronavirus 229E | 27,317      |
| 2   | NC_002306 | I     | TGEV        | Transmissible gastroenteritis virus | 28,586      |
| 3   | NC_003436 | I     | PEDV        | Porcine epidemic diarrhea virus | 28,033      |
| 4   | U00735    | II    | BCoVM       | Bovine coronavirus strain Mebus | 31,032      |
| 5   | AF391542  | II    | BCoVL       | Bovine coronavirus isolate BCoV-LUN | 31,028      |
| 6   | AF220295  | II    | BCoVQ       | Bovine coronavirus strain Quebec | 31,100      |
| 7   | NC_003045 | II    | BCoV        | Bovine coronavirus | 31,028      |
| 8   | AF208067  | II    | MHVM        | Murine hepatitis virus strain ML-10 | 31,233      |
| 9   | AF201929  | II    | MHV2        | Murine hepatitis virus strain 2 | 31,276      |
| 10  | AF208066  | II    | MHVP        | Murine hepatitis virus strain Penn 97-1 | 31,112      |
| 11  | NC_001846 | II    | MHV         | Murine hepatitis virus | 31,357      |
| 12  | NC_001451 | III   | IBV         | Avian infectious bronchitis virus | 27,608      |
| 13  | AY278488  | IV    | BJ01        | SARS coronavirus BJ01 | 29,725      |
| 14  | AY278741  | IV    | Urbani      | SARS coronavirus Urbani | 29,727      |
| 15  | AY278491  | IV    | HKU-39849   | SARS coronavirus HKU-39849 | 29,742      |
| 16  | AY278554  | IV    | CUHK-W1     | SARS coronavirus CUHK-W1 | 29,736      |
| 17  | AY282752  | IV    | CUHK-Su10   | SARS coronavirus CUHK-Su10 | 29,736      |
| 18  | AY283794  | IV    | SIN2500     | SARS coronavirus Sin2500 | 29,711      |
| 19  | AY283795  | IV    | SIN2677     | SARS coronavirus Sin2677 | 29,705      |
| 20  | AY283796  | IV    | SIN2679     | SARS coronavirus Sin2679 | 29,711      |
| 21  | AY283797  | IV    | SIN2748     | SARS coronavirus Sin2748 | 29,706      |
| 22  | AY283798  | IV    | SIN2774     | SARS coronavirus Sin2774 | 29,711      |
| 23  | AY291451  | IV    | TW1         | SARS coronavirus TW1 | 29,729      |
| 24  | NC_004718 | IV    | TOR2        | SARS coronavirus | 29,751      |
tively. The Z-curve consists of a series of nodes $P_n$, where $n = 1, 2, \ldots, N$, whose coordinates are uniquely determined by the Z-transform of DNA sequences (Zhang and Zhang, 1991, 1994)

$$
\begin{align*}
x_n &= (A_n + G_n) - (C_n + T_n) \equiv R_n - Y_n, \\
y_n &= (A_n + C_n) - (G_n + T_n) \equiv M_n - K_n, \\
z_n &= (A_n + T_n) - (C_n + G_n) \equiv W_n - S_n,
\end{align*}
$$

where $A_0 = C_0 = G_0 = T_0 = 0$ and $x_0 = y_0 = z_0 = 0$. Here $R$, $Y$, $M$, $K$, $W$, and $S$ represent the bases of purine, pyrimidine, amino, keto, weak hydrogen bonds, and strong hydrogen bonds, respectively, according to the Recommendation 1984 by the NC-IUB (Cornish-Bowden, 1985). The line that connects the nodes $P_0$ ($P_0 = 0$), $P_1$, $P_2$, $\ldots$, until $P_N$ one by one sequentially is called the Z-curve for the DNA sequences inspected.

The Z-curve defined above is a three-dimensional space curve, having three independent components, i.e., $x_n$, $y_n$, and $z_n$, which display the distributions of bases of $R/Y$, $M/K$, and $W/S$ types, respectively, along the sequence. By viewing the Z-curve, some global and local features of the sequence can be detected in a perceivable way. For almost all genome or chromosome sequences, the curves of $z_n \sim n$ are roughly straight lines (Zhang et al., 2001). For convenience, the curve of $z_n \sim n$ is fitted by a straight line using the least square technique

$$
z = kn,
$$

where $(z, n)$ is the coordinate of a point on the fitted straight line and $k$ is its slope. Instead of using the curve of $z_n \sim n$, we will use the $z'_n \sim n$ curve hereafter, where

$$
z'_n = z_n - kn.
$$

### 2.3. Algorithm

In this paper, we propose a new way to infer evolutionary distances between organisms from the whole genome sequences. As the Z-curve is a unique representation of a genome, it can be used to reflect a genome’s characteristics (Fig. 1). For convenience, we use the coordinates $(X, Y, Z)$ rather than $(X, Y, Z')$. The differences among the Z-curves of these genomes form the basis for constructing the phylogenetic tree. To study the phylogenetic relationships, the process can be separated into three stages. First, the Z-curve of each genome is described by a set of parameters; second, the distance matrix is generated based on the parameters obtained in the first stage; and finally, the phylogenetic tree can be constructed based on the distance matrix.

![Fig. 1. The three-dimensional Z-curves ($x, y, z'$) for three complete coronavirus genomes. (A–C) The Z-curves of BJ01, TOR2, and BCoV, respectively. It can be clearly seen that the Z-curves of BJ01 and TOR2 are very similar, while the Z-curve of BCoV is significantly different from the former two. This forms the basis of the present method. (D) A sketch of the three eigenvectors for a certain genome (TOR2), which illustrates the relationship between the three eigenvectors and the Z-curve.](image-url)
Table 2
The geometric center and three eigenvectors of the $Z$-curve for each of the 24 coronavirus genomes.

| $i$ | Abbreviation | $x$  | $y$  | $z$   | $C_i^x$ | $C_i^y$ | $C_i^z$ | $C_i^{x,y}$ | $C_i^{x,z}$ | $C_i^{y,z}$ |
|-----|--------------|------|------|-------|---------|---------|---------|-------------|------------|------------|
| 1   | HCoV-229E    | 313.52 | 1930.76 | 22.81 | 0.80520 | -0.16770 | -0.56879 | 0.20016     | 0.97975    | -0.00552   |
| 2   | TGEV         | -21.86 | -1185.34 | -87.64 | 0.95040 | -0.06255 | -0.30468 | 0.06443     | 0.99791    | -0.00388   |
| 3   | PEDV         | -733.04 | -1930.98 | -194.31 | 0.90732 | -0.36443 | -0.20968 | 0.37249     | 0.92804    | -0.00112   |
| 4   | BCoVM        | -272.66 | -2691.17 | -94.48 | 0.95060 | -0.13111 | -0.28044 | 0.13749     | 0.99049    | -0.00411   |
| 5   | BCoV         | -268.51 | -2658.04 | -95.32 | 0.95068 | -0.13276 | 0.28034  | 0.13965     | 0.99019    | -0.00464   |
| 6   | BCoVQ        | -257.34 | -2710.49 | -90.44 | 0.95950 | -0.13084 | -0.24949 | 0.13399     | 0.99097    | -0.00440   |
| 7   | BCoV         | -269.55 | -2643.97 | -95.53 | 0.97838 | -0.13953 | 0.15268  | 0.14190     | 0.98987    | -0.00467   |
| 8   | MHVM         | -129.63 | -2295.56 | -438.75 | 0.96108 | -0.06279 | 0.26906  | 0.06409     | 0.99794    | 0.00395    |
| 9   | MHV2         | -184.66 | -2375.26 | -428.28 | 0.98451 | -0.07886 | 0.15662  | 0.07893     | 0.99686    | 0.00582    |
| 10  | MHVP         | -197.59 | -2384.87 | -384.67 | 0.98847 | -0.08099 | 0.12835  | 0.08133     | 0.99668    | 0.00255    |
| 11  | MHV          | -124.73 | -2284.70 | -436.33 | 0.95624 | -0.06424 | 0.28543  | 0.06586     | 0.99782    | 0.00393    |
| 12  | IBV          | 142.03  | -1500.55 | -289.60 | 0.72139 | 0.02288 | 0.69215  | -0.02895    | 0.99958    | 0.00286    |
| 13  | BJ01         | -150.58 | -627.60 | -274.85 | 0.68326 | -0.36116 | -0.63460 | 0.46347     | 0.88610    | 0.00527    |
| 14  | Urbani       | -152.91 | -632.95 | -270.13 | 0.67899 | -0.35342 | -0.64348 | 0.45769     | 0.88910    | 0.00538    |
| 15  | HKU-39849    | -154.65 | -622.44 | -273.25 | 0.65662 | -0.34854 | -0.66926 | 0.46481     | 0.88539    | 0.00535    |
| 16  | CUIH-W1      | -150.87 | -623.56 | -276.66 | 0.67352 | -0.35379 | -0.64901 | 0.46128     | 0.88724    | 0.00495    |
| 17  | CUIHK-Su10   | -150.06 | -626.75 | -278.13 | 0.67051 | -0.34988 | -0.65422 | 0.45881     | 0.88852    | -0.00496   |
| 18  | SIN2500      | -148.99 | -627.23 | -277.78 | 0.67073 | -0.35031 | -0.65577 | 0.45932     | 0.88826    | -0.00472   |
| 19  | SIN2677      | -148.91 | -629.56 | -278.38 | 0.66552 | -0.34716 | -0.66073 | 0.45863     | 0.88862    | -0.00494   |
| 20  | SIN2679      | -148.22 | -627.18 | -277.99 | 0.66444 | -0.34866 | -0.65995 | 0.45800     | 0.88894    | -0.00471   |
| 21  | SIN2748      | -149.34 | -626.83 | -277.39 | 0.66698 | -0.34948 | -0.65803 | 0.46041     | 0.88769    | -0.00479   |
| 22  | SIN2774      | -148.27 | -627.13 | -277.97 | 0.66611 | -0.34675 | -0.66035 | 0.45810     | 0.88889    | -0.00467   |
| 23  | TW1          | -152.93 | -632.21 | -272.52 | 0.66820 | -0.34900 | -0.65705 | 0.45918     | 0.88833    | -0.00488   |
| 24  | TOR2         | -152.69 | -630.34 | -271.00 | 0.67042 | -0.34999 | -0.65425 | 0.45892     | 0.88846    | -0.00502   |

* Refer to the text for detailed explanation about the meaning of the mathematical symbols used in this table.
(i) The parameters of the Z-curve for each genome. Based on the Z-curve, any genome can be represented by a three-dimensional space curve composed of $N$ nodes corresponding to every base position denoted by $x_n, y_n, z'_n$ where $n = 1, 2, \ldots, N$ (Figs. 1A–C). To describe its characteristics, we calculate the following parameters. The first is the geometric center of all the $n$ nodes
\[ \bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n, \quad \bar{y} = \frac{1}{N} \sum_{n=1}^{N} y_n, \quad \bar{z}' = \frac{1}{N} \sum_{n=1}^{N} z'_n. \] (4)

Consequently, we can obtain $(\bar{x}, \bar{y}, \bar{z'})$ for each genome. Refer to Table 2 for details.

Then, the covariance matrix which describes the global distribution pattern of the three-dimensional space curve is calculated as follows:
\[ \Sigma = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{pmatrix} = (\sigma_{pq}), \quad p, q = x, y, z', \] (5)
where
\[ \sigma_{pq} = \frac{1}{N-1} \sum_{n=1}^{N} (p_n - \bar{p})(q_n - \bar{q}), \] (6)
where $p, q = x, y, z'$.

Obviously, the matrix is a real symmetric $3 \times 3$ one. Using a $3 \times 3$ matrix to represent a three-dimensional Z-curve is a very rough approximation, resulting in information loss considerably. However, the advantage is that this approximation makes it possible to compare genomes with different lengths. It is seen that a $3 \times 3$ covariance matrix is uniquely derived based on Eq. (6) for each given genome regardless of its length. From a geometrical point of view, the distribution pattern can be reduced to a three-dimensional ellipsoid approximately. Each direction of the main axis of the ellipsoid can be denoted by an eigenvector and its length should be proportional to the square root of its associated eigenvalue. The eigenvectors and their associated eigenvalues are defined as follows:
\[ \Sigma C_k = \lambda_k C_k, \quad C_k = (C_{k,1}, C_{k,2}, C_{k,3})^T, \quad k = 1, 2, 3. \] (7)

Corresponding to each eigenvalue $\lambda_k$, there’s an eigenvector $C_k$. Corresponding to $\lambda_1 < \lambda_2 < \lambda_3$, the three eigenvectors are denoted by $C_1, C_2, C_3$, respectively. It’s easy to obtain the eigenvalues and associated normalized eigenvectors using the Jacobi algorithm. The geometric center and three eigenvectors for each of the 24 genomes are obtained in the same way. Refer to Table 2 for details about the parameters.

(ii) The distance matrix derived from the above parameters. In this paper, the Euclidean distance is used to reflect the diversity between two points
\[ d_{ij} = \sqrt{ (\bar{x}_i - \bar{x}_j)^2 + (\bar{y}_i - \bar{y}_j)^2 + (\bar{z}'_i - \bar{z}'_j)^2 }, \]
\[ i, j = 1, 2, \ldots, M, \] (8)
\[ d_{ij} \] denotes the distance between the geometric centers of the $i$th and the $j$th genomes, and $M$ is the total number of all genomes ($M = 24$, here). Then we obtain a real $M \times M$ symmetric matrix whose elements are $d_{ij}$.

To reflect the differences between the trends of every two three-dimensional curves, the angles between the corresponding eigenvectors of every two genomes are used. The three-dimensional vectors are denoted as follows:
\[ C_k = (C_{k,x}, C_{k,y}, C_{k,z})^T, \quad i = 1, 2, \ldots, M, \quad k = 1, 2, 3, \] (9)
where $C_k$ is the $k$th vector of the $i$th genome. Each genome has three such eigenvectors. According to the projections on the three axes, the vectors can be divided into three groups. The three groups of vectors are represented with arrows of different styles (refer to Fig. 2A). Obviously they can be separated apart depending on their space distribution. The dark group ($X$ group) has the greatest projections on the $x$-axis, while the vectors represented with dot ($Y$ group) and grey ($Z$ group) arrows have the greatest projections on the $y$-axis and the $z'$-axis, respectively. For each genome, the three vectors can be divided into three groups, i.e., each genome has three vectors belonging to three groups, respectively.

The three groups of eigenvectors are obtained, and denoted by $C_X$, $C_Y$, and $C_Z$, respectively (see Table 2). The cosine between any two vectors in a certain group can be computed as follows:
\[ \cos \theta_{ij} = \frac{C_{x,1} \cdot C_{x,1}'}{|C_{x,1}| \cdot |C_{x,1}'|}, \quad i, j = 1, 2, \ldots, M, \quad k = x, y, z'. \] (10)

Repeating this procedure for all the three groups, we obtain three real $M \times M$ symmetric matrices. These matrices are then translated into angles, whose elements are as follows:
\[ \theta_{ij} = \arccos (\cos \theta_{ij}), \quad i, j = 1, 2, \ldots, M, \quad k = x, y, z'. \] (11)

The sum of $\theta_{ij}$ over $k$ for given $i, j$ can be used to reflect the trend information of the eigenvectors involved.

\[ \text{Fig. 2. The three groups of eigenvectors (denoted with different arrows). The vectors in the } X, \text{ and } Z' \text{ groups are denoted by dark, dot, and grey arrows, respectively. (A and B) The eigenvectors of the 24 genomes observed from different directions. It can be seen from (A) that the three groups can be separated according to their three-dimensional space distribution. (B) The vectors in } Y \text{ group of the 24 genomes are coplanar and they are almost in the } x-y \text{ plane.} \]
coordinates and the trends in Figs. 1A–C, we can see belongs to another group of coronaviruses. From the examples. TOR2 and BJ01 are SARS-CoVs and BCoV of the difference among the three-dimensional curves, we estimated by

\[ D_{ij} = d_{ij} \times \Theta_{ij}, \quad i, j = 1, 2, \ldots, M. \]  

Consequently, two sets of parameters are obtained. The first reflects the difference of center positions represented by the Euclid distance between the geometric centers. The second indicates the difference of the trends of the Z-curves represented by the related eigenvectors. The overall distance \( D_{ij} \) between the species \( i \) and \( j \) is defined by

\[ D_{ij} = d_{ij} \times \Theta_{ij}, \quad i, j = 1, 2, \ldots, M. \]  

(iii) Clustering. Accordingly, a real symmetric \( M \times M \) matrix \( D_{ij} \) is obtained and used to reflect the evolutionary distance between the species \( i \) and \( j \). The clustering tree is constructed using the UPGMA method in PHYLIP package (http://evolution.genetics.washington.edu/phylip.html). The final phylogenetic tree is drawn using the DRAWGRAM program in the PHYLIP package. The branch lengths are not scaled according to the distances and only the topology of the tree is concerned.

3. Results and discussion

3.1. The three-dimensional Z-curve for a complete genome

As mentioned above, one of the advantages of the Z-curve is its intuitiveness. The feature of a genome can be viewed intuitively regardless of how long the genome is. Therefore, global and local compositional features of a genome can be grasped quickly in a perceivable form (Zhang et al., 2003). To give an intuitive comprehension of the difference among the three-dimensional curves, we take SARS-CoV strains TOR2, BJ01, and BCoV as examples. TOR2 and BJ01 are SARS-CoVs and BCoV belongs to another group of coronaviruses. From the coordinates and the trends in Figs. 1A–C, we can see that the Z-curves of TOR2 and BJ01 are almost the same while that of BCoV is significantly different from both of them, indicating that the former two have close phylogenetic relationship, whereas the relationships between the former two and the latter are more distant. Similarity of related Z-curves implies close evolutionary relationship of the organisms involved (Zhang et al., 2003) and vice versa. This constitutes the basis of the current algorithm.

The Z-curve is approximately described by the geometric center and eigenvectors, which indicate its center position and the trends, respectively (Fig. 1D). In Fig. 1D the three arrows represent the three eigenvectors, and the point from which they start is the geometric center. The three eigenvectors of a certain genome can be divided into three groups according to their relationships with the axes (refer to Fig. 2). The trends of Z-curves carry a part of the information used to construct the phylogenetic tree, and some interesting results can be revealed by this figure. It can be seen from Fig. 2B that the vectors in the Y group, which have the greatest projections on the positive y-axis, are coplanar perfectly. They are almost in the x-y plane. As can be seen from the plot, the 24 vectors are almost superposed with each other as a single vector. The phenomenon can also be seen from the data in Table 2. All of the absolute value of \( C_{ij} \) (\( i = 1, 2, \ldots, M \)) are smaller than 0.0059. That is to say, they all have very small projections on the z-axis and are constrained into the \( x-y \) plane. The vectors in the X group and Z group (represented with black and grey arrows, respectively, in Fig. 2B) are also coplanar in the \( x-z \) plane, though their coplanarity is not as good as that of the Y group.

3.2. Phylogenetic tree of coronaviruses

As mentioned above, there are three groups of coronaviruses. Group I includes HCoV-229E, TGEV, and PEDV and group II contains BCoV, BCoVL, BCoVM, BCoVQ, MHV, MHV2, MHVM, MHVP, etc. All the viruses in these two groups are mammalian viruses. Group III contains only avian viruses, of which only the genome of IBV has been completely sequenced. Many researchers have analyzed the phylogenetic relationships among coronavirus genomes based on the 3C-like proteinase, polymerase, the structural proteins S, E, M, and N, respectively (Marra et al., 2003; Rota et al., 2003). Their results indicated that SARS-CoVs are not closely related to any of the previously characterized coronaviruses and form a distinct group (group IV) within the genus Coronavirus (Marra et al., 2003; Rota et al., 2003). As shown in Fig. 3, four groups of coronaviruses can be seen from the phylogram. The SARS-CoVs appear to cluster together and form a separate branch, which can be distinguished easily from other three groups of coronaviruses. IBV, belonging to group III, is situated at an independent branch, whereas the TGEV, PEDV, and HCoV-229E, which belong to group I, tend to cluster together. In another branch, the group II coronaviruses, including BCoV, BCoVL, BCoVM, BCoVQ, MHV, MHV2, MHVM, and MHVP tend to cluster together. First, group I and group II, which are all mammalian viruses, cluster together forming a bigger group. Second, this group joins group III, which contains only avian viruses, to form a much bigger group. Finally, SARS-CoVs join them and result in the phylogenetic tree shown in Fig. 3. The resulting monophyletic clusters agree perfectly with the established taxonomic groups. To validate the current method, a set of random sequences were used as a control. We generated 100 random sequences meeting the requirements in the method. Each time a phylogenetic analysis was done using 25 sequences including one random sequence and the 24 genomes. Consequently, 100 phylogenetic trees were obtained. Ninety-eight out of
the 100 trees showed that the random sequence formed a distinct group without disturbing the other four groups. Only two of the random sequences disturbed the four groups, suggesting that the current method is solid with respect to the situation that a random sequence is added.

3.3. Comparison with the results of previous analyses

Almost all of the previous analyses revealed that SARS-CoVs form a distinct group different from the other three groups of coronaviruses. However, the question that how SARS-CoVs emerged suddenly still remains open. Rota et al. and Marra et al. performed phylogenetic analysis based on sequence alignments using different genes. The results indicated that SARS-CoVs belong to a new group but the original group that SARS-CoVs were derived from could not be determined (Rota et al., 2003). Grigoriev found that the mutational patterns in SARS-CoV genome were strikingly different from the other coronaviruses in terms of mutation rates (Grigoriev, 2004). Phylogenetic analysis based on codon usage pattern suggested that SARS-CoV was diverged far from all the three known groups of coronavirus (Gu et al., 2004). The overall level of similarity between SARS-CoVs and the other coronaviruses is low (Rota et al., 2003). We suppose that this is due to different evolution paths. The isolation of SARS-CoV-like virus in Himalayan palm civets indicates a route of interspecies transmission. We hypothesize that some events such as the nucleotide deletion or mutation in some important genes of the precursor may have resulted in the change of host range.

3.4. Merits of the current method

Due to the lack of morphological features and frequent gene exchanges, it is highly valuable to develop methods of molecular phylogeny for viruses. Now phylogenetic analysis based on sequence alignments is well developed. Sequence alignments are always based on some special genes or some conserved fragments (Saitou, 1996). Such analysis can be done at both the amino acid level and the nucleotide level. To overcome the biases caused by individual genes or genome segments, it is valuable to develop methods of molecular phylogenetic analysis based on whole genome sequences. Being different from the sequence alignment method, the current method is a geometric approach which is based on measuring the differences of Z-curves of whole genomes, including coding and non-coding sequences. There is no need to search for similar sequences. Probably, the most remarkable advantages of the present method is its simplicity and intuitiveness.

![Fig. 3. The phylogenetic tree constructed with the current method.](image-url)
The increasing availability of complete genomes has cast doubt instead of adding details to the phylogenetic tree (Qi et al., 2004). Phylogenetic analysis based on sequence alignments is usually done on the most conservative part of a gene. These fragments are usually coding sequences, especially the sequences coding for catalytic sites or the core of proteins, because they tend to be more evolutionarily conserved. It was said by a virologist that people could not simply assume that a virus can be represented by its polymerase (http://www.ncbi.nlm.nih.gov/ICTV/). A virus must be viewed as a whole. Non-coding sequences also play an important role in the virus, so do the less conserved genes. In addition, analyses based on different genes may lead to different results. Consequently, by using complete genomes one can avoid choosing which genes to be aligned. Therefore methods that are based on the whole genome are likely to be more objective. Recently, a k-string composition approach was proposed to analyze prokaryote phylogeny based on the whole proteome and satisfactory results were obtained (Qi et al., 2004); however, such analysis must rely on the annotation information. In contrast, the complete genome sequence is the only input of the current method; neither the annotation information, nor any adjustable parameters are needed. It is noteworthy that the current method can be performed automatically without any human intervention.

The Z-curve, which serves as the foundation of the present method is a powerful tool to study the complete genome sequence. The Z-curve contains all the information that the corresponding DNA sequence carries. Many characteristics of a genome with biological meaning can be observed from the corresponding Z-curve, such as the replication origins and genomic islands for some bacterial and archaeal genomes (Zhang et al., 2003). We can inspect a genome in an intuitive way regardless of the gene content and gene order, even though the sequences are of different lengths. If the Z-curves of two species show similar pattern even though the genomes have different lengths, one may infer that they are evolutionarily close organisms, and vice versa. In this paper, we use the geometric center and the eigenvectors to describe the pattern approximately. Although this is only a rough approximation, it represents just an attempt to apply the Z-curve method to the phylogenetic analysis and the results obtained agree well with previous analyses.

3.5. Limitations of the current method

This method is aimed to analyze the phylogeny of the genomes which have close phylogenetic relationships. Phylogenetics analysis is based on the differences among the three-dimensional Z-curves. In this paper, the 24 genomes under study all belong to the same genus Coro-

navirus. Additionally, the differences of length among genomes are not very large. If the genomes under study have much farther phylogenetic relationships, and the differences in length are considerably large, the present method may not work. Consequently, cautions must be taken when using the present method to study the phylogeny of organisms with far evolutionary distances. In addition, unlike the estimation based on comparison of orthologous genes, the Z-curve approach is also sensitive to genome rearrangements: a single large-scale inversion can change the form of Z-curve drastically. Therefore, the method presented here is considerably limited in the cases of genome rearrangements. In addition, as mentioned above, the three-dimensional Z-curve is approximately depicted by a few parameters, such as the geometric center and the associated three eigenvectors. Consequently, information contained in the Z-curve is lost considerably in so doing. It is reasonable to suppose that the more information is extracted from the Z-curve, the more accurate result can be gained. Therefore, the current method can be improved if new and more effective algorithms are proposed to extract information contained in the Z-curves. In summary, although the present method has some advantages, it is still in its premature stage. The method may not be applied to some general cases, therefore the applications of it are considerably limited at present.

4. Conclusion

A geometric approach to infer phylogenetic relationships based on the Z-curves of complete genomes is proposed in this article. Phylogenetic analysis of the 24 coronaviruses shows that SARS-CoVs belong to a new cluster, named group IV, and this result is consistent with those of previous analyses. The method has much room to be improved because of the possibility to extract information from the whole genome, instead of some individual genes. Although having some limitations, the current whole-genome-based geometric approach represents a new direction to infer phylogenetic relationships of organisms in the post-genome era. However, the method is still in its premature stage and its applications are considerably limited at present.

Acknowledgments

We are indebted to both referees, whose comments are critical for improving the quality of the paper. We thank Ren Zhang for invaluable assistance. We are thankful to Prof. Jingchu Luo (Peking University) and Prof. Xi-Tai Huang (Nankai University) for their invaluable help. Discussions with Feng-Biao Guo and Bin-Guang Ma are acknowledged. The present study
was supported in part by the National Natural Science Foundation of China (Grant 90408028).

References

Chen, L.L., Ou, H.Y., Zhang, R., Zhang, C.-T., 2003. 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. 307, 382–388.

Cork, D.J., 2003. Achieving congruence of long genomic sequences with the W-curve. In: Lapafote, F., McMorris, F.R., Janowitz, M. (Eds.), Bioconsensus, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol. 61. American Mathematical Society, Providence, RI, pp. 123–134.

Cork, D.J., Hutech, T.B., Marland, E., Zmuda, J., 2002. Achieving congruency of phylogenetic trees generated by W-curves of genomic sequences. In: Valafar, F. (Ed.), Techniques in Bioinformatics and Medical Informatics. Ann. N. Y. Acad. Sci. 980, 23–31.

Cork, D.J., Toguem, A., 2002. Using fuzzy logic to confirm the integrity of a pattern recognition algorithm for long genomic sequences: the W-curve of genomic sequences. In: Valafar, F. (Ed.), Techniques in Bioinformatics and Medical Informatics. Ann. N. Y. Acad. Sci. 980, 32–40.

Cornish-Bowden, A., 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendation 1984. Nucleic Acids Res. 13, 3021–3030.

Drosten, C., Gunther, S., Preiser, W., van der Werf, S., Brodt, H.R., Becker, S., Rabenau, H., Panning, M., Kolesnikova, L., Fouchier, R.A., Berger, A., Burgueri, A.M., Cinatil, J., Eickmnn, M., Escrion, N., Grywna, K., Kramme, S., Manenguerra, J.C., Muller, S., Rickerts, V., Sturmer, M., Vieth, S., Klenk, H.D., Osterhaus, A.D., Schmitz, H., Doerr, H.W., 2003. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N. Engl. J. Med. 348, 1967–1976.

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

Grigoriev, A., 2004. Mutational patterns correlate with genome organization in SARS and other coronaviruses. Trends Genet. 20, 131–135.

Gu, W., Zhou, T., Ma, J., Sun, X., Lu, Z., 2004. Analysis of synonymous codon usage in SARS coronavirus and other viruses in the Nidovirales. Virus Res. 101, 155–161.

Guan, Y., Zheng, B.J., He, Y.Q., Liu, X.L., Zhuang, Z.X., Cheung, G.A., Yang, J., Guan, Y.J., Butt, K.M., Wong, K.L., Chan, K.W., Lim, W., Shortridge, K.F., Yuen, K.Y., Peiris, J.S., Poon, L.L., 2003. Isolation and characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 300, 1394–1399.

Saitou, N., 1996. Reconstruction of genes trees from sequence data. Methods Enzymol. 266, 427–449.

Stavrinides, J., Guttman, D.S., 2004. Mosaic evolution of the severe acute respiratory syndrome coronavirus. J. Virol. 78, 76–82.

Tsang, K.W., Ho, P.L., Ooi, G.C., Yee, W.K., Wang, T., Chan-Yeung, M., Lam, W.K., Seto, W.H., Yam, L.Y., Cheung, T.M., Wong, P.C., Lam, B., Ip, M.S., Chan, J., Yuen, K.Y., Lai, K.N., 2003. A cluster of severe acute respiratory syndrome in Hong Kong. N. Engl. J. Med. 348, 1977–1985.

Wu, D., Roberge, J., Cork, D.J., Nguyen, B.J., Grace, T., 1993. Computer visualization of long genomic sequences. Proc. IEEE Visualization 93, 308–315.

Zhang, C.-T., Wang, J., Zhang, R., 2001. A novel method to calculate the G + C content of genomic DNA sequences. J. Biol. Struct. Dyn. 19 (2), 331–345.

Zhang, C.-T., Zhang, R., 1994. Z-curves, an intuitive tool for visualizing and analyzing DNA sequences. J. Biol. Struct. Dyn. 11, 767–782.

Zhang, C.-T., Zhang, R., Ou, H.Y., 2003. The Z-curve database: a graphic representation of genome sequences. Bioinformatics 19, 593–599.