Assessment of the Evolutionary Origin and Possibility of CRISPR-Cas (CASS) Interference Pathway in *Vibrio cholerae* O395

Sajib Chakraborty\(^a\), T.M. Zaved Waise\(^a\), Faizule Hassan\(^a\), Yearul Kabir\(^b\), Mark A. Smith\(^c\) and Mohammad Arif\(^{a,*}\)

\(^a\)Department of Biochemistry and Molecular Biology, University of Dhaka, Bangladesh
\(^b\)Department of Family Sciences, College for Women, Kuwait University, Kuwait
\(^c\)Department of Pathology, Case Western Reserve University, Cleveland, OH, USA

Edited by E. Wingender; received 27 December 2008; revised 25 February 2009; accepted 21 May 2009; published 14 June 2009

ABSTRACT: Bacteria have developed several defense mechanisms against bacteriophages over evolutionary time, but the concept of prokaryotic RNA interference mediated defense mechanism against phages and other invading genetic elements has emerged only recently. Clustered regularly interspaced short palindromic repeats (CRISPR) together with closely associated genes (*cas* genes) constitute the CASS system that is believed to provide a RNAi-like defense mechanism against bacteriophages within the host bacterium. However, a CASS mediated RNAi-like pathway in enteric pathogens such as *Vibrio cholerae* O395 or *Escherichia coli* O157 have not been reported yet. This study specifically was designed to investigate the possibility and evolutionary origin of CASS mediated RNAi-like pathway in the genome of a set of enteric pathogens, especially *V. cholerae*. The results showed that *V. cholerae* O395 and also other related enteric pathogens have the essential CASS components (CRISPR and *cas* genes) to mediate a RNAi-like pathway. The functional domains of a *V. cholerae* Cas3 protein, which is believed to act as a prokaryotic Dicer, was revealed and compared with the domains of eukaryotic Dicer proteins. Extensive homology in several functional domains provides significant evidence that the Cas3 protein has the essential domains to play a vital role in RNAi like pathway in *V. cholerae*. The secondary structure of the pre-siRNA for *V. cholerae* O395 was determined and its thermodynamic stability also reinforced the previous findings and signifies the probability of a RNAi-like pathway in *V. cholerae* O395.

KEYWORDS: CRISPR, *cas* genes, prokaryotic RNA interference, *Vibrio cholerae* O395

INTRODUCTION

Clustered regularly interspaced short palindromic repeats (CRISPR) are found commonly in the bacterial and archaeal genome and are thought to be involved in the bacterial immunity against bacteriophages and other invading foreign DNA elements [1]. CRISPR sequences are composed of direct repeats of 21 to 47 base pairs. Within the CRISPR loci where two adjacent direct repeats are separated by a
spacer of similar length. The locus is generally flanked on one side by a common leader sequence of 200–350 bp [2]. The AT-rich leader sequence may act as a promoter for the transcription process of CRISPR loci [3]. CRISPR-associated genes known as the cas genes are frequently found closely linked to the CRISPR loci. These distinct repeated sequences was first observed in the E. coli genome [4] and later sequence having similar characteristics have been identified in many bacterial and archaeal genome such as Mycobacterium tuberculosis [5], Thermotoga maritima [6], Methanococcus [7].

Four CRISPR associated cas genes (cas1-cas4) are found commonly in association of the CRISPR loci. These four genes encoding four core Cas proteins, represent the core Cas family [2]. In most cases these genes are organized in a head to tail arrangement in a conserved order: cas3-cas4-cas1-cas2. Recently, two new Cas proteins (Cas5 and Cas6) have been identified, which belong to the Cas core protein family. Apart from the core Cas families two other associated protein families (subtype specific proteins and modular proteins) were identified [3]. Haft et al., 2005 [3] have proposed eight CRISPR-associated subtypes, and one CRISPR-associated module by analyzing a number of bacterial and archaeal genomes. The cas genes together with the CRISPR loci constitute the CASS system, which is thought to mediate the prokaryotic RNAi that confers resistance against bacteriophages and foreign genetic elements [8]. The CASS system represents an analogous process to the eukaryotic RNA interference (RNAi) pathway. Several components of the CASS system are found to be analogous to the essential proteins that are involved in the eukaryotic RNAi system such as Dicer, Slicer and RNA-dependent RNA polymerase. A model of prokaryotic RNAi system was predicted by Makarova et al. [1] where the functional roles of core Cas proteins, Cas1, Cas2 and Cas3 and Cas4 in the putative prokaryotic RNAi pathway have been predicted. These proteins were previously thought to be associated with the DNA repair mechanism [9]. There is increasing evidence that these proteins are somehow involved in the defense system within the host bacterium [1]. Recent findings raise the possibility of occurrence of prokaryotic RNAi pathway in different bacterial species such as Streptococcus thermophilus [8,11] and showed that the unique spacers in some of the CRISPR loci are homologous to fragments of bacteriophage and plasmid genes [10].

However, a CASS mediated defense mechanism against bacteriophages in V. cholerae O395 and other related enterobacteria has not been reported to date. Here, we identified and analyzed the necessary components of the CRISPR-Cas (CASS) system, which may mediate the RNAi pathway in V. cholerae O395 and other enteric pathogens. We also investigated the relationship that exists between different CRISPR subtypes and the DR sequences in the corresponding CRISPR loci.

MATERIALS AND METHODS

Analysis of CRISPR sequences

The complete V. cholerae O395 genome and the protein sequences were retrieved from the National Center for Biotechnology Information (NCBI, NIH, Bethesda) FTP site and the Swiss-Prot/TrEMBL database. For the identification of the CRISPR locus we applied CRISPRs finder tools, which enabled the easy detection of CRISPRs in user-submitted sequence data (allows sequences up to 67,000,000 bp). The DR and spacer sequences of V. cholerae O395 were compared to all DRs and spacers deposited in the CRISPRdb database [13,14]. The CRISPR utilities tool was used to make the phylogenetic tree of DR sequences (http://crispr.u-psud.fr/crispr/CRISPRUtilitiesPage.html).
Identification of cas genes and conserved domains

The HMMs (hidden Markov models) for cas genes were obtained from the TIGRFAM database (http://www.jcvi.org/cms/research/projects/tigrfams/overview/). To identify cas genes, all coding sequences were searched with Cas HMMs using hmmpfam (http://hmmer.janelia.org), with the thresholds of an E-value < 0.001 and a positive score. NCBI PSI BLAST program was used to identify the putative gene products. The conserved domains were searched by Interproscan tool developed by EMBL-EBI (http://www.ebi.ac.uk/Tools/InterProScan).

Constructing phylogenetic tree

Multiple sequence alignments (MSA) of the Cas protein sequences were constructed by using the CLUSTALW program (Version 2) of the EMBL-EBI (http://www.ebi.ac.uk/clustalw). The tool TreeTop of GeneBee Molecular Biology Server was used for phylogenetic tree construction (http://www.genebee.msu.su/genebee.html).

Predicting secondary structure of psiRNA (prokaryotic siRNA)

Predictions of the secondary structure of the precursor prokaryotic siRNA were performed by the “Vienna Package server” (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) [15]. For the determination of the stability of the RNA species minimum free energy (entropy) was calculated.

RESULTS

Identification and characterization of CRISPR locus in Vibrio cholerae O395 genome

One CRISPR locus was found in the chromosome 2 of the V. cholerae O395 genome by using the CRISPR finding tool. Forty direct repeats (DRs) were identified within the CRISPR locus. Two adjacent DRs are separated by a unique spacer sequence. A total of 39 unique spacer sequences of almost similar length were identified. The CRISPR sequence lies between 2836101 base pairs to 2838508 base pairs. The DR consensus sequence (GTCTTCCCCACGCAGGTGGGGGTGTTTC) was found to be 28 bp in length. This DR consensus sequence of the V. cholerae O395 strain was subjected to BLAST program in order to find homologous sequences in other bacteria. BLAST output showed that the DR consensus of V. cholerae O395 has significant sequence homology with the DR sequences of several other bacteria. A phylogenetic tree was also constructed by using this MSA to reveal the evolutionary distances among these sequences (Fig. 1).

The spacer sequences were found to be unique as they did not show any similarities with the bacterial genome sequence deposited in the NCBI and TrEMBL database. MSA of the spacer sequences revealed that these sequences do not share any significant homology among themselves. Sequence comparison and BLAST searching revealed that the DR consensus sequence of V. cholerae O395 shares extensive homology with that of C. tepidum TLS (only 2 nucleotide mismatched).
Identification and analysis of the cas genes and proteins

Eight putative cas genes were identified upstream of the CRISPR region in the V. cholerae O395 genome. For the identification of the Cas proteins, the non-redundant database of protein sequences at the National Center for Biotechnology Information (NIH, Bethesda) was iteratively searched by the PSI-BLAST program. Detailed BLAST analyses revealed that Cas3 and Cas1 proteins belong to some “Cluster of Orthologous Group” families (COG1203 and COG1518, respectively) and Cas2 protein belongs to the DUF48 (Domain of Unknown Function) superfamily. We identified 3 Cas core genes (Cas1, Cas2 and Cas3) and 5 E. coli subtype specific genes (Cse1, Cse2, Cse3, Cse4, Cas5e) in the V. cholerae O395 genome. Similar approaches were applied to identify cas genes in the genomes of other bacteria whose DR sequences showed extensive homology with that of V. cholerae O395. All core cas genes (except cas4) and the E. coli subtype specific CRISPR genes were identified in the genomes of V. cholerae O395, Chlorobium tepidum TLS, E. coli O157:H7, Lactobacillus delbrueckii, Salmonella typhimurium and Thermus thermophilus. Dehalococcoides sp. lacks the 2 genes (cas4 and cas2) of the Cas core family.

Predicting evolutionary divergence of Cas 1 and Cas 3 proteins among the selected set of bacteria

MSA for Cas1 and Cas3 proteins of the selected set of bacteria were constructed and utilized to generate phylogenetic tree for Cas1 and Cas3 proteins (Fig. 2A and 2B). In V. cholerae O395 and C. tepidum TLS, the origin for both Cas1 and Cas3 proteins were found to be same as these proteins derived form a common ancestral source. This finding is consistent with the cas gene arrangement and DR homology analysis.
Identification of the conserved domains in Cas3 proteins and comparison to eukaryotic Dicer protein

The conserved domains (CDs) in the Cas3 proteins of *V. cholerae* O395 and *C. tepidum* were identified, analyzed and compared with eukaryotic Dicer protein domains from three distinct eukaryotic species (*Schizosaccharomyces pombe*, *Aspergillus fumigates*, *Drosophila melanogaster*). Both bacteria were found to share seven functional CDs (DEAD-like helicase N-terminal, DNA/RNA helicase C-terminal, metal-dependent phosphohydrolase HD region, Cas3 CRISPR-associated core, helicase associated ATP-binding domain, dsRNA-binding domain and ribonuclease domain). Interestingly several Cas3 domains were found to be exactly homologous to eukaryotic Dicer proteins. The arrangement of the functional domains in the Cas3 proteins and eukaryotic Dicer proteins from selected set of bacterial and eukaryotic organisms are given in Fig. 3.

The helicase domain in both Cas3 and eukaryotic Dicer were found to be derived from the same helicase superfamily. Both the dsRNA-binding domain and ribonuclease domain play a key role for the functionality of the Dicer proteins. These two domains were also found in the Cas3 proteins of *V. cholerae* O395 and *C. tepidum* which provides strong evidence that these proteins play a role in slicing the precursor SiRNA molecule to form a mature prokaryotic SiRNA. Some conserved motifs (nucleotide binding motif and ATP-binding motif) were also found in the Cas3 and Dicer helicase domain indicating the evolutionary conservation of this domain in both types of proteins. The PAZ domain of the eukaryotic
Table 1
List of functional domains of CAS3 and eukaryotic Dicer proteins

<table>
<thead>
<tr>
<th>Domain name</th>
<th>Interpro id</th>
<th>Cas3 proteins</th>
<th>Dicer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>V. cholerae 0395</td>
<td>C. tepidum</td>
</tr>
<tr>
<td>Metal-dependent phosphohydrolase, HD region</td>
<td>IPR003607</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Helicase Cas3, CRISPR-associated, core</td>
<td>IPR006474</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>DEAD-like helicase, N-terminal</td>
<td>IPR014001</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Helicase superfamily associated ATP-binding domain</td>
<td>IPR014021</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>DNA/RNA helicase, C-terminal</td>
<td>IPR001650</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Ribonuclease III</td>
<td>IPR000999</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Double stranded RNA binding domain</td>
<td>IPR001159</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>PAZ domain (Argonaute protein family domain)</td>
<td>IPR003100</td>
<td>N</td>
<td>N</td>
</tr>
</tbody>
</table>

Domains identified are given along with their interproscan identity number. One unregistered domain was found in three eukaryotic Dicer proteins. Their occurrence or absence in a particular organism is indicated by “Y” and “N” respectively.
Fig. 3. Arrangement of functional domains of CAS3 and Dicer proteins of different bacteria and eukaryotes. Metal-dependent phosphohydrolase-HD region, \( \text{\textbullet} \); CRISPR-associated core, \( \text{\textbullet} \); DEAD-like helicase, N-terminal/DNA/RNA helicase, \( \text{\textbullet} \); helicase associated ATP-binding domain, \( \text{\textbullet} \); DNA/RNA helicase, C-terminal, \( \text{\textbullet} \); ribonuclease III, \( \text{\textbullet} \); PAZ domain (Argonaute protein family domain), \( \text{\textbullet} \); double-stranded RNA-binding domain, \( \text{\textbullet} \); unregistered domain, \( \text{\textbullet} \).

Dicer is absent in the Cas3 protein, whereas the HD phosphohydrolase domain is only present in the Cas3 not in Dicer. The extensive homology of the Cas3 proteins of \textit{V. cholerae} O395 and \textit{C. tepidum} with Dicer proteins in terms of functional domains is shown in Table 1.

\textit{Prediction of pSiRNA precursor secondary structure}

In this study we predicted the nature of the secondary structure of the precursor prokaryotic siRNA (psiRNA), transcribed from the CRISPR locus in the chromosome 2 of \textit{V. cholerae} O395 (Fig. 4). Prediction of minimum free energy and thermodynamic ensemble suggested that the secondary RNA structure was thermodynamically stable. We predicted the MFE (minimum free energy) structure, TE (thermodynamic ensemble of RNA structures), and the centroid structure. Any significant disparity was not observed among these structures which can be considered as a proof of the stability of the secondary structure of pre-siRNA.

\textit{DISCUSSION}

Our results suggest that the CRISPR-Cas (CASS) system of \textit{V. cholerae} O395 and \textit{C. tepidum} TLS, a thermophilic green sulfur bacteria, have almost identical features in terms of direct repeat (DRs) sequences, \textit{cas} gene arrangement and Cas protein functions and thus provide strong evidence of their evolutionary relevance. Phylogenetic analyses of the DRs of both \textit{V. cholerae} O395 and \textit{C. tepidum} TLS showed that these repeat sequences have diverged from a common ancestral node (Fig. 1). Although significant homology was found between the DRs of these two bacteria, no sequence identity was observed among the spacer sequences in their respective genomes. This finding reinforced the previous observations that the spacers are acquired by consequent phage infection and/or plasmid invasion [8].
Fig. 4. The predicted RNA secondary structure of the CRISPR loci of *Vibrio cholerae* O395 (A) and a mountain plot (B) representation of the MFE structure, the thermodynamic ensemble of RNA structures, and the centroid structure. Additionally the positional entropy for each position is also shown.

Spacers are the most dynamic component of the CASS system which is reflected by their diversity. There is growing body of evidence that the unique spacers in some of the CRISPR loci are homologous to fragments of bacteriophage, suggesting that spacer sequences within the CRISPR loci may be derived from bacteriophages [10,16].

A close evolutionary relationship between the CASS system of *V. cholerae* O395 and Chlorbium tepidum TLS was delineated when the phylogenetic tree was constructed for Cas1 and Cas3 proteins from the selected set of bacteria (Fig. 2A and 2B). Although the taxonomic classification showed that *V. cholerae* O395 and *C. tepidum* TLS are distantly related to each other, significant homology in their CASS components raises the possibility of horizontal transfer of CRISPR-Cas locus between these two bacteria. The expansion of the CRISPR loci largely depends upon the ability of the bacteria to incorporate foreign genetic elements within the repeat clusters [17]. Previous studies showed that the bacteria capable of incorporating phage DNA sequence as spacers into the CRISPR loci becomes resistant to that phage infection [18,19]. Recently it was hypothesized that the CASS mediated bacterial immunity to phage infection and reproduction may be achieved by a eukaryotic RNAi-like pathway [20]. In this study we evaluated this hypothesis in the perspective of *V. cholerae* O395 genome. Furthermore, to assess the possibility of the RNAi-like pathway in *V. cholerae* O395 we identified the essential proteins of the prokaryotic RNAi pathway. PSI BLAST results showed that the Cas3 and Cas1 proteins of *V. cholerae* O395 belong to clusters of orthologous groups, COG1203 and COG1518 respectively. COG1518 is considered as the universal marker of the CASS system. Previous findings showed that the *cas3* gene encodes a highly conserved α-helical protein of the COG1203 family which is believed to act as nuclease/integrase whereas the Cas3 protein (COG1203) is usually long and mostly contains a fused
The Cas3 protein plays a similar role as eukaryotic Dicer in the putative pRNAi pathway and hence is considered as prokaryotic Dicer (pDicer) [1]. In line with previous observations we found Cas3 gene in *V. cholerae* O395 encodes an 837 amino acid long protein with a fused HD hydrolase domain. We conducted detailed analyses of the conserved domains of related Cas3 proteins of different bacteria including *V. cholerae* O395 and *C. tepidum* and compared these prokaryotic domains with those of eukaryotic Dicer (eDicer) proteins. Certain conserved domains of the Cas3 proteins (N- and C-terminal helicase domains, dsRNA-binding domains and ribonulease domain) were identified in the Dicer proteins of different eukaryotic species such as yeast (*Schizosaccharomyces pombe*), fungus (*Aspergillus fumigatus*) and arthropods (*Drosophila melanogaster*) (Table 1). This finding signifies the possibility of the involvement of Cas3 proteins in the pRNAi pathway. Moreover the presence of a dsRNA-binding domain cancels the previous hypothesis that these Cas3 proteins may be associated with the DNA repair system in bacteria because a dsRNA-binding domain is unusual for proteins associated with DNA repair mechanism. These conserved domains of Cas3 also found to harbor some common motifs like ATP-binding motif and nucleotide binding site that are also present in the eukaryotic Dicer (eDicer). Considering all the important findings regarding the relevance between Cas3 and eDicer, we assume that the Cas3 protein (COG1203) in *V. cholerae* O395 may play a direct role in unwinding the dsRNA and processing of the pre-psiRNA to produce mature prokaryotic siRNA (psiRNA). A hypothesis has emerged recently that the palindromic nature of the direct repeats might be indicative of a functional RNA secondary structure within the CRISPR loci [12]. This hypothesis was supported by experimental evidence that CRISPR are transcribed to produce a precursor RNA species which is processed into non-messenger RNA in several Archaea [12]. The hypothetical secondary structure of the pre-psiRNA species of *V. cholerae* O395 was found to be thermodynamically stable and viable which provides evidence that the CRISPR loci in the chromosome 2 of *V. cholerae* O305 may be transcribed and processed by Cas3 (pDicer). Considering the above findings we hypothesize that a CASS mediated RNAi pathway may provide immunity in *V. cholerae* O395 and other enteric water borne pathogens such as *E. coli* O157 against bacteriophages and thus possesses a threat to the re-emerging concept of phage therapy against diarrheagenic pathogens. This new concept of prokaryotic RNAi mediated phage resistance may also limit the use of phages as an antimicrobial agent to keep the surface water free from pathogenic bacteria. Due to the absence of effective cholera vaccine, the concept of bacteriophage therapy for the treatment of diarrheal disease caused by *V. cholerae* has been drawn much attention recently [21,22]. The idea of using bacteriophages as antimicrobial agents to kill the water borne pathogens in aquatic environment is also considered as a major approach to combat these pathogenic bacteria [23]. In our study we have shown that a RNAi mediated defense mechanism, analogous to eukaryotic RNA interference pathway, has evolved in enterobacteria such as *V. cholerae* which may provide immunity against bacteriophages. Thus, to make the concept of phage therapy more realistic a thorough investigation of the phage-host interaction should be carried out to better understand the putative molecular mechanisms by which bacteria can resist phage infection.

**ACKNOWLEDGEMENTS**

We like to thank Md. Moazzem Hossain and Md. Fakhrul Kabir for their guidance and support.
REFERENCES


