BIOINFORMATICS

Vol. 19 no. 5 2003, pages 593–599
DOI: 10.1093/bioinformatics/btg041

The Z curve database: a graphic representation of genome sequences

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

[1]Department of Physics, Tianjin University, Tianjin 300072, People's Republic of China and [2]Department of Epidemiology and Biostatistics, Tianjin Cancer Institute and Hospital, Tianjin 300060, People's Republic of China

Received on July 11, 2002; revised on September 12, 2002; October 26, 2002; accepted on November 11, 2002

ABSTRACT

Motivation: Genome projects for many prokaryotic and eukaryotic species have been completed and more new genome projects are being underway currently. The availability of a large number of genomic sequences for researchers creates a need to find graphic tools to study genomes in a perceivable form. The Z curve is one of such tools available for visualizing genomes. The Z curve is a unique three-dimensional curve representation for a given DNA sequence in the sense that each can be uniquely reconstructed given the other. The Z curve database for more than 1000 genomes have been established here.

Results: The database contains the Z curves for archaea, bacteria, eukaryota, organelles, phages, plasmids, viroids and viruses, whose genomic sequences are currently available. All the 3-dimensional Z curves and their three component curves are stored in the database. The applications of the Z curve database on comparative genomics, gene prediction, computation of G+C content with a windowless technique, prediction of replication origins and terminations of bacterial and archaeal genomes and study of local deviations from the Chargaff Parity Rule 2 etc. are presented in detail. The Z curve database reported here is a treasure trove in which biologists could find useful biological knowledge.

Availability: The Z curve database is freely available at the website: http://tubic.tju.edu.cn/zcurve/.

Contact: ctzhang@tju.edu.cn

INTRODUCTION

With the completion of more and more genome projects, a large number of genomic sequences are available in public databases, resulting in an urgent need to find new mathematical approaches to analyze these sequences. There exist two basic mathematical approaches in the studies of theoretical physics. One is algebraic and the other is geometrical. They are complementary in most cases. In the area of genome studies, the algebraic approach is widely used in analyzing genomic sequences currently, whereas the geometrical approach for analyzing genomic sequences has been ignored for a long time. The Z curve is a three-dimensional curve which is a *unique* representation for a given DNA sequence in the sense that each can be *uniquely* reconstructed given the other (Zhang and Zhang, 1991, 1994). Therefore, the Z curve contains all the information that the corresponding DNA sequence carries. The analysis of a DNA sequence can be performed through studying the corresponding Z curve. Historically, various methods for the graphical representation of DNA sequences were proposed, including the H curve (Hamori and Ruskin, 1983), the game representation (Jeffrey, 1990), the W curve (Cork and Wu, 1993), and the 2-dimensional DNA walk (Lobry, 1996) etc. It was shown that most of them are special cases of the Z curve, and an extensive comparison between the Z curve and other representations was detailed in (Zhang and Zhang, 1994). One of the advantages of the Z curve is its intuitiveness. The Z curve of a genome can be viewed on a computer screen or on a piece of paper, regardless of how long the genome is. Therefore, global and local compositional features of genomes can be grasped quickly in a perceivable form. By jointing the methodology of the Z curve with those of statistics, better results could be obtained. To make the Z curve convenient for biologists to use, the Z curves for whole genomes, chromosomes or complete DNA sequences of archaea, bacteria, eukaryota, organelles, phages, plasmids, viroids and viruses were pre-calculated and displayed in the database reported here, which contains more than 7500 records. It is the authors' hope that the database would be a useful tool for genome studies in the post genome era.

THE Z CURVE

The Z curve is a *unique* three-dimensional curve representation for a given DNA sequence in the sense that each

*To whom correspondence should be addressed.

can be *uniquely* reconstructed given the other (Zhang and Zhang, 1991, 1994). The Z curve is composed of a series of nodes $P_0, P_1, P_2, \ldots, P_N$, whose coordinates x_n, y_n and z_n ($n = 0, 1, 2, \ldots, N$, where N is the length of the DNA sequence being studied) are *uniquely* determined by the Z-transform of DNA sequence (Zhang and Zhang, 1994)

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

where A_n , C_n, G_n and T_n are the *cumulative* occurrence numbers of A, C, G and T, respectively, in the subsequence from the 1st base to the nth base in the sequence. 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 $P_0, P_1, P_2, \ldots, P_N$ one by one sequentially with straight lines. Note that the Z curve always starts from the origin of the three-dimensional coordinate system. Once the coordinates x_n, y_n and $z_n (n = 1, 2, \ldots, N)$ of a Z curve are given, the corresponding DNA sequence can be reconstructed from the so-called inverse Z-transform

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

where the 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 completely describe the DNA sequence being studied. The three components x_n, y_n and z_n display the distributions of purine/pyrimidine (R/Y), amino/keto (M/K) and strong-H bond/weak-H bond (S/W) bases along the sequence, respectively. In the subsequence constituted from the 1st base to the nth 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, $y_n < 0$, when the numbers of amino (A/C) and keto bases (C/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, $z_n = 0$. For more detail about the Z curve, refer to Introduction to the Z curve in the website: http://tubic.tju.edu.cn/zcurve/.

The Z curve defined above is generally not smooth at each node. Sometimes, a smooth procedure is needed. The B-spline functions are used to smooth the Z curve (Zhang and Zhang, 1994). Suppose that the smoothed x_n, y_n and z_n are represented by $x(n)$, $y(n)$ and $z(n)$, respectively, then

$$u(n) = \frac{1}{6}u_{n-1} + \frac{2}{3}u_n + \frac{1}{6}u_{n+1}, \quad u = x, y, z, n = 1, 2, \ldots, N-1. \quad (3)$$

To strengthen the smooth effect, Equation (3) may be used repeatedly.

APPLICATIONS OF THE Z CURVE DATABASE

The coordinates of the Z curves for more than 1000 genomes are pre-calculated, and the Z curve for each genome is displayed in the Z curve database. The Z curve database is an alternative version of the genomic sequences stored in GenBank/EMBL/DDBJ databases. The former represents a genome using a curve, whereas the latter represents a genome using a sequence of four letters. The two approaches are complementary. One important advantage of the Z curve is that some features of global and local nucleotide composition of a genome can be displayed in a perceivable form. Although the screen resolution is insufficient to convey the details of the curve, the software provided here can help a user to display local features of the Z curve involved at single nucleotide level (see the discussion later). In the following we will concentrate on the benefits of the Z curve database to Bioinformatics community. The symbols 3D, X, Y, Z (or Z′) and XY in the database represent the three-dimensional Z curve, the x_n, y_n and z_n (or z'_n) component curves and the AT- and GC-disparity curves, for a genome or chromosome, respectively. They will be explained in the following sections.

Analysis and comparison of genomes based on a visual inspection of the Z curves involved

Since the Z curve contains all the information of its corresponding DNA sequence, multiple genomes can be compared by comparing their Z curves. For example, although the genome lengths of the two strains of *E.coli*, K-12 and O157: H17, are 4.6 and 5.5 Mb, respectively, their corresponding three-dimensional Z curves (see Items 16 and 17 in the bacterial section of this database) show similar pattern, indicating that they are evolutionarily close organisms. Another example is the comparison of different assemblies of human chromosome 22, which is shown in Figure 1. The blue and green Z curves denote the assemblies of the human chromosome 22 based on NCBI build 29 and UCSC August 2001 freeze, respectively. All the Z curves in this plot have been smoothed for 10 000 times in order to show a global distribution of nucleotides. It is clearly seen that the

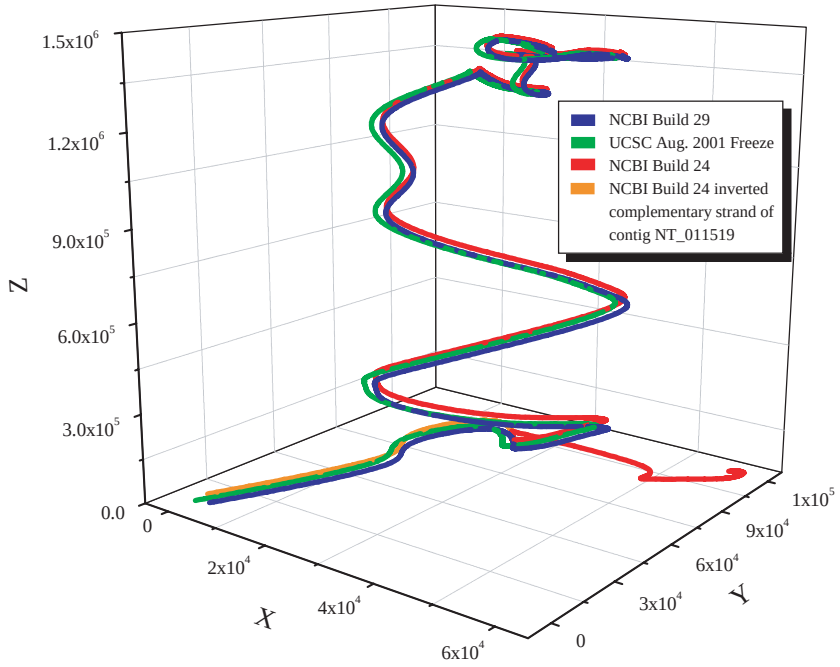

Fig. 1. The Z curves for the human chromosome 22 based on different assemblies. The blue and green Z curves denote the genome sequences of the human chromosome 22 based on NCBI build 29 and UCSC August 2001 freeze, respectively. The two Z curves for the two assemblies overlap very well, indicating that the two assemblies are very close to each other globally, and no large systemic errors occur. However, if the Z curve of UCSC (in green) is compared with that of the NCBI build 24 (in red), it is shown that most parts of Z curves overlap, while the beginning parts of the Z curves undergo different directions. According to the symmetry of the Z curve, it is easy to identify that the beginning part of the red curve (corresponding to a 3.5 Mb contig, NT_011519) shows an inversely complementary symmetry with that of the green one. The curve in orange is the Z curve for the inverted complementary strand of NT_011519, which overlaps with the corresponding part of the Z curve for UCSC August 2001 freeze.

two Z curves for the two assemblies overlap very well, indicating that the two assemblies are very close to each other globally. Therefore, they confirm the result of each other, strongly suggesting that no large systemic errors occur for these two assemblies. However, if the Z curve of UCSC (in green) is compared with that of NCBI build 24 (in red), it is clearly seen that most parts of the Z curves overlap, while the beginning parts of the Z curves undergo different directions. According to the symmetry of the Z curve (Zhang, 1997), it is easy to identify that the beginning part of the red curve shows an inversely complementary symmetry (Zhang, 1997) with that of the green one. Consequently, it is found that the difference is caused by a 3.5 Mb contig, NT_011519, which was not placed according to its coordinates in NCBI build 24. The curve in orange is the Z curve for the inverted complementary strand of NT_011519, which overlaps with the corresponding part of the Z curves for UCSC August 2001 freeze. Note that the finding is reached by simply observing the related Z curves involved. The above examples show that the Z curve methodology provides a simple and intuitive tool for comparative genomics.

Calculation of the G+C content with a windowless technique

The traditional method for the G+C content calculation is based on a sliding-window technique. As a result, only a mean G+C content averaged over a large window size is obtained. The window size may be called ‘resolution’ for the G+C content computation. High resolution of the G+C content needs the window size to be small. However, small window size will lead to a larger statistical fluctuation in the G+C content, indicating that the result is unreliable in this case. Therefore, the resolution of the G+C content computation cannot be high using a window method. A windowless method to calculate the G+C content based on the Z curve was proposed (Zhang *et al.*, 2001). Using this method, the calculation of G+C content can be performed at any resolution. In an extreme case, the G+C content may be computed at a ‘point’ (a base position) in a genome (Zhang *et al.*, 2001). The z'_n curve describes the distribution of the G+C content along the DNA sequences and is defined by

$$z'_n = z_n - k \times n, \quad (4)$$

where z_n is the z-component of the Z curve, and k is the slope of the straight line, which is used to fit the z_n curve using the least square method. In the Z curve database, a user can click Z′ to see the z'_n curve for a genome, where the value of k is also given. Based on the z'_n curve, the G+C content at a point n in a genome is calculated using

$$\text{G+C} \equiv \frac{1}{2}\left(1 - k - \frac{dz'_n}{dn}\right), \qquad (5)$$

where dz'_n/dn is defined in (Zhang *et al.*, 2001). Therefore, a jump in the z'_n curve indicates an A+T-rich region, whereas a drop means a G+C-rich region. A sudden change in the z'_n curve might imply a transfer of foreign DNA sequence from other species. A typical example is the z'_n curve for the smaller chromosome of *Vibrio cholerae*, where the position of the integron island (Zhang *et al.*, 2001) is precisely identified by observing a sudden jump in the z'_n curve, as shown clearly in Figure 2. Viewing the z'_n curves of bacterial and archaeal genomes presented in this database, biologists could find similar phenomena for other organisms. We emphasize the importance of the z'_n curves for genome studies. The distributions of the G+C content along a genome or chromosome sequence using the windowless technique presented are generally different from those obtained using a sliding window method.

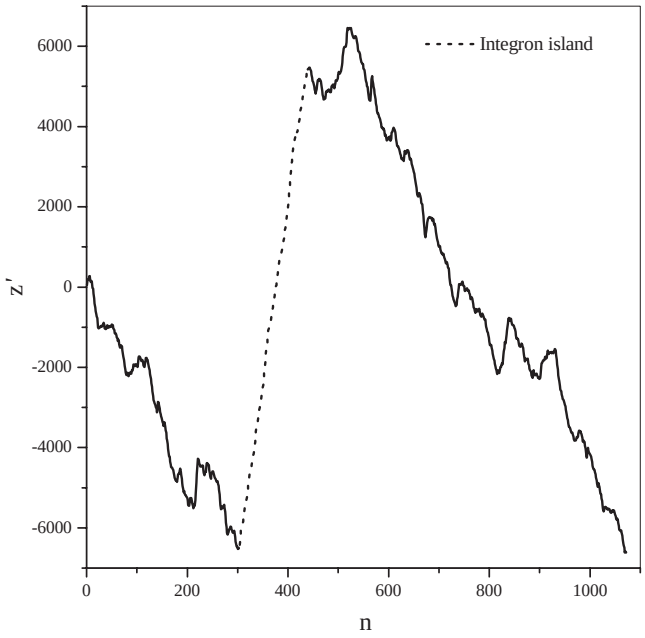

Fig. 2. The $z'_n \sim n$ curve with $k = 0.0678$ for the smaller chromosome of the *Vibrio cholerae* genome, where k is defined in Equation (4). The dash line indicates a long A + T-rich region on this chromosome and the integron island (125 kb) is located exactly at this region.

Analysis of local deviations from Chargaff Parity Rule 2 using the AT- and GC-disparity curves

The Chargaff Parity Rule 1 shows that for a double-strand DNA molecule globally %A = %T and %G = %C (Chargaff, 1950). The rigorous validation of the rule constitutes the basis of Watson–Crick pairs in the DNA double helix model. Interestingly, about 18 years later, Chargaff and co-workers found that this observation still hold even within each single strand DNA (Rudner *et al.*, 1968). This phenomenon is often called the Chargaff Parity Rule 2 or PR2. The PR2 shows that globally both %A ≈ %T and %G ≈ %C are valid for each of the two DNA strands. Using our notation, the same fact can be stated as $A_N \approx T_N$ and $G_N \approx C_N$, where N is the number of nucleotides in the sequence studied. The Chargaff Parity Rule 2 can be expressed as (see Equation (1))

$$(x_N + y_N)/2 \approx 0, \quad (x_N - y_N)/2 \approx 0. \qquad (6)$$

Although the PR2 is roughly correct for almost all genomes, the above equation does not necessarily lead to

$$(x_n + y_n)/2 \approx 0, \quad (x_n - y_n)/2 \approx 0, \quad \text{for } n \in [1, N], \qquad (7)$$

where $(x_n + y_n)/2$ and $(x_n - y_n)/2$ are called AT- and GC-disparity curves, respectively. A user may click XY to view the two curves for each genome. The Chargaff Parity Rule 2 describes only a global feature of the base composition in a single DNA strand, whereas the AT- and GC-disparity curves describe the local excess of A over T and G over C, respectively, in a single strand DNA sequence. The AT- and GC-disparity curves provide a basic tool to study organizations of genomes. The AT- and GC-disparity curves for more than 1000 genomes are displayed in this database. Viewing these curves one by one, we have found that genomes can be roughly classified into three classes, according to the patterns of disparity curves. For the first class (AT-type), $A_n \approx T_n$ and $G_n \neq C_n$, the second class (GC-type), $A_n \neq T_n$ and $G_n \approx C_n$, and the third class (mixture-type), $A_n \neq T_n$ and $G_n \neq C_n$ for any $n \in [1, N]$. The genomes of the Kennedya yellow mosaic virus (AC: D00637) and soybean chlorotic mottle virus (AC: X15828), in the Items 364 and 550 in the virus section of this database, respectively, are typically AT- and GC-type. The AT- and GC-disparity curves for the former and latter are shown in Figure 3a and b, respectively. The genomes of *Escherichia coli*, *Mycoplasma pneumoniae* and *Ureaplasma urealyticum*, (AC: U00096, U00089, AF222849, with the Items 16, 28 and 41, respectively, in the bacterial section of this database), are typically AT-, GC- and mixture-type, respectively.

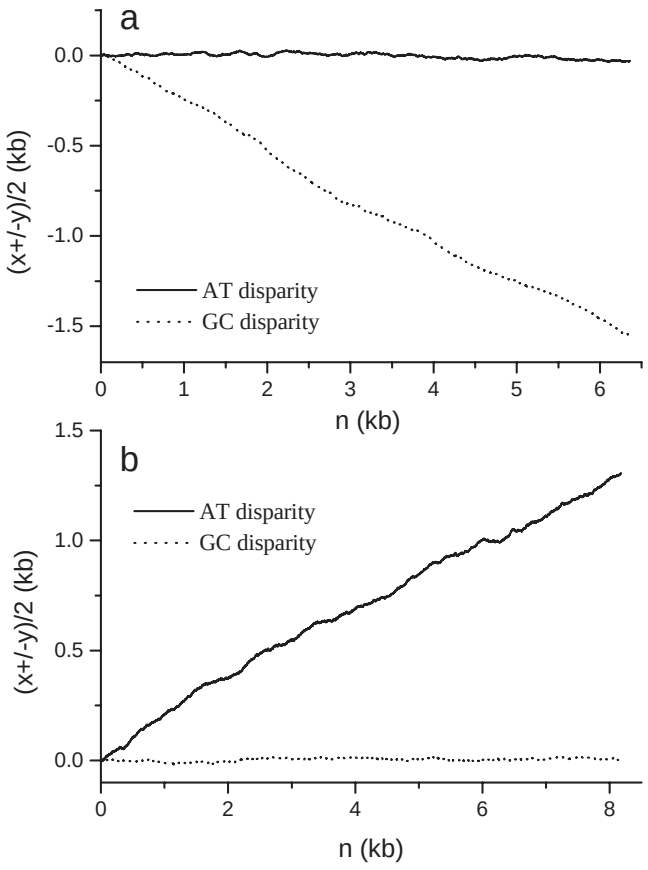

Fig. 3. The AT- and GC-disparity curves for the genomes of (a) the Kennedya yellow mosaic virus and (b) the soybean chlorotic mottle virus. The former is a typical AT-type genome, in which $A_n \approx T_n$ and $G_n \neq C_n$ for any $n \in [1, N]$, as indicated by the almost constant solid line and linearly decreasing broken line, respectively, in Figure 3a. The latter is a typical GC-type genome, in which $A_n \neq T_n$ and $G_n \approx C_n$ for any $n \in [1, N]$, as indicated by the almost constant broken line and linearly increasing solid line, respectively, in Figure 3b.

Prediction of replication origin and termination for some bacterial and archaeal genomes

The deviations of PR2 have been shown to be associated with the replication origin and termination for some bacterial or archaeal genomes (Lobry, 1996; Francino and Ochman, 1997; Grigoriev, 1998; Frank and Lobry, 1999; Zhang and Zhang, 2002). The AT- and GC-disparity curves, as well as the x_n and y_n components of the Z curve, show a change of polarity at the sites of replication origin and termination for some bacterial or archaeal genomes. The genome of *E.coli* is a typical example. Traditionally, the GC-skew analysis is often used to explore the nucleotide compositional asymmetry around replication origin. The GC-skew is defined to be $(C - G)/(C + G)$, where C and G are the occurrence numbers of C and G residues in a sliding window (Lobry, 1996). In this method, the size of the sliding window is difficult to be chosen. To overcome the drawback, the method of cumulative GC-skew was proposed in which no sliding windows are used (Grigoriev, 1998). For the genome of *Methanosarcina mazei*, however, the cumulative GC-skew failed to show compositional asymmetry around the replication origin. Since the Z curve is the unique representation of a given DNA sequence, it contains all the information that the DNA sequences carries. Therefore, almost all DNA walks are special cases of the Z curve or functions of x_n, y_n and z_n (Zhang and Zhang, 1994). For instance, the trajectory resulting from the DNA walk (Lobry, 1996) can be obtained by an appropriate rotation on the projection of the relevant Z curve onto the x–y plane (Zhang and Zhang, 1994). Furthermore, the cumulative GC-skew, a function of x_n, y_n and z_n, is equal to $(y_n - x_n)/(n - z_n)$ [see Equation (1)]. For almost all the replication origins of the bacterial and archaeal genomes that were identified using the GC-skew method, the Z curves have a change of polarity around replication origins. However, for some genomes, e.g. *Methanosarcina mazei*, the GC-skew failed to show compositional asymmetry around the replication origin that is detected based on the Z curve (Zhang and Zhang, 2002). In summary, replication origins and terminations of some bacterial and archaeal genomes can be predicted within the frame of the Z curve in a unified manner, using the x_n, y_n, AT- and GC-disparity curves presented in this database.

Gene recognition using the Fourier transform performed on x_n, y_n and z_n

It is well known that there exists a 3-periodicity in coding regions. The 3-periodicity can be easily detected by studying the power spectrum of the Fourier transform performed separately on x_n, y_n and z_n (Yan *et al.*, 1998). A special algorithm based on the Fourier transform of x_n, y_n and z_n to identify shorter exons in human genome was proposed (Yan *et al.*, 1998). A website (http://www.imtech.res.in/raghava) was established to provide the service of gene prediction using the Fourier transform of x_n, y_n and z_n for a user DNA sequence (Issac *et al.*, 2002). A more powerful gene-finding method using the Z curve approach is based on the phase-specific Z curves (Zhang and Wang, 2000; Wang and Zhang, 2001). In this database, software services using the phase-specific Z curves are provided to predict genes in the yeast genome (Zcurve_Y) and bacterial or archaeal genomes (Zcurve_C).

Software Services for drawing and manipulating the Z curve

To facilitate the use of the Z curve database, several software services are provided, which are described in detail as follows.

(i) The software, Zplotter online, has been developed to draw and manipulate the Z curve online based on a user's input sequence. The functions of this software include: (a) Display the x_n, y_n, z_n, z'_n, AT- and GC-disparity curves of a user's DNA sequence in the forward (5′–3′), inverted (3′–5′) directions and their complementary strands, respectively. Therefore, totally $6 \times 4 = 24$ curves can be displayed on the computer screen for a given DNA sequence. (b) The resolution of any local parts of each curve can be arbitrarily adjusted by using the built-in zoom function. Using the cursor to frame the part of interest, then it will be amplified, and the detail will be displayed. The procedure can be repeated many times until the highest resolution (single nucleotide level) is achieved. (c) The coordinates of any site in a curve can be displayed by putting the cursor at the site of interest. The jointing utilization of the above three functions is particularly useful in identifying the sites of interest precisely. For example, the site at which the G+C content of a given DNA sequence undergoes an abrupt change can be determined with the accuracy at single nucleotide level. Simply click Zplotter in the home page of the database to activate the software.

(ii) The functions of the above software can be further strengthened by incorporating the following programs. The first one is for changing the origin of a circular genome arbitrarily. The second is to extract the sub-sequence from an input DNA sequence at the given start and end positions. The former is particularly useful for studying bacterial and archaeal genomes. Click miscellaneous program on the home page to access these programs.

(iii) To find the Z curve for the genome of interest quickly, click Search in the web page of the database. A user can either input the complete Latin name of the organism of interest or the accession number of the corresponding genome to search the Z curve database for relevant information.

(iv) In addition to the above services, a user can submit the sequence and get the Z curve coordinates back by using the Z curve coordinate calculator. Simply click Calculate Z curve coordinates on the web page to read the instruction about this software service. Once the Z curve coordinates are received, users can draw the Z curve using their own plotting software.

(v) Finally, a user can also download the Zplotter program free of charge and run it from user's own computer.

ACKNOWLEDGEMENTS

The present study was supported in part by the 973 Project of China (grant 1999075606). We thank the anonymous referees for their constructive comments on the early version of the manuscript.

NOTE ADDED IN PROOF

The z'_n curve, also termed cumulative GC profile, can be used to identify the isochore structures of genomes (Zhang and Zhang, 2003). In addition, an *ab initio* gene-finding system based on the *Z* curve method has been developed, which is also accessible from the *Z* curve database (Guo *et al.*, 2003).

REFERENCES

Chargaff,E. (1950) Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, **6**, 201–240.

Cork,D.J. and Wu,D. (1993) Computer visualization of long genomic sequences invited IEEE special issue. *Visualization in the Sciences*, 308–315.

Francino,M.P. and Ochman,H. (1997) Strand asymmetries in DNA evolution. *Trends Genet.*, **3**, 240–245.

Frank,A.C. and Lobry,R. (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene*, **238**, 65–77.

Grigoriev,A. (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Res.*, **26**, 2286–2290.

Guo,F-.B., Ou,H.-Y. and Zhang,C.-T. (2003) ZCURVE: A new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res.*, **31**, in press.

Hamori,E. and Ruskin,J. (1983) H Curve—a novel method of representation of nucleotide series especially suitable for long DNA sequence. *J. Biol. Chem.*, **258**, 1318–1327.

Issac,B., Singh,H., Kayr,H. and Raghava,G.P.S. (2002) Locating probable genes using Fourier transform approach. *Bioinformatics*, **18**, 196–197.

Jeffrey,H.J. (1990) Chaos game representation of gene structure. *Nucleic Acids Res.*, **18**, 2163–2170.

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

Rudner,R., Karkas,J.D. and Chargaff,E. (1968) Separation of *B. subtilis* DNA into complementary strands. III. Direct Analysis. *Proc. Natl Acad. Sci. USA*, **60**, 921–922.

Wang,J. and Zhang,C.T. (2001) 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.*, **268**, 4261–4268.

Yan,M., Lin,Z.S. and Zhang,C.T. (1998) A new Fourier transform approach for protein coding measure based on the format of the Z curve. *Bioinformatics*, **14**, 685–690.

Zhang,C.T. (1997) A symmetrical theory of DNA sequences. *J. Theor. Biol.*, **187**, 297–306.

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

Zhang,C.T., Wang,J. and Zhang,R. (2001) A novel method to calculate the G+C content of genomic DNA sequences. *J. Biomol. Struc. Dyn.*, **19**, 333–341.

Zhang,C.T. and Zhang,R. (1991) Analysis of distribution of bases in the coding sequences by a diagrammatic technique. *Nucleic Acids Res.*, **19**, 6313–6317.

Zhang,R. and Zhang,C.T. (1994) Z curves, an intuitive tool for visualizing and analyzing DNA sequences. *J. Biomol. Struc. Dyn.*, **11**, 767–782.

Zhang,R. and Zhang,C.T. (2002) Single replication origin of the archaeon *Methanosarcina mazei* revealed by the Z curve method. *Biochem. Biophys. Res. Commun.*, **297**, 396–400.

Zhang,C.-T. and Zhang,R. (2003) An isochore map of the human genome based on the Z curve method. *Gene*, in press.