Journal of
Proteomics & Bioinformatics

Anupama et al., J Proteomics Bioinform 2014, S9
DOI: 10.4172/jpb.S9-005

Research Article Open Access

Evolutionary Analysis of CRISPRs in Archaea: An Evidence for Horizontal Gene Transfer

Anupama S, Aswathy Rajan MP, Gurusaran M, Radha P, Dinesh Kumar KS, Chitra R, Hima Vyshanavi AM and Sekar K*

Laboratory for Structural Biology and Bio-computing, Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore–560012, India

Abstract

Akin to the eukaryotic immune system, prokaryotes harbor CRISPRs, a lineup of DNA direct repeats and spacers to foster immunity against the invading phages and plasmids. A CRISPR (Clustered Regularly Interspaced Short Palindrome Repeats) unit found within the genome of an organism consists of an array of repeating sequences interspaced by unique spacers and is associated with special genes that reside adjacent to the array. The spacer sequences are nucleotide fragments integrated from an invading organism. The comparative genomic analysis of CRISPR sequences affirms huge variation within the CRISPR–CAS systems among different prokaryotes. Here, an analysis of the complete archaeal species is directed for their CRISPR sequences along with a case study. The phylogenetic analysis sketched from the CRISPR sequences signifies a harmony along the direct repeats of the analyzed organisms with no trace of spacer similarity. Further, novel CRISPR elements are procured, aside from those formerly present in the database. The CRISPRs are then subjected for local alignment using BLAST to ensure whether any of the sequences showed similarity with the human and viral genomes available in NCBI.

Keywords: CRISPR; Direct repeats; Spacers; CAS genes; CRISPRFinder; Horizontal gene transfer

Introduction

CRISPRs are a kind of defense mechanism featured by the prokaryotic species, both archaeal and bacterial domain. About 90% of archaea and 40% of bacteria holds CRISPR loci in their genome [1]. A CRISPR unit is made up of repeating sequences known as the direct repeats, which are separated by spacer sequences and is preceded by a 500-550 bps leader sequence [2,3]. The direct repeats of a CRISPR are partially palindromic and lie within the range of 24-48 bps [4,5]. These repeats tend to form a dyad symmetry that results in the formation of hairpin structures [6-8] including the spacers. Similarity searches on spacers authenticate that they are the captured segments of the genome sequences of the invaders which are derived either from their sense or anti-sense strand [9]. A CRISPR imparts immunity against the invading organisms by the mechanism catalyzed by the products of the CRISPR associated (CAS) genes.

A CRISPR unit is activated when a foreign genome gets conjugated [10-12]. The genome of the invading strain gets disintegrated through the CRISPR mechanism with the help of the CAS proteins, thereby the broken foreign nucleotide fragments gets integrated into the host as a spacer at the leader end of the CRISPR unit [13-15]. When the prokaryote encounters the same predator once again after the first attack, the CRISPR unit in the host genome is transcribed into a pre-CRISPR RNA (pre-crRNA) molecule [16,17]. These pre-crRNA molecules are then processed into CRISPR RNA (crRNA) with the aid of the proteins encoded by the CAS genes [18,19]. The resulting crRNAs contain a spacer flanked by the fragments of repeats on either side. The crRNAs scan the invading genome for the fragment that matches with the bound spacer sequence. The fragment in the invading genome that matches with any of the spacer in the host crRNAs is called a protospacer [20]. The crRNAs along with the CAS proteins (CRISPR-CAS complex) gets bound to the protospacer in the foreign genome sequence by a complementary sequence pairing method. The complex shears the invading nucleotide genome into small fragments which then gets inserted into the host CRISPR unit as a spacer [21-23].

The whole CRISPR mechanism is catalyzed by the CAS genes that are found in the vicinity of the CRISPR arrays [13,24]. These genes encode the enzymes involved in the processing of the CRISPR transcripts. The genes also aid in the recognition and neutralization of foreign genetic elements with the inclusion of new spacers. CAS genes can be classified into different categories depending on their role of action. There are altogether six types of core CAS genes associated with the CRISPR mechanism, out of which Cas5 and Cas6 are newly added [25]. Excluding the newly added genes, the four core genes are aligned as Cas3-Cas4-Cas1-Cas2. The Cas2 is a sequence-specific endoribonuclease [26], Cas3 acts as a helicase [27], Cas4 resembles the RecB family of exonucleases and contain a cysteine rich motif and Cas1 found in all the organisms harboring a CRISPR unit is highly basic. Apart from the core genes, there are a few subtype genes that belong to the RAMP (Repair Associated Mysterious Proteins) family of proteins [28].

Each of the genomes analyzed for this study hold varied CRISPR units, based on their length, repeats and spacers. CRISPRs appear to share similar direct repeats within the studied archaeal strains [17]. The similarity in the repeats of a CRISPR unit indicates a possible horizontal gene transfer between the strains [6,29]. This horizontal gene transfer may be mediated by plasmids, mega plasmids and even prophages which carry the CRISPR units [8,30]. Similarity in spacers seems to have originated when two different organisms encounter invasion by the same phage or plasmid. CRISPR arrays within the chromosomal and plasmid genome of the same strain having similar spacers protects the genome from degradation due to 5' overlap of the repeat [15]. The

***Corresponding author:** Sekar K, Laboratory for Structural Biology and Bio-computing, Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore–560012, India, Tel: 080-22933059/22933060/23600551; Fax: 080-23600683; E-mail: sekar@physics.iisc.ernet.in/sekar@serc.iisc.ernet.in

Received June 17, 2014; **Accepted** August 12, 2014; **Published** August 15, 2014

Citation: Anupama S, Aswathy Rajan MP, Gurusaran M, Radha P, Dinesh Kumar KS, et al. (2014) Evolutionary Analysis of CRISPRs in Archaea: An Evidence for Horizontal Gene Transfer. J Proteomics Bioinform S9: 005. doi:10.4172/jpb.S9-005

analysis of CAS genes shows that the common CAS genes found in the vicinity of all the CRISPR units are Cas1 and Cas6. The present study focuses on the computational and phylogenetic analysis of the CRISPR units present in all the 110 species of archaea. The main objectives of this work are the analysis of the direct repeats and the CAS proteins associated with the CRISPRs for substantiating their diversity.

Materials and Methods

Even though, there are efficient software packages to extract the repeats from genome and protein sequences [31-34], one has to employ a dedicated software package to extract CRISPRs. Hence, we have used Online CRISPRFinder [35] (URL-http://crispr.u-psud.fr/Server/), a program to enumerate all the mandatory details of CRISPRs. Among the tools available for retrieving the CRISPR units, CRISPRFinder is found to have better efficiency in CRISPR investigation.

Retrieving the CRISPR units from the available archaeal strains

The archaeal domain accommodates a total of 110 species. The 110 archaeal species and their strains were retrieved from the taxonomic databases and their whole genome sequences were obtained from the NCBI/GenBank database. Search for all accurate CRISPR units in the archaeal genomes was accomplished by a web interface (with default parameters) that offers elementary crossing points for the CRISPR identification with precision, allowing a factual definition of the direct repeat consensus boundaries and related spacers. This program was developed in Perl under Debian Linux [35] and was implemented to obtain the CRISPRs along with the flanking sequences. The CRISPRFinder output displays CRISPRs with the repeats, the intervening spacers along with their accurate positions in the genome and the referenced genes found within the sequence.

CRISPRFinder employs a stringent filter to cull out the confirmed CRISPRs. Confirmed CRISPRs are the ones that have at least three motifs and two exact identical direct repeats (DRs), while the remaining candidates are tagged as questionable CRISPRs. For our analyses, we have strictly considered confirmed CRISPRs and have validated it against CRISPRdb [36], a database that catalogues the confirmed CRISPRs. This database serves as a reliable source of complete CRISPR information.

CRISPR analysis

The finalized direct repeats representing 191 diverse CRISPR clusters from different organisms were analyzed for inferring the total percentage they make up in the entire genome of the organism and to find the specific GTTTG/C and GAAAC motifs in the direct repeat sequences. These motifs were responsible for the palindrome nature of the direct repeat sequences within the CRISPR array. Then the chosen direct repeats were aligned with the help of ClustalW [37] (URL - http://www.ebi.ac.uk/Tools/msa/clustalw2/), a multiple sequence alignment program. The alignment results were used as an input for constructing the phylogenetic tree using MEGA (Molecular Evolutionary Genetic Analysis), an offline toolkit for conducting alignments and drawing the relationship trees with an accurate branch distances [38].

CAS protein retrieval and phylogenetic analysis

CAS genes form the fundamental part of the CRISPR machinery that encodes the necessary DNA manipulating enzymes needed for the accomplishment of the defense mechanism. The references to the genes related to each CRISPRs in the CRISPRdb is pointed to the GenBank and the amino acid sequences of the proteins encoded by these genes are retrieved from NCBI. Phylogenetic investigation was carried out by aligning the protein sequences from different species in ClustalW. The phylograms constructed using the tool MEGA showed a clear-cut picture of the close relationships within the species, corroborating the fact that they may be the products of Horizontal Gene Transfer.

Results and Discussion

The structural attributes of CRISPRs are deduced to vary with a presumable rate within and between the species. From the examination of the CRISPR units and their organization within the genome, a comparative substantiation is made on the diversity of the CRISPR units within the archaeal domain. All the retrieved chromosomes and plasmid genomes were scanned using the CRISPRFinder for the presence of the CRISPR loci in them.

The output retrieved from the CRISPRFinder

A genome FASTA file was uploaded into the CRISPRFinder that gave an output of the CRISPR sequences with the direct repeats, spacers, CAS genes and leader sequences. The resulting output from the CRISPRFinder was cross-checked with the data available in the CRISPRdb to filter only the confirmed CRISPR sequences. The screening process revealed that some of the CRISPRs that were presented as questionable sequences in the CRISPRFinder were treated as confirmed groups. These CRISPRs are also combined along with the other confirmed sequences that are common between the tool and the repository, adding to a grand total of 391 CRISPRs. The number of repeat elements per CRISPR unit varies with the species. Among the analyzed archaeal strains *Metallosphaera cuprina Ar-4* (1,840,348 bps) in Crenarchaeota, harbors the longest CRISPR of 12,176 bps with 25 direct repeats and 189 spacers, followed by 11,632 bps CRISPR in *Methanococcus voltae A3* (1,936,387 bps) in Euryarchaeota with 31 direct repeats and 171 spacers. *Methanotorris igneus Kol5* (1,854,197 bps) has the shortest CRISPR of 85 bps with 31 direct repeats spaced by a single spacer (Supplementary Table 1). In general, all the CRISPRs comprise direct repeats in the range of 2-100 integrated with 1-99 spacer(s). The number of direct repeats is higher in the methanogens with a segment size lying between the ranges of 24 to 46 bps.

A total of 44 plasmid sequences are also retrieved along with the chromosomal genomes of the 110 archaeal species and only ten plasmid genomes showed the presence of CRISPR units in them. Out of the ten plasmids, one belongs to the methanogens and the remaining comes under the halophilic archaeal group. These plasmid genomes display a total of 15 CRISPR units, thereby giving a net count of 391 CRISPR units including those present in the chromosomal genomes.

Novel CRISPRs

A total of 33 novel CRISPRs are seen in many of the species along with the other CRISPR clusters that are commonly displayed in both the CRISPRFinder output and the database. These CRISPRs satisfied the criteria for the sequences to form CRISPR-like units and are displayed under the category of confirmed CRISPRs in the CRISPRFinder output (Tables 1a and 1b). The majority of the novel CRISPRs is discovered in the methanogens. Some of the CRISPRs are not included in the database but yet seemed to be satisfying the criteria to become a CRISPR and thus, they are also labeled as confirmed ones for the further investigation. To examine the conservation of the novel CRISPRs, we aligned them using ClustalW and constructed a circular phylogram using the Interactive Tree of Life [39]. From Figure 1, it is evident that even though the novel CRISPRs are conversed along the

Species	Genome Size	Crispr Position		Direct Repeats	Dr Length	No of Spacers	Crispr Length	% In Genome
		Start	End					
Methanosarcina barkeri str. Fusaro	4837408	1967688	1967935	GTCAAGCCTTTTTTGAAAAGGGTTG	25	3	247	0.01
		3090113	3090480	CTACCACCAGAAATTGGGAAACT	23	5	367	0.01
Methanosarcina acetivorans str. C2A	5751492	1779366	1779592	ATAATATTGCAAATTCAAACAGG	23	3	226	0.00
Methanosaeta thermophila PT	1879471	677280	677982	GTCGAAGAGCGAGTTCCAGGAAAACAAGGATTGAAAC	37	9	702	0.04
Methanosaeta harundinacea 6Ac	2559043	1830362	1833302	CTATCCATGGCTGAAAAGTCGTGGCCCCATTGAAAC	36	39	2940	0.11
Methanoregula boonei 6A8	2542943	408294	408879	GACATCATCATCATGCAGTCGCACATGGACTTCATCATGG	40	6	585	0.02
Methanosphaerula palustris E1-9c	2922917	1642478	1642689	CGGTTCATCCCCACGCTTGTGGGGAACTC	29	3	211	0.01
		1642843	1649035			101	6192	0.21
Methanoplanus petrolearius DSM 11571	2843290	1554943	1555291	TATCATCCGCGTTCACCAACTGTTCGGC	28	4	348	0.01
Methanoculleus marisnigri JR1	2478101	1554414	1554604	TCCGGGGTTTCCCGGGGCGTCTCCTC	26	3	190	0.01
Methanococcus aeolicus Nankai-3	1569500	86886	87164	ATAACCATAAATGGAAATGCAGGA	24	4	278	0.02
Methanobrevibacter ruminantium M1	2937203	592392	592735	TGTTGTTTGTGAATGTGTTGTTTATTATCTCTCC	34	4	343	0.01
		592857	593121			3	264	0.01
Methanobacterium sp. SWAN-1	2546541	785808	786234	AATTCAAATAACAACACAATATCCGGAAACA	31	6	426	0.01
Caldivirga maquilingensis IC-167	2077567	232508	232870	CTTTCTAATCCCTTTTGGGATTTTC	25	5	362	0.02
		1583088	1583382			25	294	0.01
		288199	288427	AACTTTCTAATCCCTTTTGGGATTTTC	27	3	228	0.01
		1258730	1259356	GAAAATCCCAAAAGGGATTAGAAAG	25	9	626	0.03
		1588509	1589276			11	767	0.04
		1611612	1612046			6	434	0.02
		1700150	1700375	CTTTCTAATCCCTCTTGGGATTTTCT	26	3	225	0.01
Desulfurococcus fermentans DSM 16532	1384116	973322	974943	CTTTCAATTCTTTCTATTGTATTC	24	24	1621	0.12
Pyrobaculum aerophilum str. IM2	2222430	268866	269081	GACGAAACAAATCAAAGAATTGAC	24	3	215	0.01
Sulfolobus solfataricus 98/2	2668974	2054517	2061910	GATTAATCCCAAAAGGAATTGAAAG	25	116	7393	0.28
		2076433	2080895	CTTTCAATTCCTTTTGGGATTAATC	25	70	4462	0.17
		2092271	2100245	GATAATCTCTTATAGAATTGAAAG	24	126	7974	0.3
		2499490	2499902	GATAATCTACTATAGAATTGAAAG	24	6	412	0.02
Sulfolobus islandicus HVE10/4	2655201	2266259	2266502	TCTAGTCTTTCAATATCTTGCTTAGTAGCCA	31	3	243	0.01
Archaeoglobus veneficus SNP6	1901943	665147	669770	GTTGAAATCAGACTAATGTAGGATTGAAAG	30	67	4623	0.24
Haloarcula marismortui ATCC 43049 chromosome I	3131724	3050405	3050562	GGCGGTCCCTGTTCGCTCTGGTT	23	3	157	0.00
Halogeometricum borinquense DSM 11551 chromosome	2820544	147680	148091	TCTGTCTCGTTCGACGACTCTGTCTCAGTGG	31	6	411	0.01
Halogeometricum borinquense DSM 11551 plasmid pHBOR01	362194	147158	147398	CTAACAGACGAAATGAGGGGTGTG	24	4	240	0.06
Haloquadratum walsbyi C23	3148033	403466	406637	GTTGCAACGAAGAGAAAACCCGCTAAGGGATTGAAAC	37	43	3171	0.10
		1391076	1392481	GTTTCAGATGAACCCTTGTTGGGTTGAAGT	30	21	1405	0.04
		1403321	1404737	GTTTCAGATGAACCCTTGATGGGTTGAAGT	30	21	1416	0.04
Natrialba magadii ATCC 43099 (chromosome)	3751858	1454137	1454296	GAGGTGCTGTAGTTGAGGGTGCTGTG	26	3	159	0.00
		1647603	1649414	GTTCCAGAACTACCCTTGTGGGATTGAAGC	30	27	1811	0.05
Natronomonas pharaonis DSM 2160 plasmid PL131	130989	97468	97717	GCACCCCTCTATCGATGTGTACT	23	3	249	0.19

Table 1a: Novel CRISPR sequences found within the strains. **a:** A total of 33 new CRISPRs are identified and listed below.

Sl. No	Phylum	Percentage of novel CRISPRs (%)
1	**Euryarchaeota**	
	Methanogens	36.4
	Halobacterium	15.1
	Archeoglobus	3
2	**Crenarchaeota**	
	Thermoprotei	33.3
3	**Archaeal Plasmids**	12.1

Table 1b: Distribution of novel CRISPRs across different archaeal phyla.

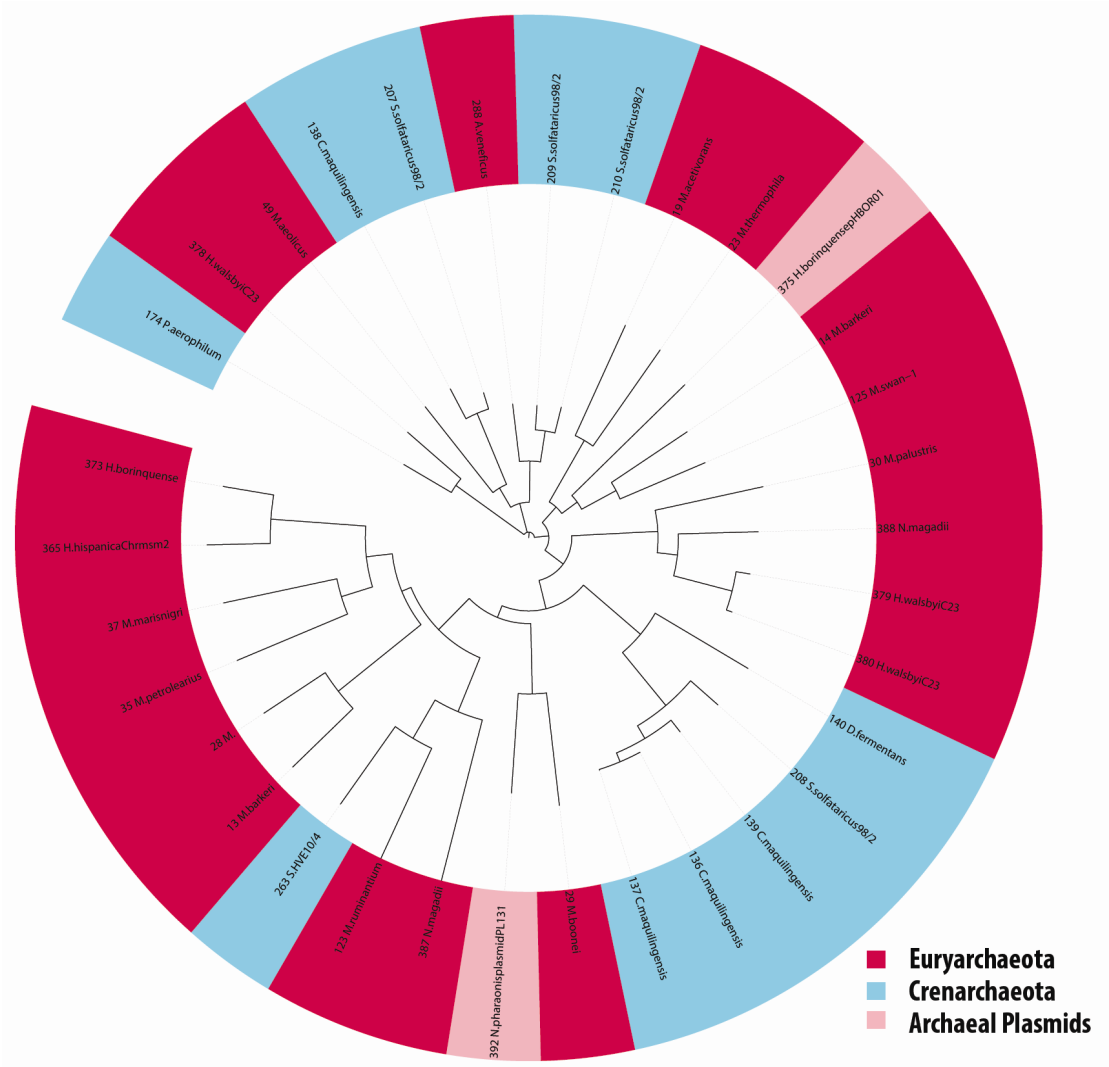

Figure 1: Phylogenetic tree for novel CRISPRs across 110 archaeal species. The tree has been colour coded to highlight different archaeal phyla.

nodes, it is segmented across the phylogram. Different archaeal phyla are highlighted in the phylogram (Figure 1) to observe its distribution associated with the novel CRISPRs.

Palindromicity in the sequences

Repeat sequences in different CRISPR loci are not completely conserved, although the existence of certain partially conserved sequences such as GTTTG/C motif at the 5' end and the GAAAC motif at the 3' end of the direct repeat have been detected which imparts a partial palindromic character to the direct repeat unit. Some of the direct repeats of *Thermococcus sp. CL1* display palindromicity in their sequences which are shown below:

- ***GTTTC***CAAACATTATGTGGTTCT***GAAAC***
- ***GTTTC***AGAACCACATGATGTTTG***GAAAC***
- ***GTTTC***AGAACCACATAATGTTTG***GAAAC***

The advantage of having these special motifs in their genomes is that

they bestow the specific RNA secondary structures to these sequences which might enlighten a better way of understanding the RNA-based CRISPR defense mechanism.

Phylogenetic approach

The broad distribution of the similar CRISPR/CAS system among various organisms irrespective of their origin can be considered as the result of a horizontal gene transfer they undergo during the microbial evolution. To prove this, evolutionary analysis of the direct repeats is carried out using ClustalW.

Similarity between the repeats

To acquire the evolutionary relationships and sequence similarities among the direct repeat sequences in the 391 CRISPR loci, the repeat units are aligned using ClustalW. The alignment scores are examined to retain only those sequences that displayed high sequence similarity with each other. The homogeneity in the analyzed sequences is drawn in the form of phylogenetic trees or phylograms with the aid of MEGA tool. Alignment of the 391 direct repeat sequences is carried out in two different steps. An alignment of the total direct repeat units is made and the sequences showing score more than the maximum score (95) are selected for analyzing the Horizontal Gene Transfer possibilities in them (Table 2). Figure 2 represents the phylogram for all 391 CRISPRs across 110 species of archaea, with the novel CRISPRs tagged by a black dot. From the phylogenetic tree of the direct repeats (Figure 2), it is evident, that the CRISPRs are well conserved within the archaeal phyla. This corroborates the fact that the CRISPRs can be used to infer the evolutionary relationship for 110 archaeal species. The tree also serves an evidence for horizontal gene transfer across archaeal phyla as exemplified by Figure 3. For more detailed information, the tree (Figure 2) is colour

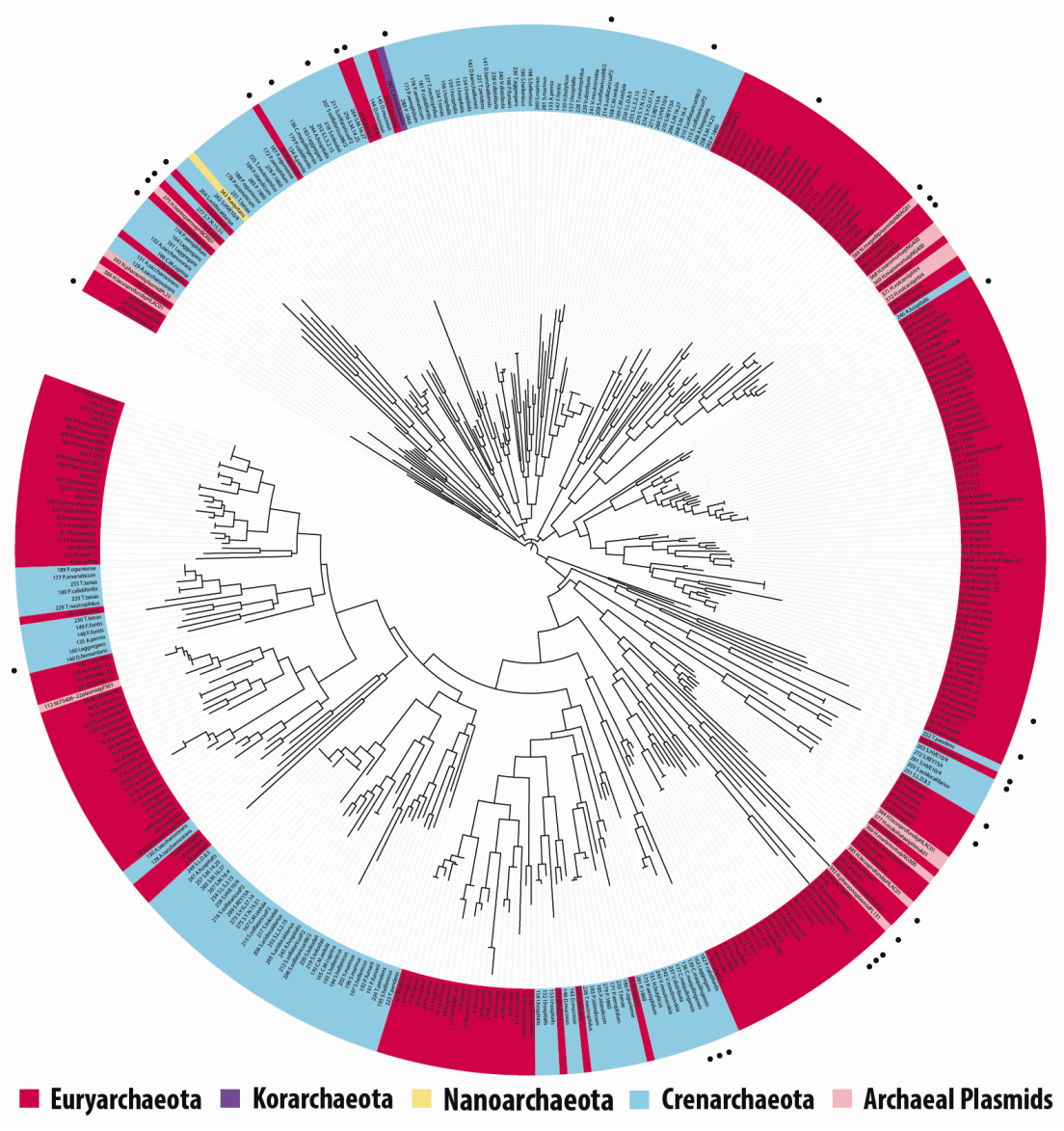

Figure 2: Phylogenetic tree for all CRISPRs across 110 archaeal species. The tree has been colour coded to highlight different archaeal phyla. In addition, the black dots tag the novel CRISPRs represented in figure 1.

Groups	Direct Repeat No.	Direct Repeat Sequence
Methanogens	>6M.mazei >17M.acetivorans	GTTTCAATCCTTGTTTTAATGGAT---------- GTTTCAATCCTTGTTTTAATGGATCTTGCTCTCGAAT
	>48M.aeolicus >58M.igneus	GTCTAAAAGACACATCCATTAAAACAAGGATGGAAAC ----------ATCCATTAAAACAAGGATGGAAAC
	>74M.jannaschii >94M.fervens	-ATTTCCATTCCGAAACGGTCTGATTTTAAC AATTTCCATTCCGAAACGGTCTGATTTTAAC
	>85M.jannaschii >100M.FS406-22 >112M.FS406-22plasmidpFS01G	-TTTCCATCCTCCAAGAGGTCTGATTTTAAC-- GTTTCCATCCTCCAAGAGGTCTGATTTTAACA- -TTTCCATCCTCCAAGAGGTCTGATTTTAAC--
	>114M.thermautotrophicus >115M.marburgensis	GTTAAAATCAGACCAAAATGGGATTGAAAT GTTAAAATCAGACCAAAATGGGATTGAAAT
Thermococci	>293P.abyssi >310P.furiosusCOM1 >336P.NA2 >315P.horikoshii	CTTTCAATTCTATTTTAGTCTTATTGGAAC CTTTCAATTCTATTTTAGTCTTATTGGAAC CTTTCAATTCTATTTTAGTCTTATTGGAAC CTTTCAATTCTATTTTAGTCTTATTGGAAC
	>295P.abyssi >296P.furiosus3638	GTTCCAATAAGACTAAAATAGAATTGAAAG GTTCCAATAAGACTAAAATAGAATTGAAAG
	>328T.onnurineus >333T.kodakarensis	GTTTCAATTCTCTTAGAGTCTTATTGCAAC GTTTCAATTCTCTTAGAGTCTTATTGCAAC
	>331T.gammatolerans >344T.4557	GTTGCAATAAGACTCTAGGAGAATTGAAAG GTTGCAATAAGACTCTAGGAGAATTGAAAG
	>342T.4557 >353T.CL1	CTTTCCACACAATTCTGTCCTACGGAAAC CTTTCCACACAATTCTGTCCTACGGAAAC
Halophiles	>391N.pharaonis >393N.pharaonisplasmidPL23	GTCGAGACGGACTGAAAACCCAGAACGGGATTGAAAC GTCGAGACGGACTGAAAACCCAGAACGGGATTGAAAC
Crenarchaeota	>171P.aerophilum >279P.1860	GTTTCAACTATCTTTTGATTTCTGG GTTTCAACTATCTTTTGATTTCTGG
	>172P.aerophilum >278P.1860 >187P.oguniense	CCAGAAATCAAAAGATAGTTGAAAC CCAGAAATCAAAAGATAGTTGAAAC CCAGAAATCAAAAGATAGTTGAAAC
	>189P.oguniense >235T.tenax >177P.arsenaticum	CTTTCAATCCTCTTTTTGAGATTC CTTTCAATCCTCTTTTTGAGATTC CTTTCAATCCTCTTTTTGAGATTC
	>178P.arsenaticum >188P.oguniense	--CAAAATCAAAAGATAGTTGAAAC GTCAAAATCAAAAGATAGTTGAAAC
	>183P.islandicum >226 P.neutrophilum >185P.islandicum	GTTTCTACTATCTTTTGATTTCTGG GTTTCTACTATCTTTTGATTTCTGG GTTTCTACTATCTTTTGATTTCTG
	>184P.islandicum >225P.neutrophilum	CCAGAAATCAAAAGATAGTAGAAAC CCAGAAATCAAAAGATAGTAGAAAC
	>210S.solfataricus98/2 >248A.hospitalis >258S.M.14.25	GATAATCTACTATAGAATTGAAAG GATAATCTACTATAGAATTGAAAG GATAATCTACTATAGAATTGAAAG

Table 2: Alignment of direct repeats displaying 100% similarity.

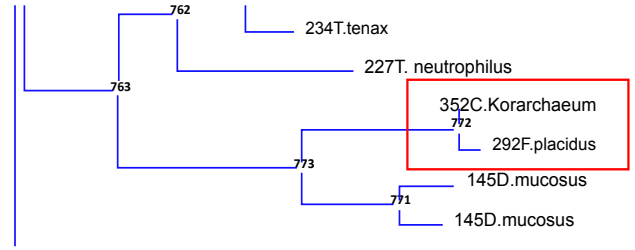

Figure 3: Similarity in direct repeats between the organisms belonging to two different Phyla; *Ferroglobus placidus* DSM 10642 (Phylum– Euryarchaeota) and *Candidatus korarchaeum cryptofilum OPF8*.

coded corresponding to different archaeal phyla. Similarity between the species of two different phyla is seen in the direct repeats of *F. placidus* and *C. Korarchaeum* (Figure 3). In addition, the program MEME [40] was employed (with default parameters) to identify and analyze the consensus motif from the multiple sequence alignment of all 391 direct repeats (Figure 4) and a total of 33 novel CRISPRs are identified. The consensus motif (Figure 4) was also defined in the novel CRISPRs.

Similarity between the spacers

Spacer acquisition occurs in between the repeat sequences of the CRISPR array, when the host genome is infected by a pathogen. This addition of spacers takes place amidst the leader sequence and the prior repeat unit by the assistance of a special feature on the incoming genome sequence known as the Protospacer Adjacent Motif (PAM) and the CAS gene products. Spacers are occasionally repeated, sometimes more than once within a cluster, and can also appear in different arrays within the same chromosome. In some cases, interspecies repetition of a particular spacer can also be visualized which may suggest that the same predator can attack two entirely different organisms. In Archaea, few spacers are found to be repeated at different positions in different CRISPRs within the same organism. The Crenarchaeal strains *Pyrococcus furiosus DSM 3638* and *Pyrococcus furiosus COM1* share 187 spacer matches in their CRISPR unit with *Sulfolobus solfataricus 98/2* and *Sulfolobus solfataricus P2* showing 159 spacer matches in their CRISPR loci. The CRISPR loci in halophilic archaea *Natronomonas pharaonis DSM 2160* and its extra chromosomal plasmid PL23 share seven spacer sequences in common. This repetition of the spacer units in different CRISPR clusters reveals the fact that different clusters get activated and integrate these spacers

Spacer	Size	1. Organism	Crispr No_spacer Position In Array
Spacer similarity within a strain			
AAAGAGGTATCTTTCAACTCAGGAAGTTCATTTGCAG	37	Methanothermococcus okinawensis H1	**4**_42, **9**_17
CAATTTTCCGTCAGATTAAAATTTTGAAATCATTATATCC	40	Methanotorris igneus Kol 5	**18**_4, **19**_4
TAGTCGTAACATCTTCTTATAGTTTTATCGTTAAATA	37		**18**_3, **19**_3
TTTCACACTTGTTTATTATGAAATCCATACAAATCTT	37		**18**_2, **19**_2
ATTGTAAATTGGTTCAAAGATTATGGCATTGATGT	35		**18**_1, **19**_1
AAACAAGGATAGAAAATACTGCACTCGAAAACTCGAAAGTTCGAACACCTA	51		**14**_2, **7**_5
TGAGTAGATATATAGTGGATTTAATAGGCGATGGTTGTTTGAATACCTC	49		**14**_1, **7**_4
GTGGCTGCAGATCGTATATCTGATCATCTCAGAATAATGA	40	Methanosaeta concilii GP6	**5**_8, **7**_2
GCTCTGGCCGCACCCCGGACAATGGCCAAGAAGATGGCGGT	41		**5**_7, **7**_1
CGGCTCATACGGCATGGATGCCCTAACCAAGATGGT	36		**5**_5, **6**_17
AGCTGATCAACGATTCTGCCAGCCAGGGCTTCAAAAAG	38		**5**_4, **6**_16
AGAAGCCCCACAAGCAGAGCCACCGGGAAGAATGCA	36		**3**_3, **4**_6
GTTCACTTTGTAATGCCACGGGCTGTACGCATTTC	35		**3**_2, **4**_5
ATGGAAACGTCTTTGAACAAATGGCAAAGGAAGTTGCCAC	40		**3**_1, **4**_4
TTAAAATAAACGATCCATTTGGTTGGGGTTCGAATAA	37	Methanocaldococcus vulcaniusM7	**10**_16, **20**_1
AAGGGTTCCCCGTCAAAGTCGTTGGAGACAACAGC	35		**10**_15, **20**_2
ACATAGAACAGATTAAAATAAGTGATTTAAAAAGGAAG	38		**10**_11,12,13,14 **20**_3
AGCTAAGCAACCGCAACAACAACAACAAAGCACACTTGC	39	Fervidicoccus fontiskam940	**4**_10,**5**_8
ACTACCTTTGGCAGGAGTTATACGAGAAGAAACTAA	36		**4**_11,**5**_7
AAAAACAATCGAAGAGGACATAGAAAACTTAAGAGAA	37		**4**_12,**5**_6
GGATCCCCAGCGTTGACATATACGAAACGAACTCGTC	37		**4**_13,**5**_5
CGAAGACAAAAGTAAGCAAGATGAAATAGCGGATTT	36		**4**_14,**5**_4
TTGGTAGGCGCAATTACGTTTATTAGCCAACTGAACGC	38		**4**_9,**5**_9
AACGAGGGTAGTGAATTCGAGGAGGGTATCATTATGATGCCCGAGAG	47	Vulcanisaeta distributa DSM 14429	**8**_10,**9**_2
CTCGAGCTGGCTTACTGGTGGTGGCGGTAGTGACTTCGGCACGAG	45		**8**_9,**9**_3
TTCTAATAATCACGAGTACGACGTCTTCGTCGTCAT	36	Pyrococcus yayanosii CH1	**3**_4, **4**_4
TCGTTTTCTTCATTATTATCTATGTCATTACACCTAT	37		**3**_5, **4**_5
CGGGGCGCGAGCCGGCGGGAAATCCGCCCCCGCGGGG	37		**3**_6, **4**_6
CTGGATCTCTCTGAATTCTTCTGGATCACTAGTAACCT	38		**3**_7, **4**_7`
CCCGCCCACTTGTTGCTCATGTAGACCTTTCCTGAC	36	Haloarcula marismortui plasmid pNG400	**1**_1,**3**_40
TCGTTTTCTTCATTATTATCTATGTCATTACACCTAT	38	Haloquadratum walsbyi C23	**2**_6,**3**_12
Spacer similarity among the strains			
TTCTTCGACAGGCAGGGAGAGGAGAGGTACGCGCTGTACAAG	42	(A) Pyrobaculum arsenaticum DSM13514 & (B) Pyrobaculum oguniense TE7	**A3**_88 **B4**_2
ATTAGTATTTGGTCTAGCAGGTCTTTTATTTGTAGATATTCT	42		**A3**_89 **B4**_1
AAATATGTGAAGACGAAGCAATTGCAAGGGTAATATACACTAA	43	(A) Sulfolobus islandicus L D 8 5 & (B) Sulfolobus islandicus L S 2 15	**A3**_1 **B2**_1
TAGGACTATAAATGAAATTAAGAACACGA	29	(A) Sulfolobus islandicus Y N 15 51& (B) Sulfolobus islandicus REY15A	**A1**_1 B**8**_3
CAAGGAGATTGAGGAAACCAAGAAAATAA	29		**A1**_2 B**8**_2
CAATTGCGTCAATGAGTTTATTTAAATCGTCAGCAATTAA	40	(A) Sulfolobus islandicus Y G 57 14 & (B) Sulfolobus islandicus Y N 15 51	**A1**_1 B**4**_65
AAAATAACAAATTTCAAAGAAGGTAAAGTGAAAAGTGAC	39		**A1**_2 **B4**_64
Spacer similarity within chromosome and plasmid			
ATTACGGCGAATGGACGCTCGATATGACGTTTGACGAT	38	Natronomonas pharaonis DSM 2160 & Natronomonas pharaonis DSM 2160 plasmid PL23	**5**_2,**1**_1
GTTGCGGTCGACGATTCAGGCGACCAACTCGAAC	34		**5**_3,**1**_2
AAAGATACGCTCAGGCGTTGTCTCCCGCAACCCAT	35		**5**_4,**1**_3
AACGTCCCTGATTCAACGCGGATTGGGTCGGTTGCA	36		**5**_5,**1**_4
TGAAGGTAAAGATATCGTCGTCAATCATGAAAACC	35		**5**_6,**1**_5
TCGCTAGTCATCGGTCAGGCCCTCCGGGCTGGTGGT	36		**5**_7,**1**_6
CGTCGGCGTCTCGCTCGACGACCGTGCCCGTCGA	34		**5**_8,**1**_7

Table 3: Organisms with CRISPRs sharing same spacer sequences.

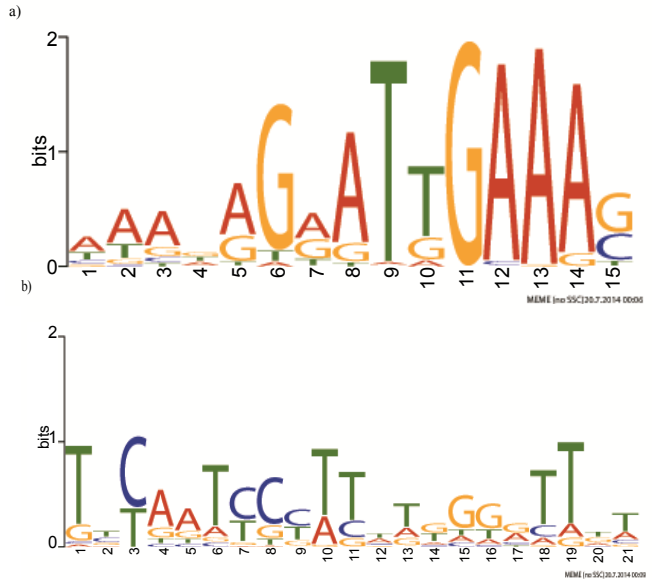

Figure 4: Repeat consensus motif for CRISPR sequences. (a) Consensus motif for entire CRISPR sequence. Out of 391 sequences, 342 sequences were found in the consensus motif region. (b) Consensus motif for novel CRISPRs. Out of 33 novel CRISPRs, 31 CRISPRs sequences were found in the motif regions.

on having an encounter with the foreign elements each time. Although identical groups of spacer-repeat units have been observed in the closely related strains, however, they have not been detected in other species. Within the analyzed CRISPR units of archaea, none of them tends to bare similarity between the spacers except in halophiles. In halophiles, some of the spacers of plasmid CRISPR tend to match with the spacer of the chromosomal CRISPR of the same strain (Table 3).

On the core CAS proteins

A set of genes known as the core CAS genes (Cas1-Cas6) encode a set of enzymes such as the helicases and nucleases which help in the manipulation of the DNA strands. These proteins which are inevitable for the functioning of the CRISPRs have also been analyzed phylogenetically for sequence similarity and predictable evolutionary homology. The FASTA sequences of the core CAS proteins are obtained for all the species of interest from the GenBank. Alignment is carried out between a single CAS protein family found in all the organisms at one stretch (for example; Cas1 protein family of all the species are aligned). The alignment score and the distance guide tree developed as a result of the Clustal alignment are analyzed and used in order to construct the phylograms for each CAS protein family by making use of MEGA. Among the CAS genes, Cas4 and Cas1 are seen in most of the strains (Table 4). Table 5 represents the distribution of CAS gene across 110 archaeal species. By aligning the core CAS proteins of the CRISPRs under study, it is observed that no significant score is given when the Cas3 proteins are aligned and this is also similar to Cas6 proteins. A decent similarity is observed in Cas1, Cas3, Cas4, Cas6 group of protein families. Cas2 gene of *A. hospitalis, S. islandicum HVE10/4* and *S. islandicum REY15AQ* displays 100% similarity. Cas5 genes of *S. solfataricus, S. islandicus LS215* and *YG5714* tends to share 100% similarity.

Organism	Cas 1	Cas 2	Cas 3	Cas 4	Cas 5	Cas 6
Methanosarcina barkeri str fusaro	1			1		1
Methanosalsum zhilinae DSM 4017	1	1	1		1	
Methanococcoides burtonii DSM 6242	2		2	1	1	1
Methanosaeta thermophila PT	2	1	1	2	2	
Methanosaeta concilii GP6	2	1	4	1	3	
Methanosaeta harundinacea 6Ac	1	1		1		
Methanospirillum hungatei JF-1	7			2	2	
Methanoculleus marisnigri JR1	1					
Methanocorpusculum labreanum Z	1			1		
Methanothermococcus okinawensis IH1	1	1		1		1
Methanococcus maripaludis X1	1	1	1	1	1	
Methanococcus maripaludis C5	1	1	1	1	1	
Methanococcus Vol.tae A3	2	2	2	2	2	
Methanococcus vannielii SB	2	2	1	1	1	
Methanococcus aeolicus Nankai-3	1	1	1	1	1	
Methanotorris igneus Kol 5	1	1		1		
Methanocaldococcus vulcanius M7	1	2	1	1	1	
Methanocaldococcus infernus ME	2	2	2	1	2	2
Methanocaldococcus fervens AG86	1	1	1	1	1	
Methanocaldococcus sp. FS406-22	1	1	1	1	1	
Methanobrevibacter ruminantium M1	2	2	1	1	1	1
Methanobacterium sp. SWAN-1	1	1	1	1	1	1
Methanocella sp. HZ254	1	1	1	1		
Acidilobus saccharovorans 345-15	1			2		
Aeropyrum pernix K1	1	1	2		1	
Desulfurococcus kamchatkensis 1221n	1	1		1	1	
Desulfurococcus mucosus DSM 2162	2		1	2	1	
Fervidicoccus fontis Kam940	1	1		1		
Hyperthermus butylicus DSM 5456	1	1	1	2		
Ignicoccus hospitalis KIN4	2	1	1	2		
Desulfurococcus fermentans DSM 16532					1	
Ignisphaera aggregans DSM 17230	1	1	1	2	1	
Metallosphaera cuprina Ar-4	1	1	1		1	
Metallosphaera sedula DSM 5348	1	1	3	1	1	1
Pyrobaculum aerophilum str. IM2	2	2	4	3	1	
Pyrobaculum arsenaticum DSM 13514	1	2	2	3		2
Pyrobaculum calidifontis JCM 11548	1	1	1	2		
Pyrobaculum islandicum DSM 4184				1		1
Pyrobaculum oguniense TE7	3	2	1	2		1
Pyrolobus fumarii 1A	1	1	1	2	2	
Staphylothermus hellenicus DSM 12710				1		
Staphylothermus marinus F1	1	1	1	2	1	
Sulfolobus acidocaldarius DSM 639	1	1		1		6
Sulfolobus solfataricus 98/2	2	2	1	3	1	1
Sulfolobus solfataricus P2	2	2	5	3	3	5
Sulfolobus tokodaii str. 7	3	2	4		1	
Thermofilum pendens Hrk 5	1	1	1	2	1	
Pyrobaculum neutrophilum- strain V24Sta	2	2	2	3	2	
Thermoproteus tenax Kra 1			2	2		
Thermosphaera aggregans DSM 11486	1	1	1	1		
Vulcanisaeta distributa DSM 14429	1	2		2	2	2
Vulcanisaeta moutnovskia 768-28	1	1	1	2		2
Acidianus hospitalis W1	2	2	1	2	1	2
Sulfolobus islandicus L.D.8.5	2	1	1	2		1
Sulfolobus islandicus L.S.2.15	1	2	1	2	1	1
Sulfolobus islandicus M.14.25	2	2	2	2	2	3
Sulfolobus islandicus HVE10/4	2	1	2	2	1	2
Sulfolobus islandicus M.16.27	2	2	2	2	2	3
Sulfolobus islandicus M.16.4	1	1	1	2	1	1

Sulfolobus islandicus REY15A	1	1	1	3	1	1
Sulfolobus islandicus Y.G.57.14	1	1	1	2	1	1
Sulfolobus islandicus Y.N.15.51	1	1	1	2	1	1
Pyrobaculum sp. 1860	3	2	1	3	1	2
Archaeoglobus veneficus SNP6				1		
Ferroglobus placidus DSM 10642	1	1		1		
Pyrococcus furiosus COM1	1					
Pyrococcus yayanosii CH1	2	1	3	3	2	2
Pyrococcus sp. NA2					1	
Thermococcus barophilus MP	1		1	1		2
Thermococcus gammatolerans EJ3			1	1	1	3
Thermococcus sibiricus MM 739	1					
Pyrococcus sp. ST04		2	1	2		3
Thermococcus sp. 4557	2	1	1		1	
Thermococcus sp. AM4	1	1	1	1	1	2
Thermococcus sp. CL1	3	3	1	1	2	2
Candidatus Korarchaeum cryptofilum OPF8 chromosome	1	1	1	1	1	
Haloarcula hispanica ATCC 33960 chromosomeII	1	1	1		1	1
Haloarcula marismortui ATCC 43049 plasmid pNG400	1	1	-	1	-	-
Haloferax volcanii DS2	1	1	1	1	1	1
Halomicrobium mukohataei DSM 12286	1	1		1	1	1
Haloquadratum walsbyi C23	2	2	2	2	1	1
Halorhabdus utahensis DSM 12940	1	1	-	1	-	1
Halorubrum lacusprofundi ATCC 49239 plasmid pHLAC01	2	2	-	2	-	1

Table 4: CAS genes associated with the CRISPR found in the archaeal strains –Among 110 archaeal strains, the following holds CAS genes in their CRISPRs whereas in the remaining 27 strains CAS genes are absent but show the presence of unclassified and putative CAS genes.

Sl. No	Phylum	Number of Cas genes						Percentage (%)
		Cas1	Cas2	Cas3	Cas4	Cas5	Cas6	
1	Euryarchaeota	52	34	32	49	31	22	**40.7**
2	Crenarchaeota	53	47	51	72	33	36	**54**
3	Archaeal Plasmids	5	5	3	5	3	3	**4.4**
4	Korarchaeota	1	1	1	1	1	-	**0.9**
Percentage (%)		**20.5**	**16.1**	**16.1**	**23.5**	**12.6**	**11.3**	

Table 5: Distribution of Cas genes across different archaeal phyla.

Comparative analysis of CRISPR loci in Pyrobaculum genomes: A case study

Among the 110 archaeal species, a general comparative analysis of the CRISPR loci across the Pyrobaculum genomes has been presented. The analyzed seven genomes, namely, *P. aerophilum str. IM2*, *P. arsenaticum DSM 13514, P. calidifontis JCM 11548* , *P. islandicum DSM 4184, P. oguniense TE7, P. neutrophilum–strain V24Sta* and *Pyrobaculum sp. 1860* are rod-shaped archaeal gram-negative species that belong to the archaeal phylum, Crenarchaeota. Each CRISPR loci of the subjected species consists of short DNA sequences and a cluster of CRISPR-associated (CAS) protein coding genes. These CAS genes are associated with the activation of CRISPRs system. The different classes of CAS genes (Cas1, Cas2, Cas3, Cas4, Cas5 and Cas6) are distributed among the Pyrobaculum species. Cas1 (19.8%), Cas2 (18%), Cas3 (18%), and Cas4 (27.8%) gene classes have displayed a higher propensity among the different Pyrobaculum species. Each CRISPR loci consists of an array of repeating sequences interspaced by unique spacers.

Many of the CRISPRs shared similar spacers within the CRISPR loci and across genomes. Few of the strains share the same spacers in their chromosomal and the plasmid genomes. Among the seven species of Pyrobaculum, some species shared similar DRs (Table 2). Similarity between the spacers suggests the fact that the organisms had encountered the same phage or plasmid. While, the purpose of CRISPR loci within the chromosomal and plasmid genome of the same strain having similar spacers suggests protection of the genome from degradation due to 5' overlap of the repeat. Among the analyzed species, *P. aerophilum str. IM2* and *Pyrobaculum sp. 1860* share the similar direct repeat 'GTTTCAACTATCTTTTGATTTCTGG'. While, *P. aerophilum str. IM2, Pyrobaculum sp. 1860* and *P. oguniense TE7* had 'CCAGAAATCAAAAGATAGTTGAAAC'. Finally, *P. arsenaticum DSM 13514* and *P. oguniense TE7* shared 'CTTTCAATCCTCTTTTTGAGATTC'. In case of spacer similarity, *P. arsenaticum DSM13514 and P. oguniense TE7* had similar spacers within their CRISPR unit (Table 3).

For the analysis, the direct repeats are aligned for all the seven Pyrobaculum species using ClustalW and then constructed a phylogenetic tree (Figure 5). According to the phylogenetic tree, the species sharing the similar direct repeats are grouped under the same clade. The tree (Figure 5) suggests a horizontal gene transfer between the different species, which could have been mediated through the

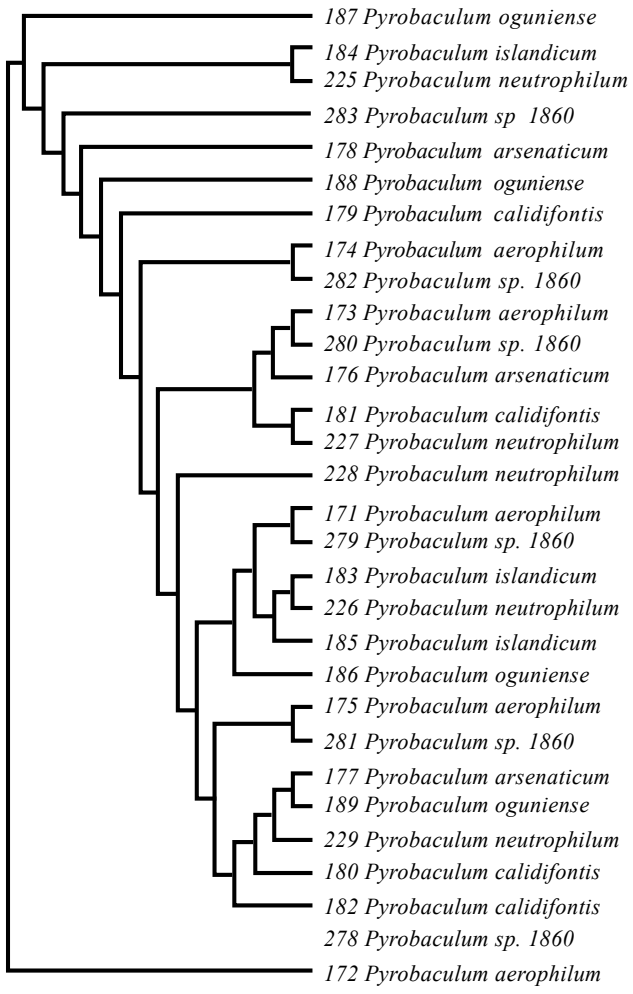

Figure 5: Cladistic analysis of CRISPRs across seven Pyrobaculum species.

plasmids, megaplasmids, and even prophages. The horizontal gene transfer plays a critical role in the distribution and the evolution of CRISPR loci [41]. The existence of DRs might assist the inclusion of DNA segments by recombination and thus suggestively contributing to the evolution of species and their genomic differentiation [30].

BLAST results of CRISPR sequences

The retrieved CRISPRs were subjected to NCBI Blast. The BLAST results did not show any match between the CRISPRs and the human genome sequences. CRISPR loci in *P. yayanosii* chromosome showed a match with the mushroom *Tuber melanosporum mel28* hypothetical protein sequence. A portion of the CRISPR in *M. voltae A3* showed high matches with that of *Leptotrichia buccalis DSM 1135*. The results also revealed that in the archaeal domain, Euryarchaeota holds 55.2% of CRISPRs followed by Crenarcheota with 39.9%, Archaeal plasmid with 3.8% and 0.25% each in Korarchaeota and Nanoarchaeota. Thaumarchaeota do not accommodate any genomes to hold CRISPRs.

Conclusion

A total of 391 confirmed CRISPR loci are detected in the genomes of 110 archaeal species, out of which 33 are found to be neoteric groups that are not marked in the CRISPRdb. The 5' and 3' palindromic motifs that supported the nomenclature of this defensive asset can pave a path for further understanding of the RNA-based CRISPR mechanism. The direct repeats of the CRISPRs may be considered as the products of Horizontal Gene Transfer since they show a phylogenetic relationship with some distant inter–genus species. A set of core protein data retrieved from the databases when aligned and phylogenetically examined, displayed a clean portrait of relationships within the species, highlighting the fact that they would have undergone Horizontal Gene Transfer. Using the results of the present study, a comparative analysis of the CRISPR contents and its functionalities in the complete archaeal domain can be carried out to shed light on the similarities and dissimilarities in the CRISPR organization in them. Many CRISPRs share same spacers within the CRISPR loci and some between the organisms. Some of the strains share the same spacers in their chromosomal and the plasmid genomes. Such spacers protect the strain from degradation due to 5' overlap of the repeat, provided the spacers should be flanked by the same repeats.

Acknowledgement

The authors gratefully acknowledge the facilities offered by the Supercomputer Education and Research Centre and the Interactive graphics facility. One of the authors (KS) thanks the Department of Information Technology (DIT) for the financial support in the form of a research grant.

References

1. Stern A, Keren L, Wurtzel O, Amitai G, Sorek R (2010) Self-targeting by CRISPR: gene regulation or autoimmunity? Trends Genet 26: 335-340.
2. Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, et al. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science 315: 1709-1712.
3. Bolotin A, Quinquis B, Sorokin A, Ehrlich SD (2005) Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology 151: 2551-2561.
4. Horvath P, Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and archaea. Science 327: 167-170.
5. Lillestøl RK, Redder P, Garrett RA, Brügger K (2006) A putative viral defence mechanism in archaeal cells. Archaea 2: 59-72.
6. Kunin V, Sorek R, Hugenholtz P (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol 8: R61.
7. Deveau H, Garneau J E, Moineau S (2010) CRISPR/CAS system and its role in phage-bacteria interactions. Annu Rev Microbiol 64: 476-486.
8. Godde JS, Bickerton A (2006) The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. J Mol Evol 62: 718-729.
9. Mojica FJ, Díez-Villaseñor C, García-Martínez J, Almendros C (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 155: 733-740.
10. Bhaya D, Davison M, Barrangou R (2011) CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annu Rev Genet 45: 273-297.
11. van der Oost J, Jore MM, Westra ER, Lundgren M, Brouns SJ (2009) CRISPR-based adaptive and heritable immunity in prokaryotes. Trends Biochem Sci 34: 401-407.
12. Terns MP, Terns RM (2011) CRISPR-based adaptive immune systems. Curr Opin Microbiol 14: 321-327.
13. Koonin EV, Makarova KS (2009) CRISPR-Cas: an adaptive immunity system in prokaryotes. F1000 Biol Rep 1: 95.
14. Mojica F J M, Diez-Villasenor C, Diez-Villasenor C, Soria E (2005) Intervening Sequences of Regularly Spaced Prokaryotic Repeats Derive from Foreign Genetic Elements. J Mol Evol 60: 174-182.
15. Brodt A, Lurie-Weinberger MN, Gophna U (2011) CRISPR loci reveal networks of gene exchange in archaea. Biol Direct 6: 65.
16. Brouns SJ, Jore MM, Lundgren M, Westra ER, Slijkhuis RJ, et al. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 321: 960-964.
17. Karginov FV, Hannon GJ (2010) The CRISPR system: small RNA-guided defense in bacteria and archaea. Mol Cell 37: 7-19.
18. Makarova KS, Grishin NV, Shabalina SA, Wolf YI, Koonin EV (2006) A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol Direct 1.
19. Jore MM, Brouns SJ, van der Oost J (2012) RNA in defense: CRISPRs protect prokaryotes against mobile genetic elements. Cold Spring Harb Perspect Biol 4.
20. Semenova E, Jore MM, Datsenko KA, Semenova A, Westra ER, et al. (2011) Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc Natl Acad Sci U S A 108: 10098-10103.
21. Sorek R, Kunin V, Hugenholtz P (2008) CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea. Nat Rev Microbiol 6: 181-186.
22. Garrett RA, Vestergaard G, Shah SA (2011) Archaeal CRISPR-based immune systems: exchangeable functional modules. Trends Microbiol 19: 549-556.
23. Barrangou R, Horvath P (2009) The CRISPR System Protects Microbes against Phages, Plasmids. Microbe 4: 224-230.
24. Jansen R, Embden JD, Gaastra W, Schouls LM (2002) Identification of genes that are associated with DNA repeats in prokaryotes. Mol Microbiol 43: 1565-1575.
25. Haft DH, Selengut J, Mongodin EF, Nelson KE (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. PLoS Comput Biol 1: e60.
26. Beloglazova N, Brown G, Zimmerman MD, Proudfoot M, Makarova KS, et al. (2008) A novel family of sequence-specific endoribonucleases associated with the clustered regularly interspaced short palindromic repeats. J Biol Chem 283: 20361-20371.
27. Sinkunas T, Gasiunas G, Fremaux C, Barrangou R, Horvath P, et al. (2011) Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. EMBO J 30: 1335-1342.
28. Carte J, Wang R, Li H, Terns RM, Terns MP (2008) Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes Dev 22: 3489-3496.
29. Makarova KS, Haft DH, Barrangou R, Brouns SJ, Charpentier E, et al. (2011) Evolution and classification of the CRISPR-Cas systems. Nat Rev Microbiol 9: 467-477.

30. Portillo MC, Gonzalez JM (2009) CRISPR elements in the Thermococcales: evidence for associated horizontal gene transfer in Pyrococcus furiosus. J Appl Genet 50: 421-430.

31. Senthilkumar R, Sabarinathan R, Hameed BS, Banerjee N, Chidambarathanu N, et al. (2010) FAIR: A server for internal sequence repeats. Bioinformation 4: 271-275.

32. Gurusaran M, Ravella D, Sekar K (2013) RepEx: repeat extractor for biological sequences. Genomics 102: 403-408.

33. Ahmed Z, Gurusaran M, Narayana P, Kumar KS, Mohanapriya J, et al. (2014) PPS: A computing engine to find Palindromes in all Protein sequences. Bioinformation 10: 48-51.

34. Poddar A, Chandra N, Ganapathiraju M, Sekar K, Klein-Seetharaman J, et al. (2007) Evolutionary insights from suffix array-based genome sequence analysis. J Biosci 32: 871-881.

35. Grissa I, Vergnaud G, Pourcel C (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res 35: W52-57.

36. Grissa I, Vergnaud G, Pourcel C (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics 8: 172.

37. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.

38. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol 28: 2731-2739.

39. Letunic I, Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics 23: 127-128.

40. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34: W369-373.

41. Horvath P, Coûté-Monvoisin AC, Romero DA, Boyaval P, Fremaux C, et al. (2009) Comparative analysis of CRISPR loci in lactic acid bacteria genomes. Int J Food Microbiol 131: 62-70.

This article was originally published in a special issue, **Computational Intelligence in Bioinformatics** handled by Editor(s). Dr. Jean-Christophe Nebel, Kingston University, London, UK