Asian Journal of
Biotechnology

Molecular Machinery of Crispr-CAS System-RNA Mediated Defense Pathway in Prokaryotes

P. Chellapandi and J. Ranjani

Department of Bioinformatics, School of Life Sciences, Bharathidasan University, Tiruchirappalli-620024, Tamil Nadu, India

Corresponding Author: P. Chellapandi, Department of Bioinformatics, School of Life Sciences, Bharathidasan University, Tiruchirappalli-620024, Tamil Nadu, India Tel: +91-431-2407072 Fax: +91-431-2407045

ABSTRACT

Small interfering RNA or silencing RNA plays a variety of roles in cellular process of organisms. Studies in bacteria and archaea had gene silencing pathways that are triggered by transgene expression or viral replication. RNA silencing mechanism is first recognized as antiviral mechanisms that protect organisms from RNA viruses, or which prevent the random integration of transposable elements. Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) are arrays of prokaryotic DNA sequences that mediate a form of acquired immunity to specific viral pathogens. CRISPR-associated (CAS) proteins play an important role in the initial recognition of phage genetic material and incorporating these proto-spacers in the CRISPR array. Phage-derived spacers are incorporated at the CRISPR 5' leader sequence so that the ordered sequence of spacers essentially gives a temporal record of the infection history in that bacterial population. Hence, CRISPRs offer a high-resolution method for molecular typing of bacterial hosts and their pathogens based on the unique CRISPR signature.

Key words: CRISPER interface, CAS genes, adaptive immune system, small RNA, RNAi, xRISC, RAMP, spoligotyping

INTRODUCTION

Small RNAs perform diverse biological functions. They function by guiding sequence-specific gene silencing at the transcriptional and/or post-transcriptional level. The process by which double-stranded RNAs set off the degradation of homologous RNA is known as RNAi (RNA interference). Several diverse classes of small RNAs, which undergo this process, have been discovered to play key regulatory roles in diverse cellular processes (Mondal, 2003). According to their origin or function, three types of naturally occurring small RNAs have been described; short interfering RNAs (siRNAs), repeat-associated short interfering RNAs (rasiRNAs) and microRNAs (miRNAs) (Ding and Voinnet, 2007). In nature, double stranded RNA (dsRNA) can be produced by RNA-templated RNA polymerization or by hybridization of overlapping transcripts. Such dsRNAs give rise to siRNAs or rasiRNAs, which generally guide mRNA degradation and/or chromatin modification (Mojica *et al.*, 2005). In addition, endogenous transcripts that contain complementary or near-complementary 20-50 bp inverted repeats fold back on themselves to form dsRNA hairpins. These dsRNAs are processed into miRNAs that mediate translational repression and they may also guide mRNA degradation (Mondal, 2003; Ding and Voinnet, 2007).

DISCOVERY OF RNAi PATHWAY

RNA-mediated gene silencing can be categorized into two partially overlapping pathways; RNAi pathway and miRNA pathway. Both pathways have been widely distributed in eukaryotes wherein they are served as a defense system to protect host genomes from foreign genetic elements. RNAi is triggered by either endogenous or exogenous dsRNA and silences endogenous genes carrying homologous sequences. The dsRNA molecules in RNAi pathway are processed by Dicer RNase III proteins into small RNAs, which are then loaded into silencing complexes either RNA Induced Silencing Complexes (RISC) for post-transcriptional silencing or RNA-induced initiation of transcriptional gene silencing (RITS) complexes for transcriptional silencing (Mondal, 2003; Lares *et al.*, 2010).

CRISPR-CAS ARCHITECTURE

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) are a peculiar family of DNA repeats widely distributed in the genomes of bacteria and archaea. CRISPR loci are present in ~40% of eubacterial genomes and nearly all archaeal genomes sequenced to date and consist of short (~24-48 nucleotide) repeats separated by similarly sized, unique spacers. They show characteristics of both tandem and interspaced repeats (Held *et al.*, 2010). There is a tendency for existing CRISPRs in thermophiles more than in mesophiles. Some of the CRISPR sequences from *Thermoanaerobacter tengcongensis* and *Thermus thermophilus* HB8 are unique to the megaplasmid on which others are shared with bacterial chromosomes. Geographic distribution of spacers and CRISPR loci and their diversity within extremophiles are sometimes providing the importance to reveal its hypothetical molecular machinery reliably (Kunin *et al.*, 2007; Diez-Villasennr *et al.*, 2010; Deveau *et al.*, 2010). CRISPR-CAS system has comprised the following genome components (Mojica *et al.*, 2005; Horvath and Barrangou, 2010). They are:

Direct Repeat (DR): A CRISPR is a succession of 21-47 bp sequences called direct repeat separated by unique sequences of a similar length. The repeats are almost always identical with respect to size and sequence. Despite being divergent between species, repeats can be clustered into at least 12 major groups. Some of the larger groups contain a short (5-7 bp) palindrome which has been inferred to contribute to an RNA stem-loop secondary structure of the repeat. It can be supported both by existence of compensatory mutations in the repeats that maintain the stem structure and by observations that repeat-spacer array is transcribed into RNA. Thus, both the structural features and the conserved 3' motif GAAA(C/G) have been suggested to act as binding sites for one or more of the CRISPR-associated (CAS) proteins (Jansen *et al.*, 2002; Mojica *et al.*, 2005; Kunin *et al.*, 2007).

Leader: Chromosomal clusters carry a flanking sequence at one end is referred to as a leader, although its function is unknown. The leader sequences tend to be rich in short homopolynucleotide sequences and AT-rich regions and they lack open reading frames. Archaeal flanking sequences range in size from 132 bp for *Haloarcula marismortui* to 564 bp for *Methanopyrus kandleri*and they directly adjoin the first repeat sequence of the cluster (Lillestol *et al.*, 2006). They invariably occur at the same end of the cluster with respect to the strand orientation of the repeat sequence. There is an approximate direct correlation between the sequence length and the optimal growth temperature of the organism. A new repeat-spacer unit is almost always added to the CRISPR array between the leader and the previous unit, which suggests that the leader could function as

a recognition sequence for the addition of new spacers. The leader has also been suggested to act as the promoter of the transcribed CRISPR array, as it is found directly upstream of the first repeat (Pougach *et al.*, 2010).

Spacer: The repeated sequences, typically specific to a given CRISPR locus, are interspaced by variable sequences of constant and similar length, called spacers. The size of them is usually 20 to 58 bp depending on the species or the CRISPR locus. Spacers vary in size from 35 to 44 bp between clusters as well as between organisms (Aklujkar and Lovley, 2010; Pougach *et al.*, 2010). All spacer sequences within a cluster and within a chromosome are generally different. Spacers are occasionally repeated, sometimes more than once within a cluster and can appear in different clusters within the same chromosome (Lillestol *et al.*, 2006).

CAS GENE FAMILY

CAS genes are always found closely linked to the repetitive sequences. A comprehensive bioinformatics analysis of the CAS system in sequenced genomes resulted in a refined classification with 25 gene families and at least nine types of the CAS operon organization (Jansen *et al.*, 2002) and examples of some CAS operons are depicted in Fig. 1. Eight CAS protein families have been predicted to possess nuclease activity and nine families have been characterized as putative RNA-binding proteins (RAMP domain proteins). Two families have been predicted to possess heliCASe and DNA/RNA polymerase (Horvath and Barrangou, 2010). Phylogenetic studies performed on the CAS protein family suggest that CRISPRs are acquired by horizontal transfer activity (Makarova *et al.*, 2006; Marraffini and Sontheimer, 2009, 2010). The list of CAS proteins and corresponding molecular functions are summarized in Table 1.

CAS1 gene encodes a highly conserved protein and that is represented in all CAS neigh borhoods, with the single exception of thermophilic archaea, *Pyrococcus abyssii* (Makarova *et al.*, 2006). CAS1 protein may act as a novel nuclease/ integrase. It has a metal-dependent DNase activity so that it may be involved in the initial recognition and acquisition of viral motifs. It is used as the best marker of the CRISPR-associated systems in prokaryotic genomes (Jansen *et al.*, 2002; Deveau *et al.*, 2010).

CAS2 is another common gene in CAS gene family, which usually located immediately downstream of CAS1 gene. The members CAS2 superfamily are small proteins with 80-120 amino acids. It has distinct structural motifs, particularly an N-terminal β-strand followed by a polar amino acid (aspartate or asparagines). CAS2 family proteins are ssRNA-specific endoribonucleases and cleave it within U-rich regions. CAS3 gene in CAS system encodes heliCASe, which is unusually long and proteins. CAS4 gene encodes a RecB family nuclease usually containing a
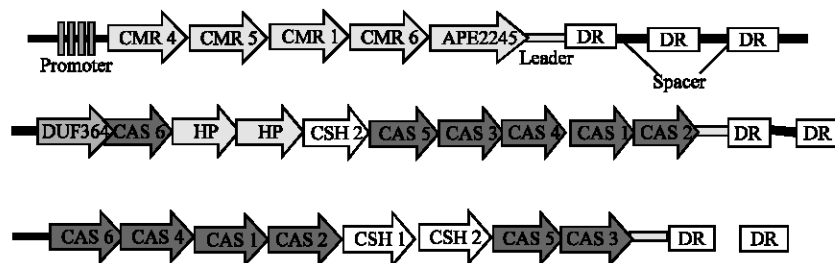

Fig. 1: Some of the CAS operon/gene clusters

Table 1: Functional description of CAS protein superfamily

Gene	Protein function
CMR1	RAMP, CAS like function
CMR4	RAMP, CAS like function
CMR5	RAMP, CAS like function
CMR6	RAMP, CAS like function
CAS1	Nuclease/ integrase
CAS2	ssRNA-specific endoribonucleases
CAS3	CRISPR-associated helicase
CAS4	RecB family exonucleases
CAS5	Unknown
CAS6	Endoribonucleases
HD domain	HD like nuclease
DevR	CRISPR-associated autoregulator
DxTHG motif	CRISPR-associated protein
APE2256	CRISPR-associated protein

C-terminal Zn cluster. RAMP family is belonged to CAS5 superfamily. CAS6 protein acts as a novel endoribonuclease. CAS6 protein interacts with a specific sequence motif in the 5' region of the CRISPR repeat element and cleaves at a defined site within the 3' region of the repeat and then release individual invader targeting RNAs. The functions of several other CAS gene families remain obscure (Makarova *et al.*, 2006; Koonin and Makarova, 2009; Garneau *et al.*, 2010).

The *Dev* and CAS genes in *Myxococcus xanthus* are co-transcribed with a short upstream gene and at least two repeats of the downstream CRISPR located at the *dev* locus (forming the *dev* operon), which is adjacent to CRISPR system (Viswanathan *et al.*, 2007).

STRUCTURE OF RAMPS

RAMPs are the most diverse class of CAS genes which have G-rich loop at the C-terminus. The crystal structure of this protein from *Thermus thermophilus* reveals that the RAMP module is a duplication of a ferredoxin-like fold domain. Each domain has a two layer $\alpha+\beta$ architecture. It is composed of four β-strands and two α-helixes. It contains a small domain preceding the HD-hydrolase N-terminal domain present in many CAS3 helices topologically arranged as a repeat of two $\beta\alpha\beta$ units. The N-terminal ferredoxin-like domain contains two additional α-helices inserted before and after the first α-helix. The C-terminal domain has two disordered regions and houses the conserved Gly-rich loop situated between the last α-helix and β-strand (Makarova *et al.*, 2006).

DEFENSE MECHANISMS AGAINST FOREIGN GENETIC ELEMENTS

Genomes are potential targets of invasion by molecular parasites such as viruses and transposable elements and organisms have evolved RNA-directed defense mechanisms to cope with the constant threat of genome invaders (Barrangou *et al.*, 2007). CRISPR-CAS systems protect microbial cells from invasion by foreign genetic elements using presently unknown molecular mechanisms. The proposed CRISPR-CAS pathway is represented in Fig. 2. Although, CRISPRs found to play a major role in replicon partitioning, DNA repair, regulation, DNA mobilization and chromosomal rearrangement. The CRISPR spacers and repeats are transcribed and processed into small CRISPR RNAs (crRNAs) that specify acquired immunity against bacteriophage infection by a mechanism that relies on the strict identity between CRISPR spacers and phage targets
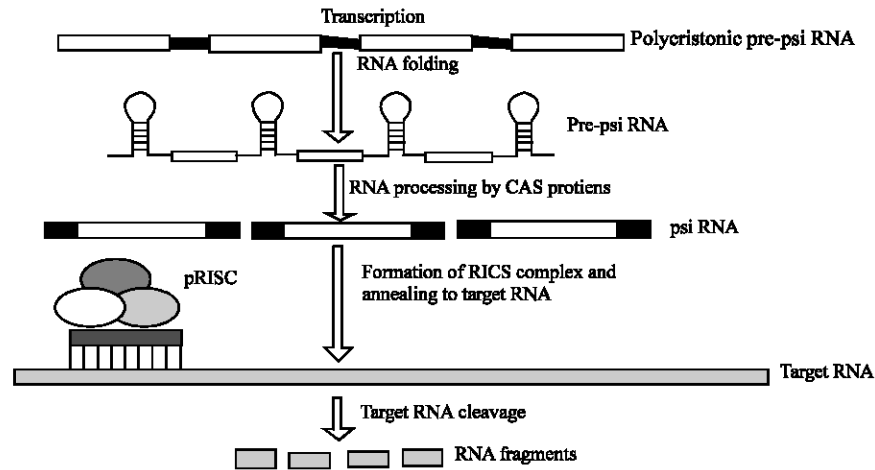
Fig. 2: The proposed molecular mechanism of psiRNA processing to cleave the target RNA molecule
psiRNA: Prokaryotic short interfering RNA; xRISC: prokaryotic RNA induced silencing complexes

Step 1: Cas proteins complexes target and cleave short recognition motifs in the phage genome and incorporate these proto-spacers into the host genome at the 5' end of the CRISPR locus.

Step 2: These incorporated spacers, flanked on either side by partial repeat sequences, are transcribed as crRNAs resulting in large-scale amplification of that specific sequence.

Step 3: Interference with phage-derived sequences is again mediated by Cas-protein complexes thereby crRNAs serve as templates to target conserved viral motifs in subsequent infections.

(Ding and Voinnet, 2007; Horvath and Barrangou, 2010; Vale and Little, 2010). Some spacer sequences match sequences in phage genomes and these spacers can be derived from phage and subsequently help protect the cell from infection (Levin, 2010). As a result, an active CRISPR repeat array may evolve rapidly. Repeat clusters are generally highly conserved in their repeat sequences and in the sizes of their repeats and spacers. Such integrity extends to an almost complete lack of insertion sequences or other mobile elements. Thus, the large number of mobile elements in some archaeal and bacterial genomes and the large variety of spacer sequences could provide potential target sites for the insertion of mobile elements. Many reports support the existence of an RNA-mediated genome defense pathway in archaea and numerous bacteria that has been hypothesized to parallel the eukaryotic RNAi pathway. However, these regulatory RNAs are not generally considered to be analogous to miRNAs because the dicer enzyme is not involved (Mojica *et al.*, 2005; Ding and Voinnet, 2007; Sorek *et al.*, 2008; Marraffini and Sontheimer, 2009; 2010).

FUTURE SCENARIO

Repetitive sequences are common in the prokaryotic genomes and their identification is increasingly facilitated by the availability of sequences of complete bacterial genomes (Chellapandi *et al.*, 2007, 2009, 2010a). Indeed, prophages account for most of the genetic variation among closely related strains and have greatly contributed to their evolution (Chellapandi, 2010). Unlike mesophiles, thermophiles have unique molecular mechanism to uptake foreign genetic

elements. However, phylogenetic analyses clearly revealed the function of identified key proteins which functions still not known in thermophiles. Phylogenetic relationship of CAS protein family in thermophilic bacteria was absolutely varied from thermophilic archaea. The length, sequence and position of the sequences of the protein families in archaeal genome are highly variable and often unique for a single strain (Karthigeyan *et al.*, 2007; Vedhagiri *et al.*, 2009; Chellapandi *et al.*, 2007, 2009, 2010a). Hence, unique extremophilic adaptations to drastically varying biosystems have aroused special interests in their respective potential in biotechnological applications. Available genome sequences of thermophiles and bioinformatics databases, resources and software tools provide an importance for revealing the molecular hypothesis on CRISPR-CAS system and metabolomes of prokaryotes (Razia *et al.*, 2010; Chellapandi and Kalaimathy, 2010; Chellapandi and Dhivya, 2010; Chellapandi *et al.*, 2010b).

APPLICATION OF CRISPR-CAS SYSTEM

CRISPR-CAS system has become a widely used tool to knock down and analyze the function of genes, especially in non-model organisms where the systematic recovery of mutants is not feasible. The potential for using this mechanism is to identify new drug targets and for designing small molecule drugs. Perhaps, this system would be useful for strain typing of pathogenic bacteria, engineered defense against virus in industrial bacteria, selective silencing of endogenous genes for revolutionize microbial-physiology research by CRISPR arrays (Spoligotyping) (Makarova *et al.*, 2006; Sorek *et al.*, 2008; Cadmus *et al.*, 2010).

GENERAL DEFINITION RELATED TO CRISPR-CAS SYSTEM

- **Spoligotyping:** A technique is used to differentiate different strains of the same species according to differences in the spacers in their CRISPR arrays
- **MiRNAs:** Genomically encoded non-coding RNAs that help regulate gene expression, particularly during development
- **RISC:** A multiprotein complex that incorporates one strand of a siRNA or miRNA
- **RITS:** A form of RNA interference by which siRNA - trigger the downregulation of transcription of a particular gene or genomic region.
- **Dicer:** An endoribonuclease in the RNase III family that cleaves dsRNA and pre-microRNA into short dsRNA fragments about 20-25 nucleotides long, usually with a two-base overhang on the 3' end.
- **RAMP:** A cytoplasmic and nuclear protein that associates with double-strand or single-strand RNAs through RNA recognition motif.
- **RasiRNA:** A subclass of Piwi-interacting RNAs (piRNAs), which are small RNA molecule which interact with Piwi proteins
- **Horizontal gene transfer:** Any process in which a bacterium incorporates genetic material from distantly-related bacteria without being the offspring of that organism

REFERENCES

Aklujkar, M. and D.R. Lovley, 2010. Interference with histidyl-tRNA synthetase by a CRISPR spacer sequence as a factor in the evolution of *Pelobacter carbinolicus*. BMC Evol. Biol., 10: 230-230.

Barrangou, R., P. Boyaval, S. Moineau, D.A. Romero and P. Horvath, 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Sci., 315: 1709-1712.

Cadmus, S., V. Hill, D. van Soolingen and N. Rastogi, 2010. Spoligotype profile of *Mycobacterium tuberculosis* complex strains from HIV positive and negative patients in Nigeria: A comparative analysis. J. Clin. Microbiol.,

Chellapandi, P., C. Karthigeyan and S. Sivaramakrishnan, 2007. Phylogenomic analysis of archaeal domain-An overview. Res. Rev. J. Biosci., 1: 160-168.

Chellapandi, P., C. Karthigeyan and S. Sivaramakrishnan, 2009. Evolutionary implication of protein secondary structure among archaea and bacteria. Internet J. Genomics Proteomics, Vol. 4,

Chellapandi, P., 2010. Molecular evolutionary mechanism of methanogens based on their protein superfamily proteins-An *in silico* approach. Ph.D. Thesis, Bharathidasan University, Tiruchirappalli, India.

Chellapandi, P. and S. Kalaimathy, 2010. Molecular aspects of b-galactosidase production system in *Aspergillus* genomes. J. Adv. Dev. Res., 1: 81-89.

Chellapandi, P. and C. Dhivya, 2010. Overview of microbial metabolomics: A special insight to cyanobacterial methylotrophy. J. Adv. Dev. Res., 1: 59-73.

Chellapandi, P., C. Karthigeyen and S. Sivaramakrishnan, 2010a. Superprotein families based phylogenomic analysis of archaeal domain. Proteomics Res. J.,

Chellapandi, P., S. Sivaramakrishnan and M.B. Viswanathan, 2010b. Systems biotechnology: An emerging trend in metabolic engineering of industrial microorganisms. J. Comput. Sci. Syst. Biol., 3: 43-49.

Deveau, H., J.E. Garneau and S. Moineau, 2010. CRISPR/Cas system and its role in phage-bacteria interactions. Ann. Rev. Microbiol., 64: 475-493.

Diez-Villaseñnr, C., C. Almendros, J. Garcia-Martínez and F.J. Mojica, 2010. Diversity of CRISPR loci in *Escherichia coli*. Microbiol., 156: 1351-1361.

Ding, S.W. and O. Voinnet, 2007. Antiviral immunity directed by small RNAs. Cell, 130: 413-426.

Garneau, J.E., M.E. Dupuis, M. Villion, D.A. Romero and R. Barrangou *et al.*, 2010. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature, 468: 67-71.

Held, N.L., A. Herrera, H. Cadillo-Quiroz and R.J. Whitaker, 2010. CRISPR associated diversity within a population of *Sulfolobus islandicus*. PLoS One, 5: e12988-e12988.

Horvath, P. and R. Barrangou, 2010. CRISPR/Cas, the immune system of bacteria and archaea. Sci., 327: 167-170.

Jansen, R., J.D. Embden, W. Gaastra and L.M. Schouls, 2002. Identification of genes that are associated with DNA repeats in prokaryotes. Mol. Microbiol., 43: 1565-1575.

Karthigeyan, C., S. Sivaramakrishnan and P. Chellapandi, 2007. Phylogenomic analysis of archaeal domain. Bioinform. Trends, 2: 37-55.

Koonin, E.V. and K.S. Makarova, 2009. CRISPR-Cas: An adaptive immunity system in prokaryotes. Biol. Rep., 1: 95-95.

Kunin, V., R. Sorek and R.P. Hugenholtz, 2007. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Gen. Biol., 8: R61-R61.

Lares, M.R., J.J. Rossi and D.L. Ouellet, 2010. RNAi and small interfering RNAs in human disease therapeutic applications. Trends Biotechnol., 28: 570-579.

Levin, B.R., 2010. Nasty viruses, costly plasmids, population dynamics and the conditions for establishing and maintaining CRISPR-mediated adaptive immunity in bacteria. PLoS Genet., 6: e1001171-e1001171.

Lillestol, R.K., P. Redder, R.A. Garrett and K. Brugger, 2006. A putative viral defence mechanism in archaeal cells. Archaea, 2: 59-72.

Makarova, K.S., N.V. Grishin, S.A. Shabalina, Y.I. Wolf and E.V. Koonin, 2006. A putative RNAinterference 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: 7-7.

Marraffini, L.A. and E.J. Sontheimer, 2009. Invasive DNA, chopped and in the CRISPR. Struct., 17: 786-788.

Marraffini, L.A. and E.J. Sontheimer, 2010. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat. Rev. Genet., 11: 181-190.

Mojica, F.J., C. Diez-Villasenor, J. Garcia-Martinez and E. Soria, 2005. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J. Mol. Evol., 60: 174-182.

Mondal, S., 2003. RNA interference-towards RNA becoming a medicine. Resonance, 8: 42-49.

Pougach, K., E. Semenova, E. Bogdanova, K.A. Datsenko, M. Djordjevic, B.L. Wanner and K. Severinov, 2010. Transcription, processing and function of CRISPR cassettes in *Escherichia coli*. Mol. Microbiol., 77: 1367-1379.

Razia, M., R.K. Raja, K. Padmanaban, S. Sivaramakrishnan and P. Chellapandi, 2010. Phylogenetic approach for assigning function of hypothetical proteins in *Photorhabdus luminescens* subsp. laumondii T101 genome. J. Comp. Sci. Syst. Biol., 3: 21-29.

Sorek, R., V. Kunin and P. Hugenholtz, 2008. CRISPR- A widespread system that provides acquired resistance against phages in bacteria and archaea. Nat. Rev. Microbiol., 6: 181-186.

Vale, P.F. and T.J. Little, 2010. CRISPR-mediated phage resistance and the ghost of coevolution past. Proc. Royal Soc. B: Biol. Sci., 10.1098/rspb.2010.0055.

Vedhagiri, K., K. Natarajaseenivasan, P. Chellapandi, S.G. Prabhakaran, J. Selvin, S. Sharma and P. Vijayachari, 2009. Evolutionary implication of outer membrane lipoprotein-encoding genes *ompL1*, *ompL32* and *lipL41* of pathogenic *Leptospira* species. Genomics Proteom. Bioinform., 7: 96-106.

Viswanathan, P., K. Murphy, B. Julien, A.G. Garza and L. Kroos, 2007. Regulation of dev, an operon that includes genes essential for *Myxococcus xanthus* development and CRISPR-associated genes and repeats. J. Bacteriol., 189: 3738-3750.