ISSN 0974 - 9373

Vol. 14 No.4 (2010) **Journal of International Academy of Physical Sciences** pp. 531-541

In silico Detection of Origins of Replication in Bacteria and Archaea

Krishna Kumar Ojha
Department of Bioinformatics
Banaras Hindu University, Varanasi -221005
E-mail: krisbhu@gmail.com

D. Swati
Department of Bioinformatics & Physics
MMV, Banaras Hindu University, Varanasi -221005
E-mail: swatid@gmail.com

(Received February 20, 2010)

Abstract: In this paper we have studied the replication machinery in three domains of life namely, Bacteria, Archaea and Eukaryotes. We have discussed the *in silico* methods which are being used for locating *Ori* (origin of replication) sites in Bacteria and Archaea. We have also shown that some nucleotide skew method namely GC skew is good enough to locate the *Ori* site in bacteria but it fails in Archaea because the Archaea do not have pronounced nucleotide skew. It has been observed that combined strategy of cumulative GC skew and DnaA boxes location can improve results for bacterial genomes. Data show that bacterial replication proteins tend to lie in cluster adjacent to each other. The use of Z curve, location of cdc6 gene, upstream AT rich region and consensus ORB sequences can help us to increase the confidence level in detecting the *Ori* sites in Archaea. We have also found that copy number and position of cdc6 gene correlated directly with the *Ori* site in Archaea.

Keywords: Z curve, disparity plot, *Ori* site, ORC, inverted repeats

Abbreviation: *Ori* (origin of replication site), Cdc6 (cell division cycle6), MCM (mini chromosome maintenance), ORC (origin recognition complex), ORB (origin recognition boxes)

1. Introduction

DNA replication is an integral part of cell cycle, essential to life. Chromosomal DNA must be duplicated during the cell proliferation to maintain a constant level of genetic material[1]. To ensure the high fidelity of the DNA replication, cells have a complicated mechanism to coordinate the replication process and initiation of DNA replication is one among them. Before DNA synthesis starts, the origin of DNA replication is recognized by

a specific factor(s)[2]. DNA must be unbound before the replication machinery is assembled, because DNA synthesis occurs on single stranded DNA.

In all prokaryotes DNA replication is thought to initiate at well defined chromosomal sites. Bacterial chromosomes typically carry a single replication site called *OriC* having a cognate initiator protein *dnaA* which bind to repeated sequences (DnaA box)[3]. The dnaA binding induces topological changes in the *Ori* region and duplex unwinding occurs at a region, which is rich in adenine and thymine. This process is dependent on the ATP bound form of DnaA. In following step, dnaB, the replicative helicase is loaded into the unwound site to extend the single stranded region. Then, DNA primase and DNA polymerase initiate DNA synthesis on the single stranded DNA.

Eukaryotic chromosomes have multiple origin replication sites, which are bound by a multi-protein origin replication complex (ORC) composed of six subunits (Orc1-6),that bind to the origin of replication[4]. The binding of ORC to replication region causes topological alteration in DNA strand. This is an ATP dependent process. Archaea, which are less studied have single as well as multi origins of replication. Archaea share more similarities with eukaryotes, since almost all the proteins involved in archaeal DNA replication have their homologous counterparts in Eukaryotes. The origin of DNA replication (*Ori*) is likely to be located in an intergenic region upstream region of cdc6 gene which is AT rich and contains many direct and inverted repeats. This intergenic region contains several ORB and mini ORB sequences which are recognized by cdc6 gene. The *Ori* in *Pyrococcus* species consist of two ORB repeats, several mini ORB repeats and AT rich region[5]. All sequenced genome of archaeal species have at least one copy of cdc6 gene. The binding of cdc6 gene on *Ori* site causes conformational changes. Though replication machinery is almost similar in all archaea, there is a major difference in the replication sites in different archaeal groups. In particular, the chromosome of *Pyrococcus abyssi* (*Euryarchaeota*) is replicated from a single *Ori*, whereas three different *Ori* are used to replicate the single chromosome of *Sulfolobus solfataricus* (*Crenarchaeota*). *Methanogens* (*Euryarchaeota*) are the only group within the archaea where replication origins have not been experimentally detected yet. With the advent of the post-genomic era, genomic data are accumulating exponentially. Out of 65 archaeal genomes deposited on NCBI there are only 6 for which the *Ori* sites has been located experimentally, so there is a big gap between the sequenced genome and experimentally determined *Ori* sites and to overcome this situation we do have to rely on *in silico* approach which is computationally fast and cost effective. GC skew and Z curve have been used to predict the origin of replication in bacteria[6, 7], but they have also

their limitations. GC skew works well in finding *Ori* in most of the bacterias but it fails on some like the cyanobacteria *Nostoc sp.* and *Thermosynechococcus* elongates. Similarly Z curve analysis has been used to predict origins of replications for a number of archaeal species including *M. thermoautotrophicus* and *Methanosarcina mazei*[8], however, these techniques have not been useful in other archaea, such as *Methanococcus jannaschii* and *Sulfolobus solfatricus*[8]. Whether or not predictions exist, experimental data are still required to confirm whether putative origins act as a site of replication initiation *in vivo.*

The replication machinery of bacteria and archaea is quite different. Here we have compared the replication proteins of both the domains. We can easily see from the Table 1 that replication machinery of archaea is more similar to eukaryotes than that of the bacteria**.** Carl Woese in his pioneering work distinguished archaea as the third domain of life on the basis of its transcription machinery[9]. This distinction as a domain disparate from both the bacteria as well eukaryotes is seen in the replication machinery also.

2. Material and Methods

Genome sequences of different archaea and bacteria have been downloaded from NCBI (http://www.ncbi.nlm.nih.gov). PDB sequence for dnaA protein and cdc6 proteins were downloaded from protein data bank (www.rcsb.org). Structural, functional and biochemical study of the proteins was performed with Abalone software package. A program script for plotting GC skew and Z curve from DNA sequence has been written on MATLAB (http://www.mathworks.com). GC skew is the skew calculated by counting the excess of G over C and divided by G+C (G- C/ G+C) in a sliding window of small length say 100 bp. This window is used for scanning the whole genome and cumulative score of this GC score for each windows is plotted for the entire genome. Z curve is a three dimensional space curve constituting the unique representation of a DNA sequence which has all the property of a given sequence. We present a briefly method of Z curve as follows.

Suppose we have a DNA sequence reading from 5' – 3' having N bases long. Beginning from the first sequence inspects the sequence one at a time. Let the number of steps be n, where n = 1, 2, 3, 4……..N. In n^{th} step count the cumulative number of bases A, T, G and C occurring in the subsequence from 1^{st} to the n^{th} base in the DNA sequence analyzed.We define the Z curve as consisting of a series of node P_n where n = 0, 1, 2, 3…. N, whose coordinates are represented by X_n, Y_n, and Z_n respectively.

$$X_n = (A_n + G_n) - (C_n + T_n) \equiv R_n - Y_n$$

$$Y_n = (A_n + C_n) - (G_n + T_n) \equiv M_n - K_n$$

$$Z_n = (A_n + T_n) - (G_n + C_n) \equiv W_n - S_n$$

Here R, Y, M, K, W, and S represent the bases of purine, pyrimidine, amino, keto, weak hydrogen bonds, and strong hydrogen bonds respectively. The three axis of the Z curve X, Y, and Z show the disparity curve of the purine- pyrimidine, amino–keto and weak- strong respectively. The Z curve described above is a three-dimensional space curve, having three independent components, i.e. Xn, Yn and Zn. Each component has a clear biological meaning. The Z curve is an intuitive and convenient approach to study DNA sequences. By viewing the Z curve, some global and local features of the sequence can be detected in a perceivable way.

3. Results

Replicon unit	Bacteria	Archaea	Eukarya
Chromosome	Linear or circular	Circular	Linear
Replication origin(s)	Single	Single or multiple	Multiple
Origin recognition	DnaA	Cdc6/ORC	ORC
Helicase	DnaB	MCM	MCM
Helicase loader	DnaA and DnaC	Cdc6/ORC	Cdc6 and Cdt1
Single-stranded DNA-binding protein	SSB	SSB or RPA	RPA
Primase	DnaG	Primase	Polα/Primase
Sliding clamp	β subunit	PCNA	PCNA
Clamp loader	γ-complex	RFC	RFC
Polymerase	PolC	PolD and/or PolB	PolB

Bacteria and archaea use different replication machinery. From the table (1) it is clear that archaeal replication proteins are similar to eukaryotes, so we may conjecture that their replication process should also be similar to the process in eukaryotes. Some archaea fulfill this assumption and show multiple origin of replication as in eukaryotes while some members shows

single origin of replication.he GC skew analysis of most of the bacteria shows that leading and lagging strand have different nucleotide composition. Generally leading strand is rich in GC while lagging strand is rich in AT. This bias is probably due to differential mutation rate on leading and lagging strand.

All the bacterial genomes analyzed show change in skew at the site of origin and termination, a sharp peak in GC skew which divides the genome in almost two equal segments. The location of this peak is the putative *Ori* site. Two bacterial genome (*E. coli* and *M. tuberculosis*) have been analyzed on basis of the GC skew (Figure 1 & 2). In *E. coli* we see a sharp peak at about 1.7 Mb in positive direction from the initial nucleotide and a sharp skew in negative direction which is about 3.9 Mb apart. It has been experimentally proved that the sharp peak in positive direction is the *Ori* site in *Escherichia coli.* In *Mycobacterium tuberculosis* we see a sharp peak in positive direction at about 2.4 Mb from initial nucleotide. Both these graphs cut the GC skew disparity zero axis only once, which shows that they have single origin of replication.

Figure 1

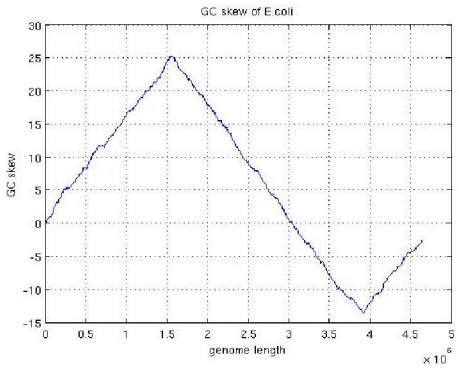
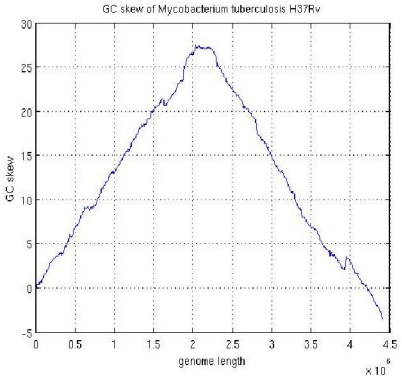


Figure 2



.

Figure (1 & 2) GC disparity curve of *Escherichia coli* and *Mycobacteriuam tuberculosis*. Both curve show a sharp peak in positive direction and cross the zero axis at once

When we analyzed the gene position involved in replication machinery we found that most of the genes responsible for the replication in E. coli tends to lie in cluster, very close to each other, as shown Figure 3 (from NCBI site

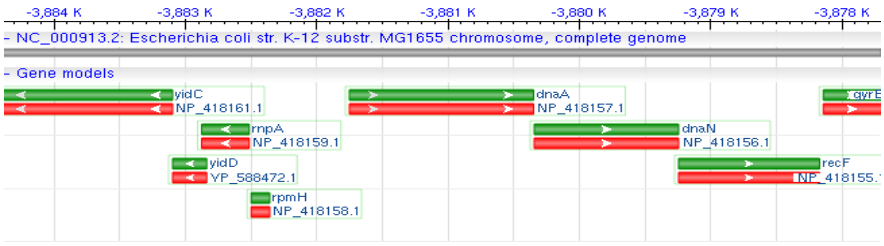


Figure 3: dnaA protein location in E.coli K-12, which is adjacent to DNA polymerase III beta subunit (dnaN) and gap repair protein (recF).

Dna A protein lies at 3880 kb on negative strand, three other proteins which also play important role in replication process namely, DNA pol. III beta sub unit and gap repair protein also lie adjacent to the dnaA protein[9].

The GC skew approach fails in finding the *Ori* site in Archaea, because they do not show well-defined strand asymmetry. So for the archaeal genome we have used the Z curve approach which is composed of three components X, Y and Z each having different biological role.

For all archaeal genome analyzed by Z curve, we get two kind of graph one having the smooth sharp extreme and other having the fuzzy appearance. The *Ori* is likely to be situated in the region of extreme either minimum or maximum. In all archaea for which the replication origins have been identified experimentally (*Sulfolobos solafataricus*) the three-dimensional Z curves and its X or Y component show an abrupt turn at the site of replication origin. But this does not guarantee the exact location of *Ori* site. Moreover there are several archaea, for example, *Aeropyrum pernix* for which Z curve plot does not give a global peak and it gives a random graph of X and Y component. So we used the cdc6 gene location. Cdc6 is the initial proteins which comes and binds with origin recognition boxes. To fine-tune the location of the putative *Ori*, the adjacent 'AT' rich region was also located and ORB consensus sequence location used to enhance the accuracy. Some archaea have single origin of replication (*P. abyssi*, *P.furiousus*) ,while most have multiple origin of replication (*H. salinarum, S. acidocaldarius)*. We have also shown that cdc6 gene copy number is, to some extent related with number of *Ori* sites. Generally archaea which have single origin of replication have single copy of cdc6 gene in their genome, while several copies of cdc6 gene are reported (shown in Table 2) for archaea having multiple *Ori* sites.

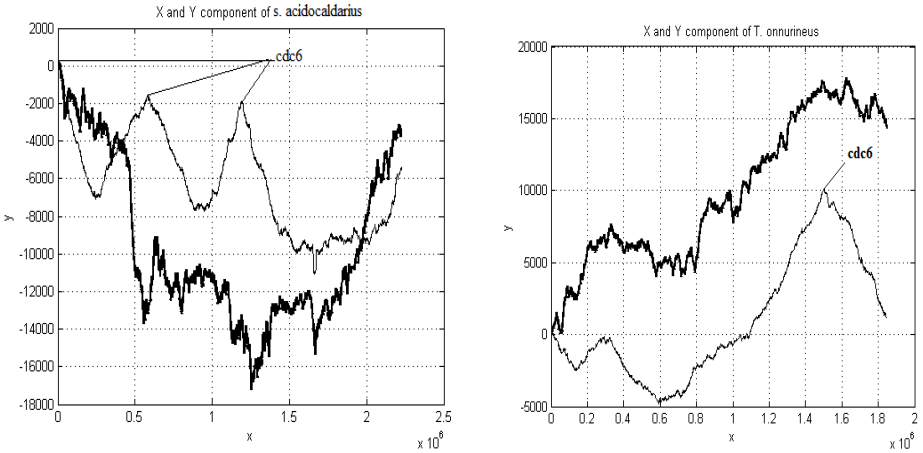


Figure (4,5) Figure 3 at left is for *S. acidocaldarius* and Figure 4 at right is for *T. onnurineus,* which shows X and Y component of Z curve. X component is shown in dark line while Y component is in light gray line. The cdc6 gene location is indicated by narrow lines; at this point abrupt changes occur in the nucleotide skew value.

The genome of *Thermococcus onnurineus* was sequenced in 2008.The X and Y axis of Z curve for *Thermococcus onnurineus* is shown in Figure 5. The Y axis of the Z curve shows a sharp global peak at 1500 kb and a broad peak in opposite direction. The location of cdc6 gene is very close to this peak. The location of cdc6 gene is 1508116bp to 1509363bp on the negative(reverse) strand. The x axis also shows a peak at this point. The upstream region of the cdc6 gene shows high AT density (about 76%) which is a favorable condition for *Ori* site, because AT has two H bond which is easy to break as compared to GC having three. The upstream region of cdc6 gene has an intergenic region which contains a 30-mer direct repeat 3'ATGTGCTCCACAGGAAGCACCGGATACAGA 5'. According to the behavior of the known Z curve we predict that *T. onnurinneus* has a single origin of replication situated at about 1.5 MB apart from initial nucleotide.

Table 2

S. No	Classification	Organisms	Genome size(Mb)	GC %	Opt. Growth Temp. (in ^{0}C)	No. of cdc6 genes	Location cdc6 gene	cdc6 gene length (Bp)
1	**Crenarcheeota**	*Aeropyrum pernix K1*	1.669	56.3	95	2	**1** 112111 - 113343	1232
							2 331447 - 332346	899
2		*Desulfurococcus kamchatkensis 1221n*	1.365	45.3	85	2	**1**127992 - 1281096	1184
							21323734 - 1324978	1244
3		*Metallosphaera sedula DSM 5348*	2.191	46.2	65	3	**1** 331 – 1521	1190
							2 638445 - 639677	1232
							31969173 - 1970426	1253
4		*Sulfolobus acidocaldarius DSM 639*	2.225	36.7	80	3	**1** 101 – 1261	1160
							2 578164 to 579357	1193
							3 724282 to 725529	1244
5	**Euryarchaeota**	*Haloarcula marismortui ATCC 43049 chromosome I*	3.131	62.4	40	8	**1** 482379 – 483554	1175
							2 937406 – 938680	1274
							3 1128689 - 1130017	1328
							4 1401414 – 1402625	1211
							5 1402965 – 1403417	452
							6 2124334 - 2125386	1052
							7 2414031 - 2415605	1575
							8 2542392 - 2543519	1127
6		*Haloarcula marismortui ATCC 43049 chromosome II*	0.288	57.2	40	2	**1** 50205 – 51485	1280
							2 6826 - 8064	1238

7		*Halobacterium sp. NRC-1*	2.014	67.9	42	5	**1** 38017 – 39432	1415
							2 1448861 – 1450033	1172
							3 1691265 - 1692389	1124
							4 1807230 - 1808786	1556
							5 920669 - 921862	1556
8		*Pyrococcus abyssi*	1.765	44.7	100	1	121402 - 122700	1298
9		*Methanocaldococcus jannaschii*	1.664	31.42	85	1	**1** 695227 - 696456	1229
10		*Methanococcus maripaludis C5*	1.780	33	82	1	**1**1583988 - 1585214	1226
11	**Korarchaeota**	*Candidatus Korarchaeum cryptofilum OPF8*	1.590	49	65	2	**1** 1016339-1017487	1148
							2 1378915 - 1380105	1190
12	**Thaumarchaeota**	*Nitrosopumilus maritimus SCM1*	1.645	34	55	2	**1** 146 - 1348	1202
							1542607 - 1643926	1319

4. Discussion

Bacteria and archaea have similar morphology and habitat and both can thrive in inhospitable conditions. Also their metabolic machinery is almost the same. The major difference is in the information processing machinery. The archaea are more akin to eukaryotes in their genome replication machinery. Almost all proteins involved in the replication process have homologues in eukaryotes or eukarya. Sequence statistical feature like GC skew can be used to map the *Ori* site in bacteria with great confidence, because they have strand asymmetry and their leading and lagging strand shows inequality in nucleotide distribution[10]. But there are exceptions in several bacteria that do not have nucleotide skew and GC skew approach to find *Ori* in these organisms fails. Nucleotide skew is due to two reasons- the first reason is difference in mutational rate on leading and lagging strand and second is single origin of replication due to which nucleotide skew is well-defined and shows up on the skew or disparity plots with clarity. In bacteria the mutation repair system is not robust so they are more prone to mutation. It is also shown in analysis that a dnaA protein lies adjacent to other proteins which are involved in DNA replication[11].

Archaeal replication machinery is similar to the eukaryotes; their error repair machinery is very robust and therefore the rate of mutation is very slow. Because some archaea have more than one *Ori* sites, the strand-switch of GC-skew and R-Y skew is very frequent muting the signal for the location of *Ori* based on nucleotide skew[11]. Z-curve with some level of certainty can reveal the overall and local features of the sequence in a perceivable way. While the Z-curve method has been used to find *Ori* site in some archaeal genomes, it is not appropriate for most of the archaeal genomes which lack clear cut nucleotide skews, and their chromosomes have mosaic structures[12, 13]. The *cdc6* gene location can act as a tag for identifying the *Ori* site in archaea. All the archaea for which experimental data is available shows that cdc6 gene lies within close proximity of the *Ori* site. The cdc6 gene codes cdc6 protein; which belongs to AAA+ protein super family that act as origin recognition proteins. It binds with *Ori* site and causes conformational changes in DNA strand. For this process to take place energy from the ATP is required. All the cdc6 proteins are rich in helix turn helix domain because helix provides greater 'bendability' during making a ring over the DNA strand.

The *Ori* region is likely to occur in intergenic region between cdc6 and some other gene. This intergenic region has high 'AT' rich (upto 80%) because of minimum energy required to break two hydrogen bond between A and T compared to three hydrogen bond between G and C. it is also seen that intergenic region between cdc6 and its adjacent gene has a number of direct and inverted repeats. What are the roles of these repeats is still not clear but some scientists suggest that they help in the binding of MCM helicase and to perform its activity. Consensus ORB sequence could be used in addition to these method to increase the accuracy but this has also some bottleneck because generally archaeal genome ORB sequence do not show a common pattern and the sequences change from species to species. Moreover, in a single species there are several copies of ORB sequences . FindingO*ori* sites in archaea is not an easy task. *In silico* methods are necessary because they are fast and cost effective and easy to use for the archaea which are not easily cultivated like the *Korarchaeota*, new group of archaea found from environmental samples.

5. Future Work

Even In this preliminary study we can conclude that the replication machinery of the archaea resembles that of the eukaryotes and is very different from that of the bacteria. There is an open question in evolution as to whether the replication mechanism of archaea and eukaryotes are related or have evolved separately, that is which of them evolved first and is closer

to LUCA (Last universal common ancestor). We plan on exploring this open question in our future work.

References

1. Baker *et al*, Helicase action of Dna B protein during replication from the Escherichia coli chromosomal origin in vitro*, J. Biol. Chem*., **262** (1987) 6877-6885.
2. R. Bernander, The archaeal cell cycle*, Mol. Microbio*.,**48** (2003) 599-604.
3. P. Worning *et al,* Origin of replication in circular prokaryotic chromosomes, *Environmental Microbiology*, **8** (2005) 353-361.
4. Bell *et al,* DNA replication in Eukaryotic cells, *Annu. Rev. Biochem*., **71** (2002) 333-374.
5. Norais *et al*, Genetic and Physical Mapping of DNA Replication Origins in Haloferax volcanii, *Plos Genetics***, 3** (2007) (77).
6. J. R. Lobry, Genomic Landscapes, *Microbiology Today*, **26**, (1999) 164.
7. R. Zhang and C.T. Zhang, The Z-curve database: a graphic representation of genome sequences, *Bioinformatics,* **19** (2003) 593.
8. R. Zhang and C.T. Zhang,Identification of replication origins in the genome of the methanogenic archaeon, Methanocaldococcus jannaschii, *Extremophiles*, **6** (2004).
9. C. Woese *et al*, Towards a natural system of organism: proposal for the domain of archaea, bacteria and eukarya, *Proc. Natl. Acad. Sci., USA***, 87** (1990).
10. L. Chen *et al.,* The Genome of Sulfolobus acidocaldarius, a Model Organism of the Crenarchaeota, *J. of Bacteriology,* **187** (2005) 4992.
11. N. Fujikawa et al**,** Structural basis of replication origin recognition by the DnaA protein, *Nucleic Acids Research*, **31(8)** (2003) 2077-2086.
12. J. Lobry, Asymmetric substitution patterns **in** the two DNA **strand**s of **bacteria,** *Molecular Biology and Evolution*, **13**, 660-665.
13. J. Lobry and N. Sueoka, Asymmetric directional mutation pressures in bacteria, *Genome Biology* **3** (2002).
14. A. C. Frank and J. R. Lobry, Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms, *Gene* **238** (1999) 65–77.
15. A. Grigoriev, Analyzing genomes with cumulative skew diagrams, *Nucleic Acids Res*. **26** (1998) 2286–2290.
16. J. Mrazek and S. Karlin, Strand compositional asymmetry in bacterial and large viral genomes, *Proc. Natl. Acad. Sci. USA* **95** (1998)3720–3725.