Archaea 1, 335–346

Identification of replication origins in archaeal genomes based on the *Z*-curve method

REN ZHANG[1] and CHUN-TING ZHANG[2,3]

[1] *Department of Epidemiology and Biostatistics, Tianjin Cancer Institute and Hospital, Tianjin 300060, China*

[2] *Department of Physics, Tianjin University, Tianjin 300072, China*

[3] *Corresponding author (ctzhang@tju.edu.cn)*

Received July 15, 2004; accepted August 31, 2004; published online November 9, 2004

Summary The *Z*-curve is a three-dimensional curve that constitutes a unique representation of a DNA sequence, i.e., both the *Z*-curve and the given DNA sequence can be uniquely reconstructed from the other. We employed *Z*-curve analysis to identify one replication origin in the *Methanocaldococcus jannaschii* genome, two replication origins in the *Halobacterium* species NRC-1 genome and one replication origin in the *Methanosarcina mazei* genome. One of the predicted replication origins of *Halobacterium* species NRC-1 is the same as a replication origin later identified by in vivo experiments. The *Z*-curve analysis of the *Sulfolobus solfataricus* P2 genome suggested the existence of three replication origins, which is also consistent with later experimental results. This review aims to summarize applications of the *Z*-curve in identifying replication origins of archaeal genomes, and to provide clues about the locations of as yet unidentified replication origins of the *Aeropyrum pernix* K1, *Methanococcus maripaludis* S2, *Picrophilus torridus* DSM 9790 and *Pyrobaculum aerophilum* str. IM2 genomes.

Keywords: Halobacterium, Methanocaldococcus jannaschii, Methanosarcina mazei.

Introduction

The Archaea are a group of prokaryotes that were recognized in 1977 as an independent monophyletic domain of life (Woese and Fox 1977). The evolutionary relationships among the Archaea and the other domains of life, the Bacteria and the Eukarya, are uncertain. However, based on similarities in the proteins involved, the process of replication in archaea appears to be more closely related to that in eukarya than in bacteria (Edgell and Doolittle 1997, Tye 2000, MacNeill 2001, Giraldo 2003, Kelman and Hurwitz 2003). Our understanding of archaeal replication mechanisms has advanced dramatically in the past few years (Bernander 2000, 2003, Kelman 2000, Tye 2000, Bohlke et al. 2002, Grabowski and Kelman 2003, Kelman and Kelman 2003), and it appears that archaea have a simplified version of the eukaryotic replication apparatus. Clarification of the archaeal replication mechanism is therefore important not only to the understanding of archaeal replication, but also for the insight it may provide into the replication mechanisms of eukarya.

Replication initiates bidirectionally at a specific locus called the origin of replication. Knowing the positions and sequences of replication origins is critical to understanding the initiation phase of replication. Replication origins have currently been identified in vivo for only four of the 19 available archaeal genomes (Myllykallio et al. 2000, Maisnier-Patin et al. 2002, Berquist and DasSarma 2003, Matsunaga et al. 2003, Lundgren et al. 2004, Robinson et al. 2004). The experimental methods for identifying replication origins in vivo are reliable, but time-consuming and labor-intensive. *In silico* analysis, however, is fast and suitable for handling a large number of genomes. In addition, in some experimental methods, e.g., as used to identify the replication origin of *Halobacterium* species NRC-1 (Berquist and DasSarma 2003), the replication origin must first be located approximately in a known sequence.

With the advent of the post-genomic era, genomic data are accumulating exponentially. High-throughput methods for genome annotations, e.g., replication origin identification, are thus needed to meet the challenge of interpreting this information. The identification of replication origins based on *in silico* analysis has been the subject of intensive study during the past few years. The GC skew method was first proposed to detect nucleotide composition asymmetry around the replication origin (Lobry 1996*a*). Other algorithms were later proposed to tackle the same task (Grigoriev 1998, McLean et al. 1998, Mrazek and Karlin 1998, Salzberg et al. 1998, Rocha et al. 1999).

The *Z*-curve is a three-dimensional curve that constitutes a unique representation of a DNA sequence, i.e., for the *Z*-curve and the given DNA sequence, each can be uniquely reconstructed from the other (Zhang and Zhang 1991, 1994). We have used *Z*-curve analysis to identify one replication origin in the *Methanocaldococcus jannaschii* genome (Zhang and Zhang 2004*b*), two replication origins in the *Halobacterium* species NRC-1 genome (Zhang and Zhang 2003*c*) and one replication origin in the *Methanosarcina mazei* genome (Zhang and Zhang 2002). One predicted replication origin of *Halobacterium* species NRC-1 is the same as the replication origin later identified by in vivo experiments (Berquist and

DasSarma 2003). The *Z*-curve analysis suggested the existence of three replication origins in the *Sulfolobus solfataricus* P2 genome, and indicated their approximate locations (Zhang and Zhang 2003*c*), the results being consistent with the results of subsequent in vivo studies (Lundgren et al. 2004, Robinson et al. 2004).

This review summarizes past applications of the *Z*-curve in identifying replication origins in archaeal genomes, and applies the same technique in the search for clues about the locations of as yet unidentified archaeal replication origins.

The *Z*-curve representation of genome sequences

The *Z*-curve is a three-dimensional curve that provides a unique representation of a DNA sequence in that the DNA sequence and the *Z*-curve can each be uniquely reconstructed from the other. Therefore, the *Z*-curve contains all the information that the corresponding DNA sequence carries. The resulting curve has a zigzag shape, hence the name *Z*-curve. A DNA sequence can be analyzed by studying the corresponding *Z*-curve. One of the advantages of the *Z*-curve is its intuitiveness; the entire *Z*-curve of a genome can be viewed on a computer screen or on paper, regardless of genome length, thus allowing both global and local compositional features of genomes to be easily grasped. By combining use of the *Z*-curve with statistical analysis, better results may be obtained.

The *Z*-curve is composed of a series of nodes, P_0, P_1, P_2, ..., P_N, with coordinates x_n, y_n and z_n ($n = 0, 1, 2, \ldots, N$, where N is the length of the DNA sequence), which are uniquely determined by the *Z*-transform of a DNA sequence (Zhang and Zhang 1991, 1994, Zhang et al. 2003):

$$\begin{cases} x_n = (A_n + G_n) - (C_n + T_n) \equiv R_n - Y_n, \\ y_n = (A_n + C_n) - (G_n + T_n) \equiv M_n - K_n, \\ z_n = (A_n + T_n) - (C_n + G_n) \equiv W_n - S_n, \end{cases} \quad (1)$$
$$n = 0, 1, 2, ..., N, x_n, y_n, z_n \in [-N, N]$$

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 first base to the nth base in the sequence. We define $A_0 = C_0 = G_0 = T_0 = 0$, and therefore, $x_0 = y_0 = z_0 = 0$. Here R, Y, M, K, W and S represent the purine, pyrimidine, amino, keto, weak hydrogen (H) bond and strong H bond bases, respectively, according to the Recommendation 1984 by the NC-IUB (Cornish-Bowden 1985). The *Z*-curve is defined as the sequential connection of the nodes P_0, P_1, P_2, ..., P_N 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:

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

where $A_n + C_n + G_n + T_n = n$.

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 components x_n, y_n and z_n display the distributions of purine versus pyrimidine (R vs. Y), amino versus keto (M vs. K) and strong H-bond versus weak H-bond (S vs. W) bases along the sequence, respectively. In the subsequence constituted from the first base to the nth base of the sequence, when purine bases (A and G) are in excess of pyrimidine bases (C and T), $x_n > 0$, otherwise, $x_n < 0$, and when the numbers of purine and pyrimidine bases are identical, $x_n = 0$. Similarly, when amino bases (A and C) are in excess of keto bases (G and T), $y_n > 0$, otherwise, $y_n < 0$, and when the numbers of amino and keto bases are identical, $y_n = 0$. Finally, when weak H-bond bases (A and T) are in excess of strong H-bond bases (G and C), $z_n > 0$, otherwise, $z_n < 0$, and when the numbers of weak and strong H-bond bases are identical, $z_n = 0$. The x_n and y_n components are termed RY and MK disparity curves, respectively. 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, respectively, along the genome. The RY and MK disparity curves, as well as AT and GC disparity curves, can be used to predict replication origins. Figure 1 shows an example of the *Z*-curves for the *M. mazei* genome. The *Z*-curve for a genome is a three-dimensional (3-D) curve (Figure 1a). To facilitate the use of the *Z*-curve, it can be plotted as two-dimensional (2-D) curves. Figure 1b is a plot based on RY and MK disparities, whereas Figure 1c is a plot based on AT and GC disparities. The most convenient method, however, is to plot one of the *Z*-curve components, i.e., RY, MK, AT or GC disparities, along the chromosome. Figure 1d shows an AT disparity curve and Figure 2d shows RY and MK disparity curves for the *M. mazei* genome. Arrows indicate the position of *cdc6* genes, and also the putative replication origin. Therefore, in the case of *M. mazei*, all 3-D, 2-D and various disparity curves (RY, MK, AT and GC) show a peak at the position of the putative replication origin.

Replication origin identification in the *Methanocaldococcus jannaschii* genome

Methanocaldococcus jannaschii is an autotroph that grows at pressures greater than 20 MPa and at temperatures up to 94 °C (Jones et al. 1983). As the first completely sequenced archaeon (Bult et al. 1996), *M. jannaschii* is notorious for the difficulty it presents to those seeking to identify its replication origins. Despite extensive efforts, the locations of the replication origins of this species remain elusive 8 years after the publication of its complete genome sequence. Ambiguous results were obtained in identifying the replication origins of *M. jannaschii* based on all *in silico* genome analyses, which usually assess biases in nucleotide, codon and oligomer usages (Salzberg et al. 1998, Lopez et al. 1999, Rocha et al. 1999). Recently, a technique called marker frequency analysis was successfully applied in vivo to identify the location of the replication origin of the archaeon *Archaeoglobus fulgidus*. It failed, however, in the case of *M. jannaschii* (Maisnier-Patin et al. 2002). Distin-

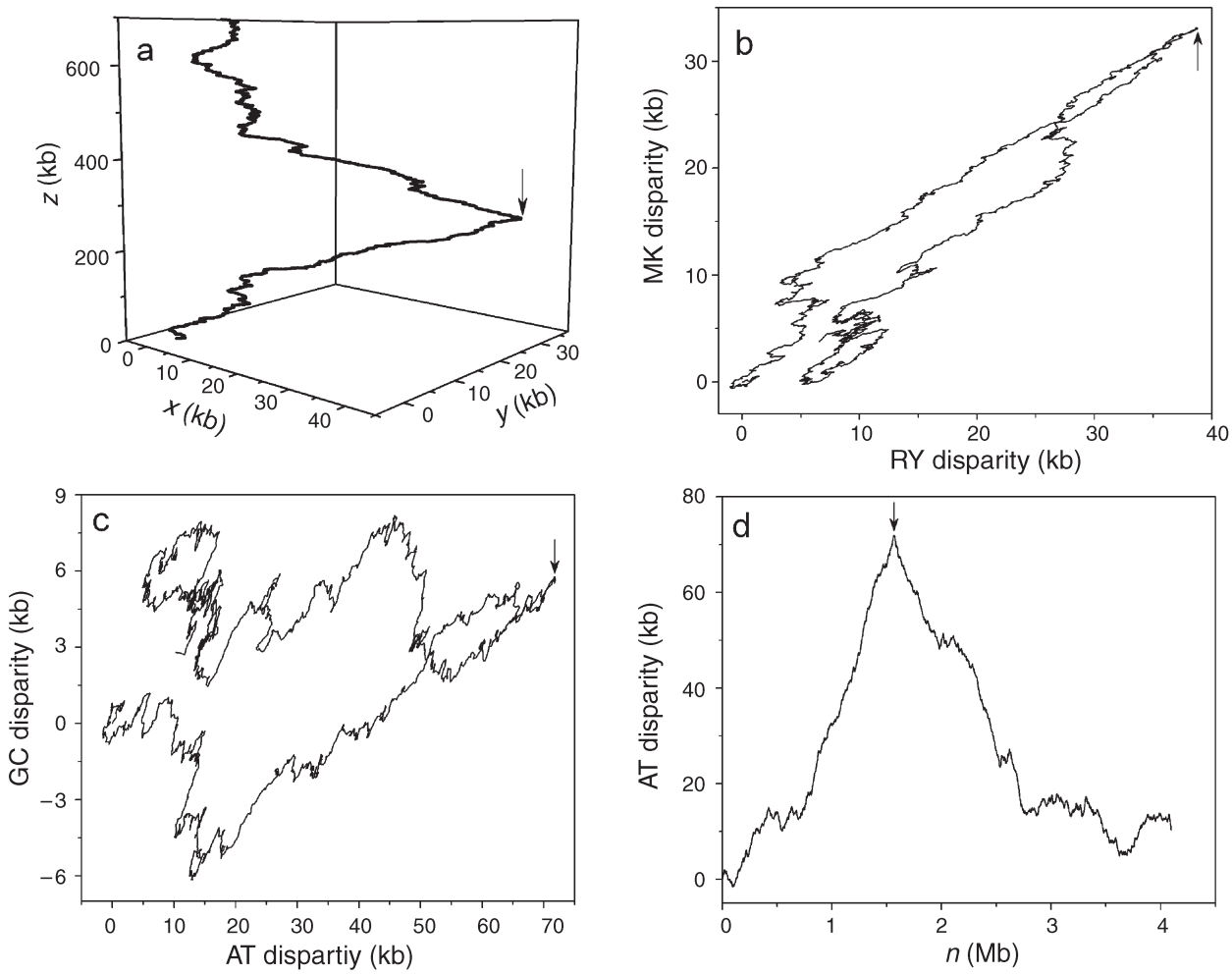

Figure 1. The *Z*-curves for the *Methanosarcina mazei* genome. (a) The 3-D *Z*-curve, (b) the 2-D *Z*-curve based on RY and MK disparity, (c) the 2-D *Z*-curve based on AT and GC disparity and (d) the AT disparity curves. Arrows indicate the positions of the *cdc6* gene, which is also the position of the predicted replication origin.

guishing it from other archaea, the genome of *M. jannaschii* was generally thought to lack a clear *cdc6* homologue (Bernander 2000).

The RY disparity curve for the *M. jannaschii* genome shows a global minimum at the position of about 695 kb, indicating that the genome changes from CT-rich to AG-rich at this site (Figure 2a). Therefore, the site around 695 kb may contain a replication origin. We scanned the region around the minimum for a potential *cdc6* gene. Surprisingly, we found that an open reading frame (ORF), MJ0774, is highly similar to the *cdc6* gene (Zhang and Zhang 2004*b*). The ORF MJ0774 encodes a 409 amino-acid-long polypeptide, and is annotated as a hypothetical protein. We searched the amino acid sequence against the NCBI Conserved Domain Database (Marchler-Bauer et al. 2003), and a Cdc6 protein was assigned to MJ0774, from amino acids 13 to 404. The alignment of the MJ0774 (13–404) with the consensus sequence of Cdc6 proteins (12–355) showed that MJ0774 is a homologue of the Cdc6 protein. In addition, a helix-turn-helix domain was found at the region from residues 327–403, and this domain is believed to be involved in the DNA binding (Liu et al. 2000).

A closer look at the region revealed that an intergenic region of about 700 bp between the *cdc6* homologue and an adjacent gene has many characteristics of a replication origin. This intergenic region is between the ORF MJ0773 and MJ0774, from 694,540–695,226 bp of the genome. The region is 687 bp in length and is highly AT-rich (80%). In addition, there are multiple copies of direct repeat elements and AT stretches. This region contains almost all the features of known replication origins and is, therefore, very likely a true replication origin, which has been designated *oriC1* (Zhang and Zhang 2004*b*).

Recently, marker frequency analysis was successfully applied in vivo to identify the location of a replication origin of *A. fulgidus*. However, *M. jannaschii* displayed a complex pattern of marker frequency distributions with multiple peaks and valleys. An intriguing explanation proposed for this pattern is that it reflects the presence of multiple replication origins (Maisnier-Patin et al. 2002). The features of the MK disparity curve for *M. jannaschii* are consistent with this hypothesis.

The MK disparity curve for *M. jannaschii* shows four ex-

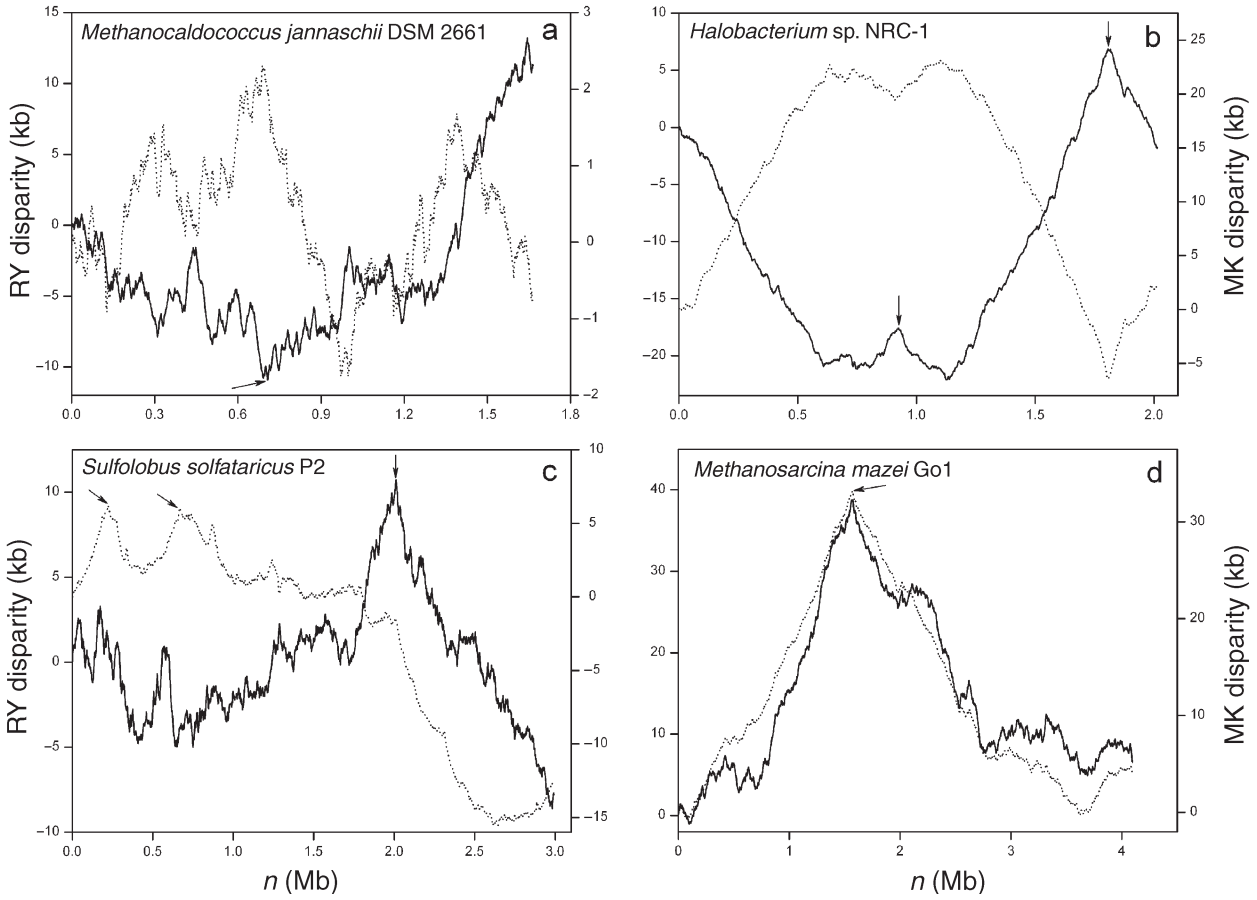

Figure 2. The *Z*-curves for the genomes of (a) *Methanocaldococcus jannaschii* DSM 2661, (b) *Halobacterium* sp. NRC-1, (c) *Sulfolobus solfataricus* P2 and (d) *Methanosarcina mazei* Go1. Unbroken lines denote RY disparity curves, and broken lines denote MK disparity curves. Arrows indicate the positions of *cdc6* genes, which are also the positions of predicted replication origins. In the *Halobacterium* sp. NRC-1 genome, Berquist and DasSarma (2003) have identified a chromosomal autonomously replicating sequence element, which is at the location of the *cdc6-3* (arrow at about 1.8 Mb). Robinson et al. (2004) have identified two replication origins in the *S. solfataricus* genome in vivo. The two replication origins, *oriC1* and *oriC2*, are close to *cdc6-1* and *cdc6-3*, respectively (the positions of the first and third arrows).

tremes, including one probable replication origin associated with the *oriC1* (Figure 2a). The locations of these maxima and minima are 695 (*oriC1*) and 1388 kb, and 127 and 986 kb, respectively. Studying the positions of the four extremes suggests the possibility that the maximum at 1388 kb is associated with another replication origin, whereas the minima at 127 and 986 kb correspond to replication termini. Supporting this hypothesis, the distances between the maximum at 1388 kb and the two predicted replication termini are exactly the same (402 kb), which is consistent with the characteristics of most identified replication origins, i.e., in genomes with a single replication origin, *oriC* and *terC* divide the genome into parts of similar length. However, we also noticed that the distances between the *oriC1* and the two predicted replication termini are different. It is known that some horizontally transferred elements are present in the genome of *M. jannaschii* (Bult et al. 1996). Although the exact amount of horizontally transferred DNA is unclear, these horizontal transfer events could explain why the two replichores have different sizes, i.e., the horizontally transferred DNA increased the length of one of the replichores. In addition, a gene coding for replication factor C (MJ1422) is situated at the position of the maximum associated with the putative *oriC2*. However, there is no evidence to suggest that the gene coding for replication factor C is close to replication origins. Nevertheless, some archaeal replication origins are indeed situated in the regions close to some replication factors, such as DNA polymerases and helicases (Salzberg et al. 1998).

Replication origin identification in the *Halobacterium* species NRC-1 and *Sulfolobus solfataricus* genomes

Halobacterium NRC-1 belongs to the obligatorily halophilic *Halobacterium* species, and is an experimental model among archaea. The exact locations of all replication origins have not been identified, although the possibility of multiple replication origins was suggested based on the GC-skew analysis (Ng et al. 2000, Kennedy et al. 2001).

The RY and MK disparity curves show two relatively sharp and two relatively broad peaks. Interestingly, two of the three *cdc6* genes are located at the positions of the two sharp peaks (Figure 2b). Furthermore, two intergenic regions immediately beside the corresponding *cdc6* genes show many features of replication origins. Therefore, the two intergenic regions were assigned as putative replication origins *oriC1* and *oriC2* (Zhang and Zhang 2003*c*).

The putative replication origin *oriC1* is at the intergenic region close to the *cdc6-1* gene, which is from 921,863–922,014 bp. The *oriC1* contains two long direct repeats. The putative replication origin *oriC2* is at the intergenic region close to the *cdc6-3* gene, which is from 1,806,444–1,807,229 bp. In addition, two helicase genes were located about 20 kb away from these two regions, respectively (Zhang and Zhang 2003*c*). Soon afterwards, a replication origin of *Halobacterium* NRC-1 was identified in vivo by Berquist and DasSarma (2003). These authors found that sequences located up to 750 bp upstream of the *orc7* gene (*cdc6-3*) translational start, plus the *orc7* gene and 50 bp downstream, are sufficient to endow the plasmid with replication ability. Further, they found that the sequence within the 750-bp region upstream of *orc7* contains a nearly perfect inverted repeat of 31 bp, which flanks an extremely AT-rich stretch of 189 bp. The region containing these inverted repeats and AT-rich stretch is within the predicted *oriC2*, 1,806,444–1,807,229 bp (Zhang and Zhang 2003*c*).

A breakthrough in the study of archaeal replication origins was the demonstration that *S. solfataricus* has multiple replication origins. This is the first archaeon found to have multiple replication origins, referred to as *oriC*1 and *oriC2*, according to the nomenclature of Lundgren et al. (2004) and Robinson et al. (2004). The replication origins *oriC1* and *oriC2* are located at sites close to *cdc6-1* and *cdc6-3*, respectively (Robinson et al. 2004). Interestingly, the RY disparity curve for the archaeon *S. solfataricus* shows a global maximum around the position of the *cdc6-3* genes, whereas the MK disparity curve shows a maximum at the position of *cdc6-1* (Figure 2c) (Zhang and Zhang 2003*c*).

Replication origin identification in the *Methanosarcina mazei* genome

The archaeon *Methanosarcina mazei* and related species have great ecological importance, because they are the only organisms that ferment acetate, methylamines and methanol to methane, carbon dioxide and ammonia. Since acetate is the precursor of 60% of the methane produced on Earth, these organisms contribute significantly to the production of this greenhouse gas (Deppenmeier et al. 2002).

Both RY and MK disparity curves for *M. mazei* show a global maximum at about 1600 kb and a minimum at about 3600 kb (Figure 2d). The maximum and minimum correspond to a sharp peak and relatively broad peak, respectively. The *cdc6* gene is located exactly at the global maximum. Based on the known behaviors of the *Z*-curves for archaea whose replication origins have been identified, we hypothesize that the replication origin and termination sites in *M. mazei* correspond to the positions of the sharp and broad peaks, respectively. We have located an intergenic region that is between the *cdc6* gene (MM1314) and the adjacent gene (MM1315), which shows many characteristics of known replication origins. This region is highly AT-rich (74%), and contains multiple copies of consecutive repeats. Our results strongly suggest that the single replication origin of *M. mazei* is situated at the intergenic region between the *cdc6* gene and the adjacent gene, from 1,564,657 to 1,566,241 bp of the genome (Zhang and Zhang 2002).

Common features of archaeal replication origins

So far, replication origins of four archaea have been identified in vivo. Two replication origins have been identified in the *S. solfataricus* P2 genome by 2-D gel analysis (Robinson et al. 2004) and the approximate location of the third was suggested by marker frequency analysis (Lundgren et al. 2004). One replication origin has been identified in *Pyrococcus abyssi* GE5 based on oligomer skew analysis, which was later confirmed in vivo (Lopez et al. 1999, Myllykallio et al. 2000, Matsunaga et al. 2003). An autonomously replicating sequence element has been identified in *Halobacterium* sp. NRC-1 (Berquist and DasSarma 2003). The marker frequency analysis showed a candidate region of a replication origin in *A. fulgidus*; however, the exact location of the replication origin has not been determined (Maisnier-Patin et al. 2002).

Common features of archaeal replication origins can be summarized based on what is known about replication origins identified in vivo. Except that of *A. fulgidus*, all identified replication origins are associated with an extreme in one of the components of the *Z*-curve. In addition, the extremes associated with replication origins are relatively sharp compared with those associated with replication termini, probably because termination sometimes occurs at multiple loci. These replication origins are located immediately beside a *cdc6* gene. This is similar to the case in bacteria, where a gene coding for DnaA is frequently close to the *oriC* (Mackiewicz et al. 2004). Replication origins are highly rich in AT content. The identified replication origins have AT stretches, as well as multiple copies of direct or inverted repeat elements. Furthermore, some replication origins, e.g., those of *S. solfataricus,* contain conserved Cdc6 binding elements.

Based on the above conserved features, some putative replication origins have been identified by *in silico* analysis, but have yet to be confirmed in vivo. These include a replication origin of *Methanothermobacter thermautotrophicus* str. Delta H (Lopez et al. 1999), a replication origin of *Methanosarcina acetivorans* C2A (Galagan et al. 2002), one of the two putative replication origins in *Halobacterium* sp. NRC-1 (Zhang and Zhang 2003*c*), a replication origin in the *M. mazei* genome (Zhang and Zhang 2002) and a replication origin in the *M. jannaschii* genome (Zhang and Zhang 2004*b*). A replication origin of *Pyrococcus furiosus* DSM 3638 and a replication origin

of *Pyrococcus horikoshii* OT3 were identified based on homologue analysis with *Pyrococcus abyssi* (Lopez et al. 1999). In addition, a replication origin of *Thermoplasma acidophilum* DSM 1728 was predicted based on different nucleotide skews; however, other conserved features of archaeal replication origins, e.g., the close proximity to a *cdc6* gene and the presence of repeat elements, were not mentioned (Ruepp et al. 2000). Furthermore, one replication origin of *Methanopyrus kandleri* AV19 was predicted based on the GC-skew analysis; however, the figure of GC-skew provided by the authors does not seem to have a clear minimum or maximum at the site of predicted replication origin (Slesarev et al. 2002). Furthermore, various components of the *Z*-curve show a complex pattern in the case of *M. kandleri* (Figure 3a). The current status of replication origin identification in the 19 available archaeal genomes is listed in Table 1.

Besides the above common features observed among replication origins, there are some differences. For instance, sometimes all disparity curves (MK, RY, AT and GC) show a global maximum or minimum for a given origin, whereas in other cases, only one or a subset of curves shows significant peaks. In addition, in the *A. fulgidus* genome, although an approximate region of replication origin was suggested by marker frequency analysis, both *Z*-curve (Figure 3b) and oligomer skew (Lopez et al. 1999) show no extremes at the site of the replication origin. Furthermore, some replication origins are not associated with *cdc6* genes, e.g., it was suggested that the third replication origin of *S. solfataricus* is about 80 kb away from the nearby *cdc6* gene (Lundgren et al. 2004), but the MK disparity curve shows a maximum at the position of the *cdc6* gene (Figure 2c). It is interesting that although the three replication origins are within the same chromosome, only two of them are close to *cdc6* genes. This may suggest different mechanisms of replication from the three origins. One reviewer of this manuscript noticed that for *S. solfataricus* and *M. jannaschii*, different DNA asymmetry is associated with replication origins. For instance, one replication origin of *S. solfataricus* corresponds to the global maximum of the RY disparity curve, whereas another replication origin corresponds to a maximum of the MK disparity curve. The different behaviors of the *Z*-curve for different replication origins are consistent with the hypothesis that the three replication origins

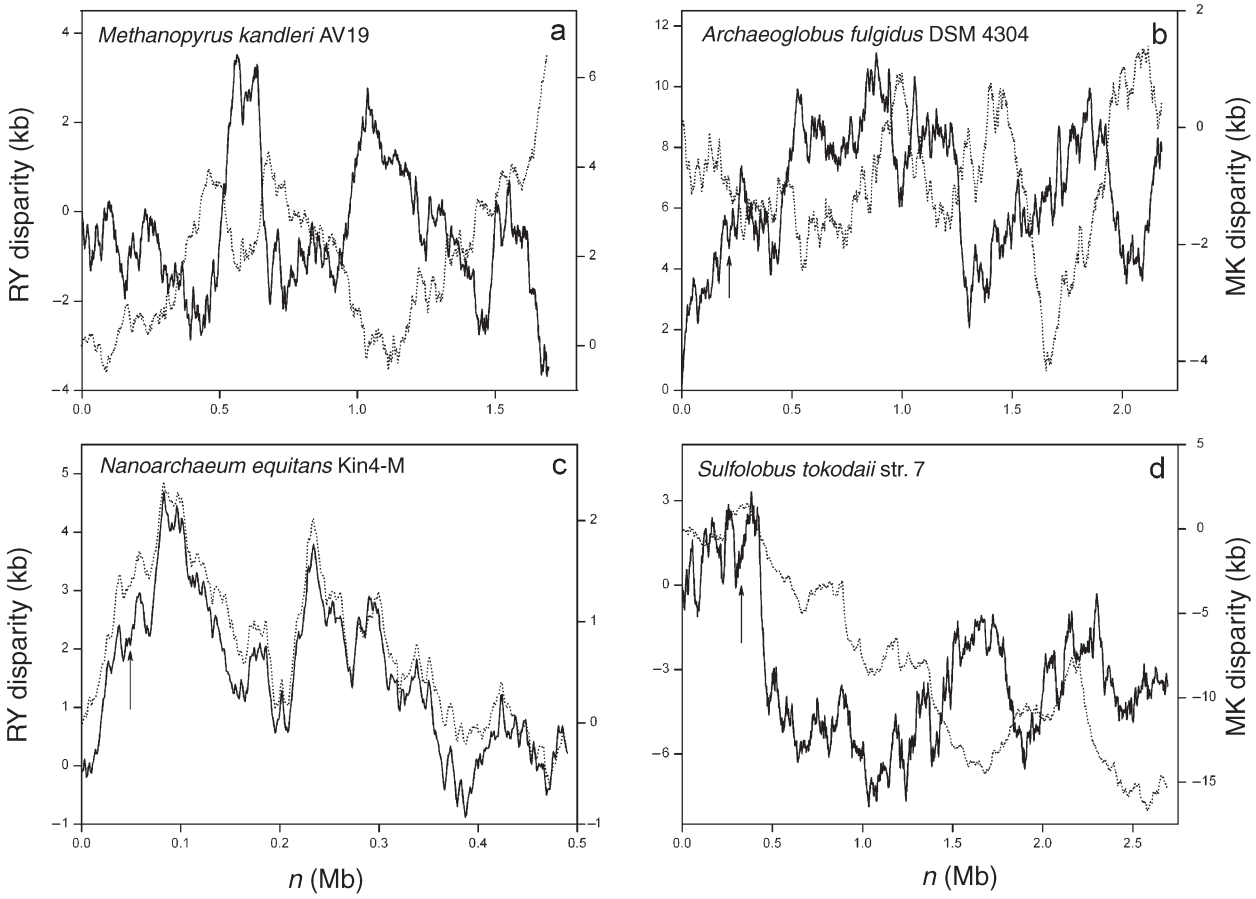

Figure 3. The *Z*-curves for the genomes of (a) *Methanopyrus kandleri* AV19, (b) *Archaeoglobus fulgidus* DSM 4304, (c) *Nanoarchaeum equitans* Kin4-M and (d) *Sulfolobus tokodaii* str. 7. Among the 19 available archaeal genomes, the *Z*-curves for these four genomes show a complex pattern, with no clear global minima or maxima. Unbroken lines denote RY disparity curves, and broken lines denote MK disparity curves. Arrows indicate the positions of *cdc6* genes in the *A. fulgidus*, *N. equitans* and *S. tokodaii* genomes. The approximate location of the replication origin of *A. fulgidus* was suggested to be at about the middle of the chromosome based on marker frequency analysis.

have different replication mechanisms. The close proximity of the *cdc6* gene and replication origin may serve to ensure that the proteins can associate with the origin as soon as they are synthesized (Kelman and Kelman 2003). It is unclear why the third replication origin of *S. solfataricus* is not adjacent to a *cdc6* gene. Lundgren et al. (2004) proposed that one of the three initiation sites might act as the master regulator, with the other two origins being subordinate and therefore different in sequence or organization, or both (Lundgren et al. 2004). Taken together, different *Z*-curve behaviors of the three replication origins of *S. solfataricus* are consistent with the hypothesis that the three replication origins have different replication mechanisms. The absence of a *Z*-curve extreme or a *cdc6* gene cannot exclude the possibility of a replication origin at a certain position of a chromosome.

A reasonable procedure for identifying replication origins by the *Z*-curve method appears to be: (1) generate RY, MK, AT and GC disparity curves for the available genomes; and (2) if there is a minimum or maximum in any of the curves, investigate the regions around each extreme for some replication origin specific features such as the presence of *cdc6* genes or AT-rich intergenic regions that contain repeats.

Z-curve analysis of archaeal genomes with unknown replication origins

In seven out of the 19 available archaeal genomes, replication origins have yet to be identified, and clues to some of their locations have not been found. These seven genomes are *Aeropyrum pernix* K1, *Methanococcus maripaludis* S2, *Nanoarchaeum equitans* Kin4-M, *Picrophilus torridus* DSM 9790, *Pyrobaculum aerophilum* str. IM2, *Sulfolobus tokodaii* str. 7 and *Thermoplasma volcanium* GSS1. Among these seven genomes, the *Z*-curves for *N. equitans* Kin 4-M and *S. tokodaii* str. 7 have a complex pattern, i.e., no global minima or maxima (Figures 3c and 3d).

The RY and MK disparity curves for *T. volcanium* GSS1 show a similar pattern to that of *T. acidophilum* DSM 1728 and have a global minimum and maximum (data not shown), suggesting the presence of a single replication origin. However, no replication origin specific features, such as the presence of a *cdc6* gene, could be found around the *Z*-curve extremes. The *Z*-curves for the remaining four genomes, *A. pernix* K1, *M. maripaludis* S2, *P. torridus* DSM 9790 and *P. aerophilum* str. IM2 show some replication origin-specific features at the extremes, and thus provide additional clues to regions that may contain replication origins.

Robinson et al. (2004) found some conserved Cdc6 binding elements across archaeal genomes. In the *A. pernix* K1 genome, such an element is located at 445 kb of the genome (Robinson et al. 2004). At 445 kb, the GC disparity curve shows a minimum, implying that the nucleotide composition changes around this site (Figure 4a). These lines of evidence suggest the presence of a replication origin around this site.

A putative replication origin has been assigned in the *M. jannaschii* DSM 2661 genome (Zhang and Zhang 2004*b*). A relative of *M. jannaschii* DSM 2661, *M. maripaludis* S2, has been sequenced recently. The AT disparity curve for *M. maripaludis* S2 shows a global minimum, suggesting the presence of a replication origin around this site. In addition, the pattern of the AT disparity curve for *M. maripaludis* is similar to the RY disparity curve of *M. jannaschii* (compare Figures 4b and 2a). However, we could not detect a *cdc6* homologue around the global minimum of the AT disparity curve of the *M. maripaludis* genome. Nevertheless, the conserved pattern of the AT disparity curve suggests the region around the global minimum needs further investigation.

The RY disparity curve for the *P. torridus* DSM 9790 genome shows a global minimum at the position 650 kb (Figure 4c), and a DNA primase gene (PTO0617) is located at the site of the extreme. In addition, immediately beside the primase gene, a 174 bp intergenic sequence between the ORF PTO0617 and PTO0616 has high AT content (81.1%). The MK disparity curve for *P. aerophilum* str. IM2 genome shows a minimum at 662 kb (Figure 4d). Two replication associated genes, a reverse gyrase gene (PAE1108) and a DNA polymerase gene (PAE1113) are all situated around the position of the minimum. In addition to *cdc6*, several replication-related genes are close to archaeal replication origins, e.g., genes encoding DNA polymerases in *M. thermautotrophicus* and *Pyrococcus* species (Lopez et al. 1999, Myllykallio et al. 2000), genes encoding replication factor C and helicases in *Pyrococcus* species (Myllykallio et al. 2000), and a gene encoding radA in *S. solfataricus* (Robinson et al. 2004). Thus, sequences around the 650 kb of the *P. torridus* DSM 9790 genome and the 662 kb of the *P. aerophilum* str. IM2 genome are good candidate regions that may contain replication origins.

Among the 19 available archaeal genomes, the *Z*-curves for the genomes of four species show a complex pattern, with no clear global minima or maxima: *M. kandleri* AV19, *A. fulgidus* DSM 4304, *N. equitans* Kin4-M and *S. tokodaii* str. 7 (Figure 3). *Methanococcus kandleri* has a high evolutionary rate and a surprisingly large number of specific insertions and deletions (Brochier et al. 2004). *Nanoarchaeum equitans* is an obligate symbiont with a small genome (490,885 bp), and is currently the only member of the archaeal kingdom Nanoarchaeota whose genome has been sequenced (Waters et al. 2003). Because of its small size and parasitic reduction, the genome of *N. equitans* may also be fast evolving. In the *S. tokodaii* genome, it was proposed that plasmid integration, rearrangement of genomic structure and duplication of genomic regions have increased the genome size (Kawarabayasi et al. 2001). Furthermore, extensive gene duplications have been found in the *A. fulgidus* genome (Klenk et al. 1997). Therefore, horizontal gene transfer, genome reduction, genome rearrangement and extensive gene duplication may explain the complex pattern of the *Z*-curves for these four genomes. Another possible explanation for the complex pattern is the presence of multiple replication origins in the genomes, or some of the above factors may act together, resulting in the complex pattern of the *Z*-curves.

Table 1. Status of replication origin identification in the currently available archaeal genomes.

	Name (reference)	Order	ID	Length (bp)	Status of replication origin identification (reference)	*Z*-curve extremes	Position of Cdc6 binding element (kb) (Robinson et al. 2004)
1	*Aeropyrum pernix* K1 (Kawarabayasi et al. 1999)	Crenarchaeota	NC_000854	1,669,695	Unknown	Yes	445
2	*Archaeoglobus fulgidus* DSM 4304 (Klenk et al. 1997)	Euryarchaeota	NC_000917	2,178,400	Approximate location is known based on marker frequency analysis (Maisnier-Patin et al. 2002).	No	1430
3	*Halobacterium* sp. NRC-1 (Ng et al. 2000)	Euryarchaeota	NC_002607	2,014,239	Two replication origins have been predicted based on the *Z*-curve and GC skew analysis (Kennedy et al. 2001, Zhang and Zhang 2003*c*). One replication origin has been identified in vivo (Berquist and DasSarma 2003).	Yes	1806
4	*Methanocaldococcus jannaschii* DSM 2661 (Bult et al. 1996)	Euryarchaeota	NC_000909	1,664,970	One replication origin has been identified based on the *Z*-curve analysis (Zhang and Zhang 2004*b*).	Yes	
5	*Methanococcus maripaludis* S2 (Unpublished)	Euryarchaeota	NC_005791	1,661,137	Unknown	Yes	
6	*Methanopyrus kandleri* AV19 (Slesarev et al. 2002)	Euryarchaeota	NC_003551	1,694,969	See footnote [1]	No	
7	*Methanosarcina acetivorans* C2A (Galagan et al. 2002)	Euryarchaeota	NC_003552	5,751,492	One replication origin has been identified based on the GC-skew analysis.	Yes	
8	*Methanosarcina mazei* Go1 (Deppenmeier et al. 2002)	Euryarchaeota	NC_003901	4,096,345	One replication origin has been identified based on the *Z*-curve analysis (Zhang and Zhang 2002).	Yes	
9	*Methanothermobacter thermautotrophicus* str. Delta H (Smith et al. 1997)	Euryarchaeota	NC_000916	1,751,377	One replication origin has been identified based on the oligomer-skew analysis (Lopez et al. 1999).	Yes	
10	*Nanoarchaeum equitans* Kin4-M (Waters et al. 2003)	Nanoarchaeota	NC_005213	490,885	Unknown	No	
11	*Picrophilus torridus* DSM 9790 (Futterer et al. 2004)	Euryarchaeota	NC_005877	1,545,895	Unknown	Yes	
12	*Pyrobaculum aerophilum* str. IM2 (Fitz-Gibbon et al. 2002)	Crenarchaeota	NC_003364	2,222,430	Unknown	Yes	
13	*Pyrococcus abyssi* GE5 (Lecompte et al. 2001, Cohen et al. 2003)	Euryarchaeota	NC_000868	1,765,118	One replication origin has been identified by 2-D gel analysis (Myllykallio et al. 2000, Matsunaga et al. 2003).	Yes	123
14	*Pyrococcus furiosus* DSM 3638 (Robb et al. 2001)	Euryarchaeota	NC_003413	1,908,256	One replication origin has been identified based on homologue analysis with *Pyrococcus abyssi* GE5 (Lopez et al. 1999).	Yes	15

Continued on facing page.

Table 1 continued. Status of replication origin identification in the currently available archaeal genomes.

	Name (reference)	Order	ID	Length (bp)	Status of replication origin identification (reference)	*Z*-curve extremes	Position of Cdc6 binding element (kb) (Robinson et al. 2004)
15	*Pyrococcus horikoshii* OT3 (Kawarabayasi et al. 1998)	Euryarchaeota	NC_000961	1,738,505	One replication origin has been identified based on homologue analysis with *Pyrococcus abyssi* GE5 (Lopez et al. 1999).	Yes	111
16	*Sulfolobus solfataricus* P2 (She et al. 2001)	Crenarchaeota	NC_002754	2,992,245	Two replication origins have been identified in vivo (Robinson et al. 2004). The location of the third replication origin was suggested by microarray-based marker frequency analysis (Lundgren et al. 2004).	Yes	222
17	*Sulfolobus tokodaii* str. 7 (Kawarabayasi et al. 2001)	Crenarchaeota	NC_003106	2,694,756	Unknown	No	323
18	*Thermoplasma acidophilum* DSM 1728 (Ruepp et al. 2000)	Euryarchaeota	NC_002578	1,564,906	One replication origin has been identified based on GC skew analysis (Ruepp et al. 2000).	Yes	
19	*Thermoplasma volcanium* GSS1 (Kawashima et al. 2000)	Euryarchaeota	NC_002689	1,584,804	Unknown	Yes	

[1] It was reported that one replication origin of *Methanopyrus kandleri* AV19 was predicted based on the GC-skew analysis, however, the figure of GC-skew provided by the authors does not seem to have a clear minimum or maximum at the site of predicted replication origin (Slesarev et al. 2002). Various components of the *Z*-curve for *M. kandleri* also show a complex pattern, suggesting that the replication origin predicted by Slesarev et al. (2002) is questionable.

Comparison of the *Z*-curve method with others

Various methods for the graphical representation of DNA sequences have been proposed, such as the H curve (Hamori and Ruskin 1983), the game representation (Jeffrey 1990), color DNA tetragram (Pickover 1992) and the two-dimensional DNA walk (Gates 1986, Lobry 1996*b*). It was shown that most are special cases of the *Z*-curve, and an extensive comparison between the *Z*-curve and other methods proposed before 1994 was detailed in Zhang and Zhang (1994). It is noteworthy that the so-called purine excess and keto excess (Freeman et al. 1998) are identical to the x and y components of the *Z*-curve, which was proposed 4 years earlier (Zhang and Zhang 1994).

Traditionally, the GC skew analysis is often used to assess the nucleotide compositional asymmetry around the replication origin. The GC skew is defined as (C – G)/(C + G), where C and G are the number of C and G residues in a sliding window (Lobry 1996*a*). Later, a method of cumulative GC skew without sliding windows was proposed, which is thought to give better resolution (Grigoriev 1998). Because the *Z*-curve provides a unique representation of a DNA sequence, it contains all the information that the DNA sequence carries. Therefore, the *Z*-curve is not any DNA walk, but almost all DNA walks are special cases of the Z-curve or functions of x_n, y_n and z_n. For instance, the cumulative GC skew is equal to $(y_n - x_n)/(n - z_n)$ (see Equation 1). Indeed, almost all the replication origins that were identified based on the GC skew, including those of bacteria, viruses and mitochondria, are indicated by a change in polarity in the *Z*-curve (Zhang et al. 2003). However, for some genomes, e.g., that of *S. solfataricus*, GC skew failed to show the compositional asymmetry around the replication origins that is detected with the *Z*-curve (Zhang and Zhang 2003*c*).

Availability of the *Z*-curve drawing software

Software has been developed to facilitate the use of the *Z*-curve. The software, Zplotter online, draws and manipulates the *Z*-curve online, based on a user's input sequence. With this software, RY, MK, AT and GC disparity curves can be shown for a user's DNA sequence in the forward (5′ to 3′) and inverted (3′ to 5′) directions and for their complementary strands. The resolution of any local parts of each curve can be arbitrarily adjusted with the built-in zoom function. The *Z*-curve coordinates can also be shown by putting the cursor at the site of interest. In addition, a user can download the local version of the Zplotter program and run it on their own computer. This software is freely available from the *Z*-curve database (Zhang et al. 2003) at http://tubic.tju.edu.cn/zcurve/.

Perspective

In bacteria, replication initiates at a unique site, whereas in eukarya, replication occurs at multiple sites along the genome. A recent breakthrough was the demonstration that the archaeon *S. solfataricus* has at least two replication origins—the

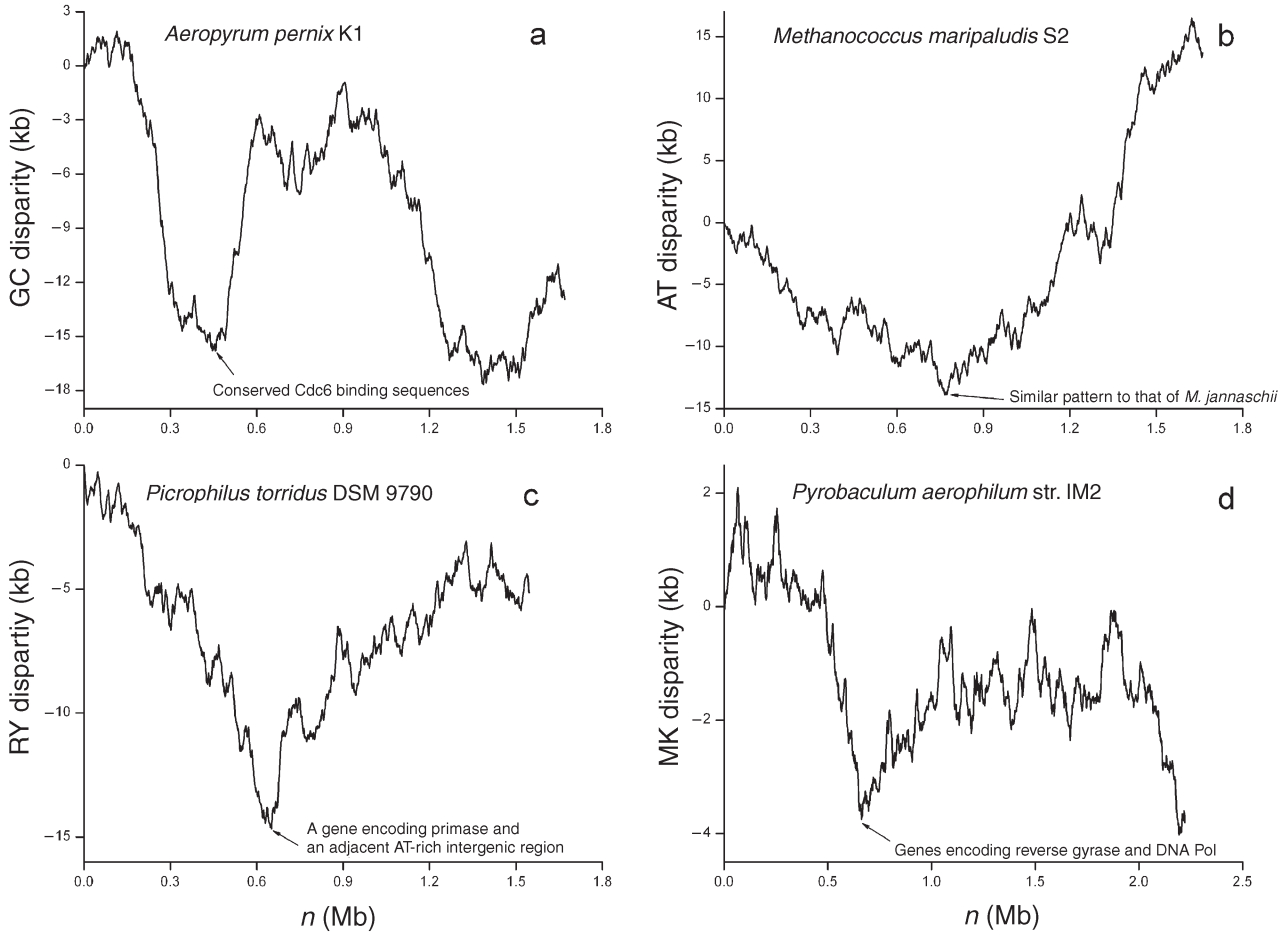

Figure 4. The *Z*-curve analysis for the genomes of *Aeropyrum pernix* K1, *Methanococcus maripaludis* S2, *Picrophilus torridus* DSM 9790 and *Pyrobaculum aerophilum* str. IM2, in which replication origins are unknown. (a) The GC disparity curve for the *A. pernix* K1 genome. Some conserved Cdc6 binding sequences are located at a minimum. (b) The AT disparity curve for the *M. maripaludis* S2 genome. The AT disparity curve shows a global minimum, suggesting the existence of a replication origin around this site. In addition, the overall pattern of the AT disparity curve is similar to the RY disparity curve of the *M. jannaschii* genome. Compare Figure 4b with Figure 2a. (c) The RY disparity curve for the *P. torridus* DSM 9790 genome. A DNA primase gene (PTO0617) is located at the site of the global minimum. In addition, immediately beside this primase gene, a 174 bp intergenic sequence between the ORF PTO0617 and PTO0616 is highly rich in AT content (81.1%). (d) The MK disparity curve for the *P. aerophilum* str. IM2 genome. Genes coding for reverse gyrase and DNA polymerase are located at a minimum. The presence of Cdc6 binding elements, AT-rich intergenic sequence, or replication-associated genes at one of the *Z*-curve extremes provides additional clues for potential candidate regions that may contain replication origins.

first example of the presence of multiple replication origins in archaea (Robinson et al. 2004). Eukaryotic genomes, such as the human genome, have thousands of replication origins, thus complicating the study of replication. In this respect, the simplified version of eukaryotic replication, i.e., archaeal replication that utilizes two or three replication origins, is an excellent model, especially for the study of how the cell coordinates replications occuring at multiple origins. The *Z*-curve analysis for the *Halobacterium* species NRC-1 and *M. jannaschii* shows the possibility that these genomes also have multiple replication origins, and some candidate sites are suggested, e.g., the second replication origin of *Halobacterium* species NRC-1 is suggested to be 921,863–922,014 bp of the genome (Zhang and Zhang 2003*c*, 2004*b*). It is hoped that further in vivo studies will confirm the multiple replication origins in the *Halobacterium* species NRC-1 and *M. jannaschii* genomes.

The *Z*-curve is a powerful tool for *in silico* identification of archaeal and bacterial replication origins. Because the *Z*-curve contains all the information that the corresponding DNA sequence carries, the DNA sequence can be studied by geometrical methods with the *Z*-curve, which is nicely complementary to widely used mathematical methods. Consequently, the *Z*-curve has been used for many purposes in addition to the identification of replication origins. For instance, algorithms based on the *Z*-curve have been used to recognize protein-coding genes in both prokaryotic (Guo et al. 2003) and eukaryotic genomes (Zhang and Wang 2000). Furthermore, it has been shown that the algorithm based on the *Z*-curve is among the best available for gene recognition (Gao and Zhang 2004). The *Z*-curve has also been used in isochore identification (Zhang and Zhang 2003*a*, 2004*a*), detection of horizontally transferred genomic islands (Zhang and Zhang 2004*c*), compara-

tive genomics (Zhang and Zhang 2003*b*), and in studying the distribution of nucleotide composition (Ou et al. 2003). With the availability of an increasing number of complete genome sequences, it is hoped that the *Z*-curve may play a more and more important role in genome research.

Acknowledgments

The present study was supported in part by the 973 Project of China (Grant 1999075606).

References

Bernander, R. 2000. Chromosome replication, nucleoid segregation and cell division in archaea. Trends Microbiol. 8:278–283.

Bernander, R. 2003. The archaeal cell cycle: current issues. Mol. Microbiol. 48:599–604.

Berquist, B.R. and S. DasSarma. 2003. An archaeal chromosomal autonomously replicating sequence element from an extreme halophile, *Halobacterium* sp. strain NRC-1. J. Bacteriol. 185: 5959–5966.

Bohlke, K., F.M. Pisani, M. Rossi and G. Antranikian. 2002. Archaeal DNA replication: spotlight on a rapidly moving field. Extremophiles 6:1–14.

Brochier, C., P. Forterre and S. Gribaldo. 2004. Archaeal phylogeny based on proteins of the transcription and translation machineries: tackling the *Methanopyrus kandleri* paradox. Genome Biol. 5: R17.

Bult, C.J., O. White, G.J. Olsen et al. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science 273:1058–1073.

Cohen, G.N., V. Barbe, D. Flament et al. 2003. An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*. Mol. Microbiol. 47:1495–1512.

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

Deppenmeier, U., A. Johann, T. Hartsch et al. 2002. The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. J. Mol. Microbiol. Biotechnol. 4:453–461.

Edgell, D.R. and W.F. Doolittle. 1997. Archaea and the origin(s) of DNA replication proteins. Cell 89:995–998.

Fitz-Gibbon, S.T., H. Ladner, U.J. Kim, K.O. Stetter, M.I. Simon and J.H. Miller. 2002. Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. Proc. Natl. Acad. Sci. USA 99:984–989.

Freeman, J.M., T.N. Plasterer, T.F. Smith and S.C. Mohr. 1998. Patterns of genome organization in bacteria. Science 279:1827.

Futterer, O., A. Angelov, H. Liesegang, G. Gottschalk, C. Schleper, B. Schepers, C. Dock, G. Antranikian and W. Liebl. 2004. Genome sequence of *Picrophilus torridus* and its implications for life around pH 0. Proc. Natl. Acad. Sci. USA 101:9091–9096.

Galagan, J.E., C. Nusbaum, A. Roy et al. 2002. The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity. Genome Res. 12:532–542.

Gao, F. and C.T. Zhang. 2004. Comparison of various algorithms for recognizing short coding sequences of human genes. Bioinformatics 20:673–681.

Gates, M.A. 1986. A simple way to look at DNA. J. Theor. Biol. 119: 319–328.

Giraldo, R. 2003. Common domains in the initiators of DNA replication in bacteria, archaea and eukarya: combined structural, functional and phylogenetic perspectives. FEMS Microbiol. Rev. 26: 533–554.

Grabowski, B. and Z. Kelman. 2003. Archaeal DNA replication: eukaryal proteins in a bacterial context. Annu. Rev. Microbiol. 57: 487–516.

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

Guo, F.B., H.Y. Ou and C.T. Zhang. 2003. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. Nucleic Acids Res. 31:1780–1789.

Hamori, E. and J. Ruskin. 1983. H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences. J. Biol. Chem. 258:1318–1327.

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

Jones, W.J., J.A. Leigh, F. Mayer, C.R. Woese and R.S. Wolfe. 1983. *Methanococcus jannaschii* sp. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent. Arch. Microbiol. 136:254–261.

Kawarabayasi, Y., M. Sawada, H. Horikawa et al. 1998. Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res. 5: 55–76.

Kawarabayasi, Y., Y. Hino, H. Horikawa et al. 1999. Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. DNA Res. 6:83–101, 145–152.

Kawarabayasi, Y., Y. Hino, H. Horikawa et al. 2001. Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain 7. DNA Res. 8:123–140.

Kawashima, T., N. Amano, H. Koike et al. 2000. Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*. Proc. Natl. Acad. Sci. USA 97:14,257–14,262.

Kelman, Z. 2000. The replication origin of archaea is finally revealed. Trends Biochem. Sci. 25:521–523.

Kelman, Z. and J. Hurwitz. 2003. Structural lessons in DNA replication from the third domain of life. Nat. Struct. Biol. 10:148–150.

Kelman, L.M. and Z. Kelman. 2003. Archaea: an archetype for replication initiation studies? Mol. Microbiol. 48:605–615.

Kennedy, S.P., W.V. Ng, S.L. Salzberg, L. Hood and S. DasSarma. 2001. Understanding the adaptation of *Halobacterium* species NRC-1 to its extreme environment through computational analysis of its genome sequence. Genome Res. 11:1641–1650.

Klenk, H.P., R.A. Clayton, J.F. Tomb et al. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature 390:364–370.

Lecompte, O., R. Ripp, V. Puzos-Barbe, S. Duprat, R. Heilig, J. Dietrich, J.C. Thierry and O. Poch. 2001. Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. Genome Res. 11:981–993.

Liu, J., C.L. Smith, D. DeRyckere, K. DeAngelis, G.S. Martin and J.M. Berger. 2000. Structure and function of Cdc6/Cdc18: implications for origin recognition and checkpoint control. Mol. Cell 6: 637–648.

Lobry, J.R. 1996*a*. Asymmetric substitution patterns in the two DNA strands of bacteria. Mol. Biol. Evol. 13:660–665.

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

Lopez, P., H. Philippe, H. Myllykallio and P. Forterre. 1999. Identification of putative chromosomal origins of replication in archaea. Mol. Microbiol. 32:883–886.

Lundgren, M., A. Andersson, L. Chen, P. Nilsson and R. Bernander. 2004. Three replication origins in *Sulfolobus* species: synchronous initiation of chromosome replication and asynchronous termination. Proc. Natl. Acad. Sci. USA 101:7046–7051.

Mackiewicz, P., J. Zakrzewska-Czerwinska, A. Zawilak, M.R. Dudek and S. Cebrat. 2004. Where does bacterial replication start? Rules for predicting the *oriC* region. Nucleic Acids Res. 32:3781–3791.

MacNeill, S.A. 2001. Understanding the enzymology of archaeal DNA replication: progress in form and function. Mol. Microbiol. 40:520–529.

Maisnier-Patin, S., L. Malandrin, N.K. Birkeland and R. Bernander. 2002. Chromosome replication patterns in the hyperthermophilic euryarchaea *Archaeoglobus fulgidus* and *Methanocaldococcus* (*Methanococcus*) *jannaschii*. Mol. Microbiol. 45:1443–1450.

Marchler-Bauer, A., J.B. Anderson, C. DeWeese-Scott et al. 2003. CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Res. 31:383–387.

Matsunaga, F., C. Norais, P. Forterre and H. Myllykallio. 2003. Identification of short 'eukaryotic' Okazaki fragments synthesized from a prokaryotic replication origin. EMBO Rep. 4:154–158.

McLean, M.J., K.H. Wolfe and K.M. Devine. 1998. Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. J. Mol. Evol. 47:691–696.

Mrazek, J. and S. Karlin. 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. USA 95: 3720–3725.

Myllykallio, H., P. Lopez, P. Lopez-Garcia, R. Heilig, W. Saurin, Y. Zivanovic, H. Philippe and P. Forterre. 2000. Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon. Science 288:2212–2215.

Ng, W.V., S.P. Kennedy, G.G. Mahairas et al. 2000. Genome sequence of *Halobacterium* species NRC-1. Proc. Natl. Acad. Sci. USA 97: 12,176–12,181.

Ou, H.Y., F.B. Guo and C.T. Zhang. 2003. Analysis of nucleotide distribution in the genome of *Streptomyces coelicolor* A3(2) using the *Z* curve method. FEBS Lett. 540:188–194.

Pickover, C.A. 1992. DNA and protein tetragrams: biological sequences as tetrahedral movements. J. Mol. Graph. 10:2–6, 17.

Robb, F.T., D.L. Maeder, J.R. Brown, J. DiRuggiero, M.D. Stump, R.K. Yeh, R.B. Weiss and D.M. Dunn. 2001. Genomic sequence of hyperthermophile, *Pyrococcus furiosus*: implications for physiology and enzymology. Methods Enzymol. 330:134–157.

Robinson, N.P., I. Dionne, M. Lundgren, V.L. Marsh, R. Bernander and S.D. Bell. 2004. Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*. Cell 116:25–38.

Rocha, E.P., A. Danchin and A. Viari. 1999. Universal replication biases in bacteria. Mol. Microbiol. 32:11–16.

Ruepp, A., W. Graml, M.L. Santos-Martinez et al. 2000. The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. Nature 407:508–513.

Salzberg, S.L., A.J. Salzberg, A.R. Kerlavage and J.F. Tomb. 1998. Skewed oligomers and origins of replication. Gene 217:57–67.

She, Q., R.K. Singh, F. Confalonieri et al. 2001. The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc. Natl. Acad. Sci. USA 98:7835–7840.

Slesarev, A.I., K.V. Mezhevaya, K.S. Makarova et al. 2002. The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. Proc. Natl. Acad. Sci. USA 99:4644–4649.

Smith, D.R., L.A. Doucette-Stamm, C. Deloughery et al. 1997. Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics. J. Bacteriol. 179:7135–7155.

Tye, B.K. 2000. Insights into DNA replication from the third domain of life. Proc. Natl. Acad. Sci. USA 97:2399–2401.

Waters, E., M.J. Hohn, I. Ahel et al. 2003. The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. Proc. Natl. Acad. Sci. USA 100:12,984–12,988.

Woese, C.R. and G.E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. USA 74:5088–5090.

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

Zhang, R. and C.T. Zhang. 1994. *Z* curves, an intutive tool for visualizing and analyzing the DNA sequences. J. Biomol. Struct. Dyn. 11:767–782.

Zhang, R. and C.T. Zhang. 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 R. Zhang. 2003*a*. An isochore map of the human genome based on the *Z* curve method. Gene 317:127–135.

Zhang, R. and C.T. Zhang. 2003*b*. Identification of genomic islands in the genome of *Bacillus cereus* by comparative analysis with *Bacillus anthracis*. Physiol. Genomics 16:19–23.

Zhang, R. and C.T. Zhang. 2003*c*. Multiple replication origins of the archaeon *Halobacterium* species NRC-1. Biochem. Biophys. Res. Commun. 302:728–734.

Zhang, C.T. and R. Zhang. 2004*a*. Isochore structures in the mouse genome. Genomics 83:384–394.

Zhang, R. and C.T. Zhang. 2004*b*. Identification of replication origins in the genome of the methanogenic archaeon, *Methanocaldococcus jannaschii*. Extremophiles 8:253–258.

Zhang, R. and C.T. Zhang. 2004*c*. A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* CMCP6 chromosome I. Bioinformatics 20:612–622.

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