Bioinformation **Volume 5** *open access*

www.bioinformation.net Issue 5 **Hypothesis**

Mapping of origin of replication in *Themococcales*

Krishna K Ojha[1], D Swati[1, 2] *

[1]Department of Bioinformatics, MMV, Banaras Hindu University, India; [2]Departments of Physics and Bioinformatics, MMV, Banaras Hindu University, India; D. Swati - Email: swatid@gmail.com; * Corresponding Author

Received June 7, 2010; Accepted September 24, 2010; Published November 1, 2010

Abstract:
Genome replication is a crucial and essential process for the continuity of life. In all organisms it starts at a specific region of the genome known as origin of replication (Ori) site. The number of Ori sites varies in prokaryotes and eukaryotes. Replication starts at a single Ori site in bacteria, but in eukaryotes multiple Ori sites are used for fast copying across all chromosomes. The situation becomes complex in archaea, where some groups have single and others have multiple origins of replication. Themococcales, are a hyperthermophilic order of archaea. They are anaerobes and heterotrophs-peptide fermenters, sulphate reducers, methanogens being some of the examples of metabolic types. In this paper we have applied a combination of multiple in silico approaches - Z curve, the cell division cycle (cdc6) gene location and location of consensus origin recognition box (ORB) sequences for location of origin of replication in *Thermococcus onnurineus*, *Thermococcus gammatolerans* and other Themococcales and compared the results to that of the well-documented case of *Pyrococcus abyssi.* The motivation behind this study is to find the number of Ori sites based on the data available for members of this order. Results from this in silico analysis show that the Themococcales have a single origin of replication.

Keywords: Ori sites, Disparity plot, Z curve, cdc6 gene, Intergenic repeats, ORB.

Background:
The path breaking work of Carl Woese and Fox established the concept of three distinct domains of living beings- Bacteria, Archaea and Eukarya**[1]**. However the debate about the evolutionary relationships between the three domains of life still continues. In terms of DNA replication it seems that the archaea are closer to eukaryotes than bacteria since proteins involved in DNA replication in archaea share more similarities with their counterparts in eukarya**[2]**. DNA replication is an integral part of cell division and a necessary condition for the equal distribution of genetic material from the parental cell lineage to its offspring. It is a highly conserved process with high accuracy and small error rate. To ensure high fidelity, the cell has an array of enzymes and a regulatory mechanism to coordinate the replication process **[3]**. Bacterial replication involves the DnaA protein binding to a stretch of DNA called dnaA boxes, located close to the origin of replication. These nucleotide sequences are also called origin recognition boxes (ORB). There are AT rich region upstream of the replication site named DUE (duplex unwinding element), which facilitate the opening of DNA double helix for the purpose of replication. Eukaryotic replication has to occur at multiple sites simultaneously and requires a complicated machinery called origin recognition complex (ORC). The exact working of this machinery is not yet fully understood. The third domain, archaea has replication proteins which are homologous to the eukaryotic proteins, but are fewer in number. Therefore a thorough study of the replication process of archaea will help in understanding the replication process in eukaryotes. The archaea however do not show a uniform pattern of replication. Some like P.abyssi **[4],** the first experimentally studied archaeon, have a single origin of replication like bacteria, but some like *Sulfolobus solfataricus* **[5]** and *Haloferax volcanii* **[6]** have multiple origins. The archaeal chromosomal Ori initiation site was first identified experimentally in *Pyrococcus abyssi*. This was located within the non coding region downstream of a gene encoding a homolog of the eukaryotic cdc6 gene. This intergenic region is very rich in AT content and has several direct and inverted repeats.

Prediction of origin of replication in archaea has attracted immense interest during past few years. Since the archaea are notoriously difficult to culture in the laboratory, in silico analysis of Ori in archaea will become more important with the availability of an increasing number of complete genome sequences. There are about 91 archaea whose genomes have been sequenced either fully or up to the draft assembly level (June 2010), but there are very few species for which experimental data is available. There is a big gap between the number of sequenced genomes and experimentally mapped Ori sites in archaea. Since in silico methods are cost effective in terms of saving man-power as well as laboratory costs, they can be used advantageously to predict replication origin(s) in archaea.

The first clue that helped identify the putative replication origins came from an observation that there are compositional asymmetries between the leading strand and lagging strand in bacteria. By using a DNA walk method, Lobry noticed that the DNA walk changes polarity at the origin of replication **[7]**. Later, various methods were proposed to measure the asymmetrical distribution of nucleotide composition around the replication origin, of which GC skew and Z curve (a combination of various nucleotides disparity plots) are extensively applied in finding Ori in bacterial and archaeal genome respectively **[8].** GC skew analysis has been widely applied for mapping Ori sites in bacteria, viruses and mitochondria. It is defined as (C−G)/(C+G) where C and G are the number of occurrences of C and G nucleotides in a subsequence covered by a sliding window scanning the DNA sequence. The cumulative graph of GC skew shows a sharp peak at the Ori site. But in archaeal genomes this approach fails because they do not have sharp nucleotide skew due to the mosaic nature of their genome. Z curve could be applied up to certain level of confidence for archaea, but it has also its limitations **[9]**. Hence nucleotide skew and location of its extremum cannot locate the Ori site in archaea. The order-specific location of ORB boxes, as well as the location of the cdc6 gene upstream of the putative origin of replication may reinforce accuracy of

Bioinformation Volume 5 open access

www.bioinformation.net Issue 5 **Hypothesis**

prediction of the origin of replication. This is the strategy applied in this work. Thermococcales are the most explored and studied order among the phylum Euryarchaeota of archaea. They thrive in thermal vents of the sea floor or deep oil wells in very hot environments. The typical optimum growth temperature for Themococcales is about 75 degrees centigrade, and it may vary from 60 to 110 degree centigrade. Being abundant in their environmental niche, they form an important constituent of the marine hydrothermal vents eco-system. Some members of this order e.g. *Pyrococcus abyssi* are well studied and the Ori site has been mapped experimentally and the data obtained is used to make prediction about other members of this order. Recently, the genome of archaeon *Thermococcus onnurineus* has been sequenced **[11]**. The genome size is 1.8 Mb long and forms a single replicon (chromosome) without any plasmid and total GC content is about 51.3%. *T. onnurineus* belongs to the order Themococcales. The complete genome of another Themococcale, *Thermococcus gammatolerance* has also been recently sequenced **[16]**. The replication origin of both the genomes has not been identified yet. The Z curve analysis of *Thermococcus onnurineus* shows one global sharp peak y curve **(Figure 1)**. It has a single cdc6 gene situated adjacent to the global peak on the complementary strand. The upstream intergenic region between cdc6 and a hypothetical protein shows several direct and inverted repeats **(Figure 2)**. Several consensus ORB sequences are also distributed in this intergenic region. For *Thermococcus gammatolerance* also there is a clear peak of the y curve, locating the putative Ori site, with the cdc6 gene being located close to the peak **(Figure 5)**.

Materials and Methodology:

The genome of seven Themococcales{*Pyrococcus horikoshii* (NC_000961), *Pyrococcus abyssi* (NC_000868), *Pyrococcus furiosus* (NC_003413), *Thermococcus kodakaraensis* (NC_006624), *Thermococcus onnurineus* (NC_011529), *Thermococcus gammatolerans* (NC_012804), *Thermococcus sibiricus* (NC_012883) were downloaded from NCBI http://www.ncbi.nlm.nih.gov/. A script/program to calculate and plot the Z curve was written in MATLAB (http://www.mathworks.com). The Z curve was plotted for all the genomes and the individual disparity curves were plotted on x, y, z axes.

The Z curve is a curve in three-dimensional space, constituting a unique representation of a DNA sequence which has all the property of the given sequence. The x curve is the purine-pyrimidine (R-Y) disparity curve, the y curve is the amino-keto (M-K) disparity curve and the z curve shows the disparity of occurrence along the DNA sequence between the strong- weak hydrogen bonded nucleotides (S-W). In this method **[9]** the given DNA sequence is scanned using a sliding window approach, and keeping track of cumulative counts. Suppose in a given subsequence the purine (A/G) count is greater than the pyrimidine (C/T) count, and then the value of x is greater than zero. For the reverse case, it is less than zero, or if it is equal, then the value of x is zero. Similarly if in a subsequence the count of amino (A/C) nucleotides are greater than the keto (T/G) nucleotides, then y for that subsequence is greater than zero. Also if the weakly bonded nucleotides (A/T) are greater than the strongly bonded (C/G) ones, then value of z is greater than zero, if it is less, than value of z is negative and if it is equal then, value of z is zero. The x or R/Y disparity curve or plot as well as the y or M/K disparity curve can be used to locate the origin of replication. These disparity plots are sometimes more useful than the widely used AT or GC disparity plots introduced by Lobry **[7]**.

The cdc6 gene location and copy number were identified with KEGG database (http://www.genome.jp/kegg). Consensus ORB sequences were taken from the data on the extensively studied thermophilic archaeon *P.abyssi* **[4]**.

Results:

The archaea whose Ori sites have been mapped experimentally show abrupt change at the site of origin of replication in either the x (signifying purine-pyrimidine skew) or y curve (signifying amino-keto skew). Two archaea for which Ori sites have been located experimentally (*P. abyssi* and *S. acidocaldarious*) have been analyzed with Z curve **(Figure 1)**. In *P. abyssi* , a sharp peak (in blue line) corresponds to the origin of replication and cdc6 gene is placed adjacent to the Ori site **(Figure 1).** In *S. acidocaldarius* we see three sharp peaks for the y curve (blue line) which corresponds to experimentally defined Ori sites. There are three cdc6 genes in *S. acidocalderius*, two of which occur close to two consecutive peaks. The third cdc6 gene is somewhat distantly situated upstream of the third peak position. The location of the 1st, 2nd and 3rd cdc6 gene is from 101 to 1261, 578164 to 579357 and 724282 to 725529 nt respectively **(Figure 1b).** Wet lab study has proved that *S. acidocaldarius* has three origins of replication **[10]**. Having tested our method on known experimental results, we applied it to locate the Ori for the remaining six Themococcale genomes, other than *P.abyssi*.

Thermococcus onnurineus genome has been recently sequenced **[11]**, and no experimental data is available for the Ori site. It has a single circular chromosome without any plasmid. All replication machinery proteins like, Cdc6, Mini Chromosome Maintenance (MCM) and Proliferating Cell Nuclear Antigen (PCNA) occur as single copy, and the y curve shows a sharp peak at 1500 kb and cdc6 gene is situated adjacent to this peak **(Figure 1)**. Several other genes crucial for DNA replication are also clustered around cdc6 gene shown in **Figure 2**.

The location of the cdc6 gene is from 1508116 to 1509363 nt on the complementary strand. The x curve also shows a peak in this region. **Figure 2** shows the compact view of the genes involved in locating putative Ori sites. In the complementary strand the cdc6 gene is contiguous to the DNA polymerase II small subunit which is followed by the DNA polymerase II large subunit. All these proteins play a crucial role in DNA replication dynamics, so it is supposed that they should lie close to each other. In the upstream region of the cdc6 gene lies a hypothetical protein after which there is an AT rich intergenic region (-1510249 to -1510837 nt) of 589 nt long, having around 64% AT in contrast to the genome-wide AT composition of 49%. There are several direct and inverted repeats with several tracts of A and T in this region. We have also found that some repeats in this region are common to the Ori region of *P. abyssi*, which has been experimentally mapped. We have established two long ORBs similar to *P.abyssi* shown below **(Figure 3)**.

TTCCAGCGGAAACTGAAGGGGTCAAAGN
TTCCAGTGGAAATGAAACTCTGGGGGGT

Further analysis of Ori sites of other members of order Themococcale shows close similarities with *P. abysii*. In all members the cdc6 gene has a single copy. The length varies between 1248 to 1348 nt and in most of the cases it is below 1300 nt (except *T. gammatolerans* and *P. horikoshii*). The location of the cdc6 gene varies from species to species. In most of the members it is situated on template strand while in some members (*Pyrococcus horikoshii*, *Thermococcus onnurineus*, *Thermococcus gammatolerans* & *Thermococcus sibiricus*) it is situated on the complementary strand. The y curve for all the members are smooth and show a single sharp peak. The cdc6 gene location around these peaks seems to be conserved in most of the members, that is, it lies in close proximity to the peak of the y curve. The results for *T.sibiricus* in this study replicate the findings in **[12]**. The region upstream of the cdc6 gene is At-rich. Presence of consensus ORB sequences varies from species to species and we have mapped the ORB in *T. onnurineus*, and *T. sibiricus* **(Figure 3, 4)**.

In most of the cases copy number of the cdc6 gene is directly linked with the number of replication sites in genome, that is if there are multiple copies of the cdc6 gene, it indicates a possibility of multiple Ori sites in that genome. All Themococcales sequenced till now have the single copy of the cdc6 gene. All Themococcales for which experimentally mapped Ori region is known show that they have a single origin of replication. In **(Table 1 see supplementary material)** we have given the details of cdc6 gene in all Themococcalesas mapped in this study.

Bioinformation Volume 5 *open access*

www.bioinformation.net Issue 5 **Hypothesis**

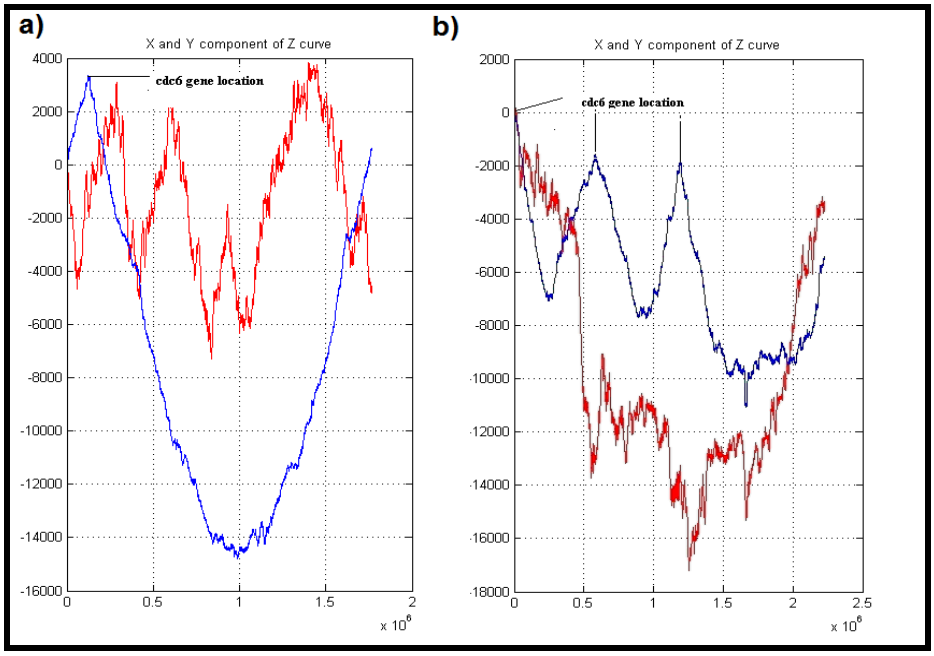

Figure 1: The **x** and **y** curves of *P. abysii* and *S. acidocalderius* respectively .The **x** curve is shown in red line while **y** curve is in blue. In Figure 1(a), **y** curve shows a sharp peak. The sharp peak corresponds to origin of replication. The site next to it is occupied by the *cdc6* gene. Similarly three peaks are seen (in Figure 1(b)) for *S. acidocaldarius* which are the three Ori sites. These peaks are also close to the *cdc6* gene position.

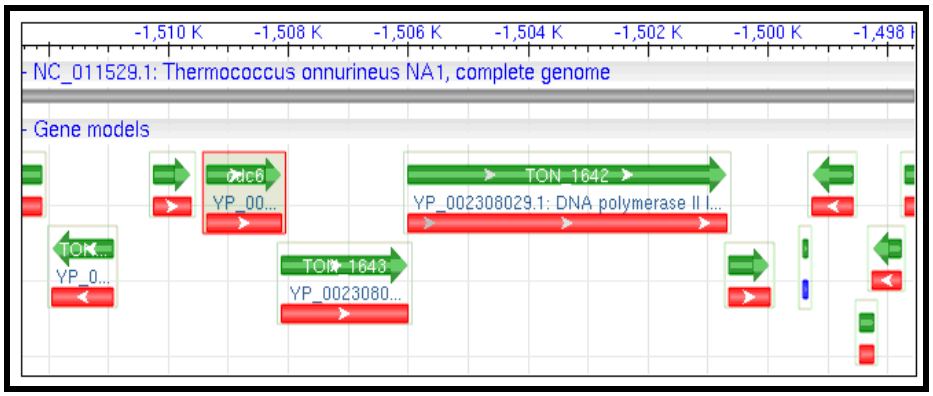

Figure 2: Graphical representation of *cdc6* gene and cluster of genes responsible for replication in *T. onnurineus*. *cdc6* gene and DNA polymerase II subunits are adjacent to each other.

TGAAATCACCAAGGGAACTTTCGATTCGAAACTTAATAACGGTATCG**TTCCAGCGG**
AAACTGAAGGGGTCAAAGAACATAATAGTGCATCAAACATTTACATTCGAAAATG
AACAATAGCATCCAGATGGGCATTTCCACTGAAAGAAGAATGAACATTTTTGGACA
GAAAATCATGTACACTAATGCTCCAAATCGTACTATTTTGGTAAACATATGTACTAA
ACACAAGACAGAACATTAAACGGTTTCGTAGTCCCCCAGAGTTCTGTTTCCGGTGGA
GGCGGGCTTTGGCTACTCCCATATTGTCCTTTGGTTTTTGGTGAAACTCTCCCGCTTT
CTTTACAATTCCCAATATTTATATTTTTTCATGAACAATGTTCATAAAAAATCTCAAA
CTCATAAAGATCATGAAAATTAGTCTAAATATTCAAAGAGTAGATTTAAAATCTGTT
GATTTTATGTTTGTAATGTCCTTTATAAAATGGAGCTTTCTGTTTTTTCTGGCATGTA
ATCCATCCATTCTCCGAAATTGGG**TTCCAGTGGAAATGAAACTCTGGGGGGTCTT**
TCAAATAATCATCGAAAAAGACCTTTGTTTCCAGTGGAAATCTAACTCCATTACTGG
ACACCACAAAGCATTAATGATACTTTTTTAAACCTGCTAACGATAGTGTACAAC

Figure 3: Upstream intergenic region of *cdc6* gene in *T. onnurineus* from -1510249 to -1510837 nt**.** Two ORBs homologous to *P. abyssi* ORB are underlined with bold capital letters. We can see that in this region there are several small tracts of A and T.

Bioinformation **Volume 5** *open access*

www.bioinformation.net Issue 5 **Hypothesis**

TAGCAATTGGAAATATCTAGTGCAAAGTTAAAAAGATTAACTTTTCCACAGGAAATA
CTGCCCCAAATTGTTGAACTAAAAGTGCCACCATATACCAATAATACAGTGATGTCC
ATAT**GTTCCAGTGGAAATAAAGGTCTTTTTC**GATACCTATATTAACGTTAAATTTT
GATCTTATAACC***CCCTCAGAGTTGTATTTCCACTGGAAG***CTAGTTAGTGAAATGGACG
GGATTTGGAATTTGATATCTAAATTTTACGTTAATTTATCATATAATTTGAACAATGT
TCATAGATAAATATATATACCCTAACGACAAAAAATATTGGGAGTGTTTATACCCGT
TTTAATCATACATACGGGTGTTCCGCAGAG**GTTCCAGTGGAAATGAAACTCTGGG**
GGGTTATAGTTCTGTTTACCAATCATCCAAAAATTAAATATTTTTAGTTAACAACAT
GAACAAAAATGTACTTAAATGATCGCTTTTTAATGAACATTTTTATACATTTTTCTTC
CAAACGAAATCCTCAAGTGAGCTCATATTAATCTCCATATGTCAAAACTTTGAGAGT
TTCTCGAAGTGTTCATCAACAAGTTTCCATTGGAACCATGGTTTTAAATACTTTCAGT
GCCTAGTTTCAT

Figure 4: *T.sibiricus* intergenic region (from 1783450 to1784178 nt) and consensus ORB**.** Three ORB are highlighted in bold and underlined, 2nd ORB is situated on complementary sequence and shown in italics.

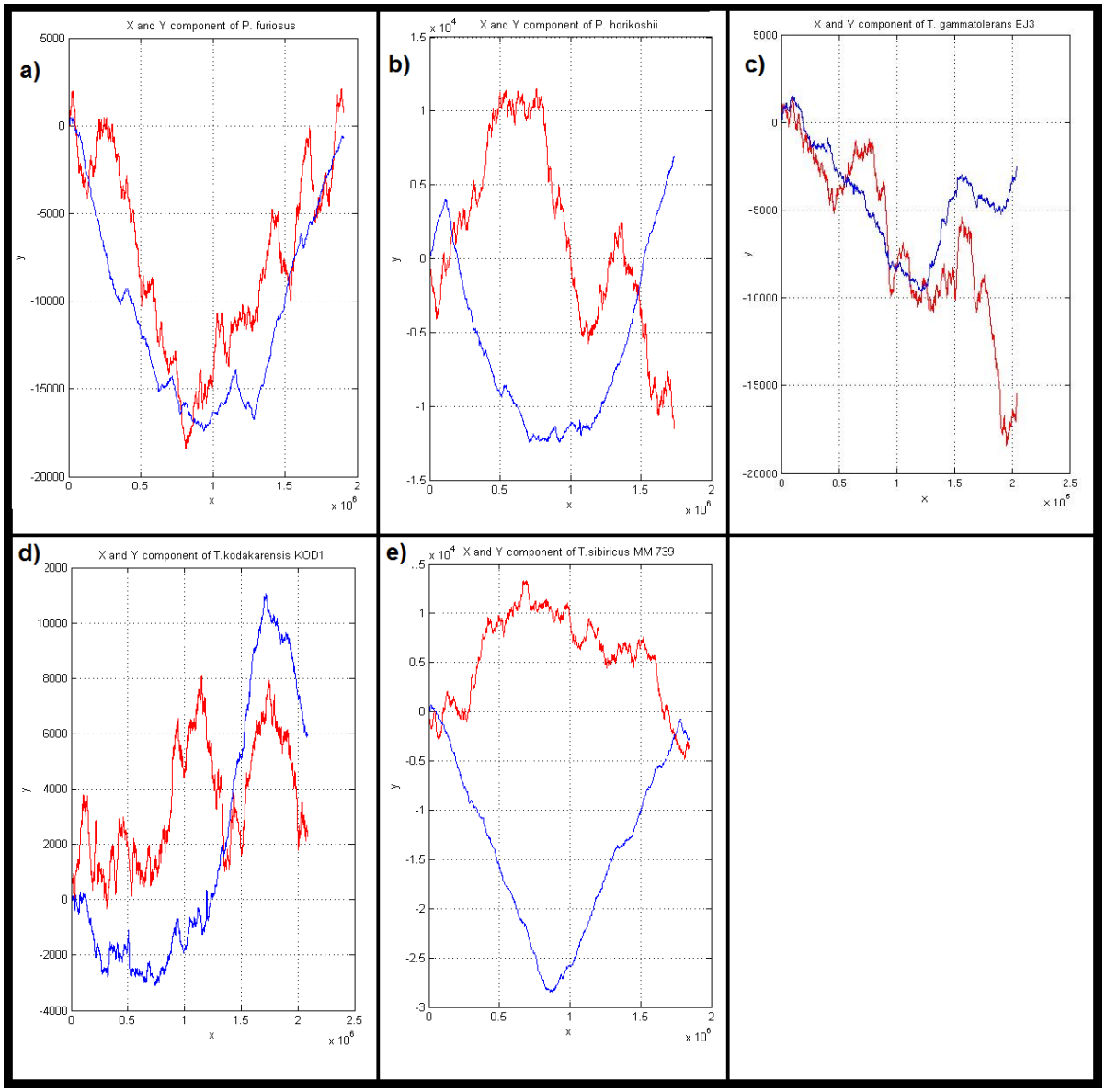

Figure 5: x and y curve of some Thermococcales.
Color code: Red line shows x curve and Blue line shows y curve. X axis shows genome length and y axis shows disparity.

Discussion:

Since their identification as a different domain of life by Woese and Fox **[1]**, the archaea have been the subject of intensive investigation. Several genomic sequences from a number of archaeal species and strains have meanwhile been published. Due to their distinct nature and habitat archaea have always intrigued scientists, though now they are no longer found to be extremophiles alone but also mesophiles and psychrophiles **[13]**. The most interesting component of the archaeal cell is its replication machinery, which shows more similarity to eukaryotic machinery than to the bacterial one. Almost all proteins involved in archaeal genome replication have their homologues in eukaryotes, so it is expected that understanding genome replication mechanism in archaea will give an insight into the functioning of the genome replication dynamics in eukaryotes as well **[14]**. However the eukaryotes have multiple Ori, but as noted earlier, the number of Ori sites in archaea varies. In some archaea like the Thermococcales there is single Ori, but in others there are multiple. Experimental determination and validation of Ori site(s) in archaea is a daunting task and there are several biological and economical constraints as well, so an in silico approach to map Ori region is not only an alternate way, but appears the only way to accelerate the mapping of Ori site up to certain level of confidence **[15]**. There are several tools available for finding putative Ori site. The widely known and applied approach is GC skew for bacteria and Z curve for archaeal genomes. GC skew works well for finding Ori in bacterial genomes, because they have pronounced nucleotide skew and show nucleotide disparity for leading and lagging strand at the site of replication origin. But this approach fails in archaea because their DNA sequences do not show a pronounced nucleotide skew in general. The combined use of nucleotide disparity curves, location of the cdc6 gene, and identification of origin recognition boxes (ORB), maps the single origin of replication in *P.abyssi, T. gammatolerans* and *T.sibericus*, as well as all the other Themococcales studied in this work. The size of the cdc6 gene is approximately 1300 nt, and the ORB of P.abyssi are found in *T. sibericus* as well as *T. onnurineus*, though with some minor mutational changes. *Thermococcus gammatolerans* is one of the most radio-resistant organisms known amongst the archaea and is of great interest to microbiologists to compare with the well known radio-resistant bacteria *Deinococcus rodiodurans*, to find the similarities and differences in their genome and proteome **[16]**.

The exact number of cdc6 does not always tally with the number of Ori, as is seen in Halobacterium sp., where the copy number of cdc6 is 13, but the number of origins found by our method is 3(data not shown). This implies that cdc6 probably performs multiple functions during archaeal replication.

The reason behind some Archaea, like, Halobacterium sp, and *H. volcanii*, *Sulfolobus solfataricus* having multiple Ori, but *P.abyssi* and *M.jannaschii* having single Ori is not clear at present. The archaea are divided into four main phyla, Euryarchaeota (e.g. *P.abyssi, M.jannaschii, Halobacterium sp*.), Crenarchaeota (e.g. *Sulfolobus solfataricus, Aeropyrum pernix*), Korarchaeota and Nanoarchaeota (with one member each at present). While no overall trend has been seen so far, this study of Ori mapping in silico of the order Themococcale is reassuringly simple and uniform.The sequence of ORB also differs from case to case but Norais *et al.***[6]** have given a consensus ORB for Euryarchaeota, which is 79% similar to the consensus ORB found here for Themococcales.

Conclusion:

The methodology used in this study, though adequate, has to be improved by looking for co-localization of the other proteins involved in the replication process, like MCM, PCNA, RFC and DNA Polymerase II, and a refinement of finding order-specific ORBs over phylum-specific ORBs. The future plan of work involves a systematization of in silico mapping of Ori sites in all sequenced archaeal genomes.

The origin of multiple Ori sites in archaea is not clear. The Sulfolobales, an example of which is *S.acidocaldarius*, have three Ori sites which appears helpful for fast replication of these extremophiles whose optimum growth temperature (OGT) is about 80 degrees centigrade. But the Themococcales, which have OGT in the same range or higher, clearly have a single origin of replication as demonstrated in this work. On the other hand, Halophiles which have lower growth temperatures, like *H.volcanii* [6] and Halobacterium sp. have multiple Ori sites. Does this imply that the replication machinery evolved repeatedly and randomly for each order of each phylum of archaea? Is Nature really that chaotic and profligate? This aspect of replication process of archaea demands concerted and co-operative work by laboratory scientists as well as Computational Biologists to lead to a definite answer to this question.

Acknowledgement:

The authors thank Rileen Sinha for carefully going over the manuscript and giving useful comments.

References:

[1] CR Woese & G Fox. *Proc. Natl. Acad. Sci. USA* 74: 5088 (1977) [PMID: 270774]
[2] B Grabowski & Kelman. *Annu. Rev. Microbiol.* **57**:487 (2003).
[3] SP Bell & B Stillman. *Nature* **357** :128 (1992)
[4] H Myllykallio *et al. Science* **23**:2205 (2000) [PMID: 10864870]
[5] I Dionne *et al. Biochem. Soc. Trans.* **31**:647 (2003) [PMID: 12773180]
[6] C Norais *et al PLoS Genetics*. **18**:729 (2007) [PMID: 17511521]
[7] Jean R. Lobry. *Microbiology Today* **26**:164 (1999)
[8] NV Sernova *et al. Briefings in Bioinformatics* **9**:376 (2008) [PMID: 18660512]
[9] R Zhang & CT Zhang. *Biochem. Biophys. Res. Commun.* **20**: 396 (2002) [PMID: 12237132]
[10] L Chen *et al. Journal of Bacteriology*. **187**: 4992 (2005) [PMID: 15995215]
[11] HS Lee *et al. J. Bacteriol.* **190** : 7491 (2008) [PMID: 18790866]
[12] AV Mardanov *et al. Applied and Environ. Microbiology* **75**: 4580 (2009)
[13] Patrick Forterre. *Theoretical Population Biology*. **61**: 409.
[14] DR Edgell & WF Doolittle. *Cell*. **89**: 995 (1997)
[15] K Bohlke *et al. Extremophile* **6** :1 (2002) [PMID: 11878556]
[16] Y Zivanovic *et al. Genome Biology* **10**: R70 (2009) [PMID: 19558674]

Edited by P Kangueane
Citation: Ojha & Swati. Bioinformation 5(5): 213- 218 (2010)

Bioinformation **Volume 5** *open access*

www.bioinformation.net Issue 5 **Hypothesis**

Supplementary material:

Table 1: Distribution of the *cdc6* gene in *Thermococcales*

S.No	Organism	*cdc6* gene copy No.	*cdc6* gene location	Length (nt)
1.	*Pyrococcus horikoshii*	single	109476..110789	1314
2.	*Pyrococcus abyssi*	single	121402..122700	1299
3.	*Pyrococcus furiosus*	single	16236..17498	1263
4.	*Thermococcus kodakaraensis*	single	1712158..1713405	1248
5.	*Thermococcus onnurineus*	single	1508116..1509363*	1248
6.	*Thermococcus gammatolerans*	single	125431..126738*	1308
7.	*Thermococcus sibiricus*	single	1434100...1435362*	1262

* Situated on complementary strand