**Genome-based analysis of rhodopsin and CRISPR function in *Halomicrobium mukohataei* and other halophiles**

Olivia Ho-Shing, Sarah Pyfrom and Katie Richeson

Biology Department, Davidson College, Davidson, NC 28035

|  |
| --- |
| **Abstract**  *Halomicrobium mukohataei* is a haloarchaea whose genome has recently been sequenced, but lacks more detailed annotation and investigation. Here, we focus our investigation on a light-driven energy component and an antiviral component in the *H. mukohataei* genome. Bacteriorhodopsin is an archaeal rhodopsin – a light-induced membrane receptor – that functions as a proton pump to generate photochemical energy. Clustered regularly interspaced short palindromic repeats (CRISPRs) are repetitive DNA structures found in bacteria and archaea that contain portions of viral sequences. CRISPRs are thought to confer immunity against viral infection. This study analyzes rhodopsins and CRISPR sequences to validate the automated genome annotation, and compares these elements to those found in other haloarchaea. We found that some bacteriorhodopsin sequences predicted by the Joint Genome Institute might be rhodopsins with different functional roles from bacteriorhodopsin. We also found that distinct CRISPR structures and potential function might also be inferred by sequence comparison among halophiles. Possible interpretations and suggestions for further study of rhodopsins and CRISPRs are discussed. |

**Introduction**

*Halomicrobium mukohataei* is an extremely halophilic archaea isolated from Argentinean salt flats in the Andes highlands. *H. mukohataei* grows strictly aerobically in 3.0 to 3.5 M salt concentrations. Based on a phylogenetic analysis of halobacterial strains using 16S rRNA, *H. mukohataei* is closely related to, yet distinct from the genus *Haloarcula*, which is also characterized by extreme halophiles. (Ihara et al, 1997) *H. mukohataei* can metabolize numerous sugars as an energy source, but also contains another potential energy component – light-detecting retinal proteins called rhodopsins that are found primarily in members of the order *Halobacteriales*. *H. mukohataei* is a species of interest because its discovery provides an opportunity to further investigate cellular mechanisms in poorly understood archaeal species, particularly this light-driven energy component that could benefit human energy production. The Joint Genome Institute (JGI) has recently assembled the genome of *H. mukohataei*, but the automated annotation for this species has not yet been confirmed by more detailed annotation analysis.

|  |  |
| --- | --- |
| Rhodopsins are light-sensitive transmembrane proteins that can initiate signal transduction and energy production within the cell. These receptors are composed of seven alpha helices that bind a vitamin A retinal molecule, which detects light wavelengths to induce a conformational change in the protein (Figure 1). There are four known functional types of rhodopsins. Two types of structurally similar sensory rhodopsins are used for phototaxis, to signal the cell to orient toward or away from a light source. Halorhodopsin (hR) is a light-driven ion pump that transports chloride ions into the cytoplasm, and bacteriorhodopsin (bR) is a light-driven proton pump that transports H+ out of the cell. Both | Macintosh HD:Users:olhoshing:Desktop:GPCR.png  **Figure 1.**Rhodopsins are trans-membrane receptors containing 7 alpha helices arranged around a retinal molecule. Signals are transduced via the cytoplasmic C-terminal (COO-) end. |

pumps function to create an electrochemical potential for ATP production. (Lanyi, 2004). Bacteriorhodopsin orthologs are found in a wide yet patchy distribution among species; primarily in haloarchaea, but also in non-halophilic bacteria and eukaryotes, strongly suggesting lateral gene transfer (Sharma et al, 2007). *H. mukohataei* has been reported to contain retinal proteins homologous to the bR proton pump, but to lack an hR ion pump (Sugiyama et al, 1994). This novel composition of rhodopsins in a haloarchaea suggests potentially different mechanisms of energy production from other bR-containing microbes.

Clustered regularly interspaced short palindromic repeats (CRISPRs) are repetitive DNA structures found in bacteria and archaea that are composed of repeat sequences (direct repeats) 21 to 37 bases long that show weak dyad symmetry. Direct repeats (DRs) separate non-repetitive spacers of similar length, which appear to contain portions of phage DNA. (Haft et al, 2005) Genes belonging to 45 different protein families called CRISPR-associated sequences (CASs) are found within the vicinity of the CRISPR structure. Four of these families, *cas1* through *cas4*, are always associated with CRISPRs and thought to play a critical role in CRISPR generation and a suggested antiviral function. Spacers are thought to confer immunity against phage infection by a mechanism of RNA interference (Makarova et al, 2006). It is thought that spacers are transcribed into small interfering RNA (siRNA), via an AT-rich leader sequence found at the beginning of the CRISPR structure (Figure 2). Upon infection, the siRNA binds to complementary viral DNA, initiating its destruction.

|  |
| --- |
| CRISPRStructure.png  **Figure 2. CRISPR sequence structure.** An AT-rich leader sequence is directly upstream of the first direct repeat (DR). Direct repeats are 21 – 37 bp that insulate unique spacers of similar length. The last direct repeat sometiimes contains polymorphisms; since it is not as strictly conserved at the other DRs, it is called a degenerate direct repeat. |

While studies have analyzed CRISPR-associated proteins and spacers, little attention has been given to the study of the direct repeats. CRISPR sequences are one of the most rapidly evolving structures, so that species with very similar genomes often differ in their CRISPR composition. Previous studies of direct repeat sequences found that the sequences are unique and highly variable between species (Jansen et al, 2002). Phylogenetic analyses indicate that CASs have undergone lateral gene transfer between distantly related organisms. CRISPRs and CASs have also been detected in plasmids and other mobile genetic elements, suggesting possibly extensive gene transfer between prokaryotes. (Godde and Bickerton, 2006) Based on these observations, it is possible that direct repeats in halophiles can indicate lateral gene transfer, or elucidate the role of direct repeats in the CRISPR structure based on observed similarities between species.

In this study, we sought to validate the JGI annotation of rhodopsins and CRISPRs found in *H. mukohataei*, and to compare direct repeat and bacteriorhodopsin sequences to those found in other halophiles. Through species comparison, we wanted to investigate phylogeny and similarities among halophiles, and better characterize the structure within the genomic DNA that might be important for function. We found that comparison of rhodopsin and CRISPR sequences in halophiles does not offer an interpretable mapping of phylogeny, but does suggest functional differences between sequences found within the same species.

**Materials and Method**

*Halophile genomes.* We received the *H. mukohataei* arg-2 genome sequence from the Joint Genome Institute (JGI). JGI sequenced the genome by whole-genome shotgun sequencing, resulting in 2 contigs (http://img.jgi.doe.gov/cgibin/geba/main.cgi?section=TaxonDetail&page=taxonDetail &taxonoid=2500395343). We obtained other halophile genome sequences from San Diego State University (http://edwards.sdsu.edu/halophiles/), for whom Roche Diagnostics sequenced eight halophiles by pyrosequencing. The sequenced halophiles were: *Haloarcula californiae, Haloferax denitrificans, Haloferax mediterranei, Haloferax mucosum, Haloferax volcanii, Haloarcula sinaiiensis*, *Haloferax sulfurifontis*, and *Haloarcula vallismortis*. For each species, we used the Large Contigs sequences from the raw data assembly.

*Rhodopsin analysis.* JGI automated annotation predicted a function for 2135 of the 3399 protein-coding sequences found. I searched the list of coding sequences for predicted rhodopsin gene sequences. I utilized the National Center for Biotechnology Information (NCBI) Conserved Domains database to find bacteriorhodopsin genes in other species. I obtained amino acid sequences from the NCBI Conserved Domains database within the pFam01036 and cl0233 bR families, and from BLASTp searches of genes found in the Conserved Domains. A total of 17 halophilic archaeal rhodopsin genes were found and compared with those in *H. mukohataei*. I used NCBI BLASTp with default parameters to compare amino acid sequences. Finally, I used the European Bioinformatics Institute (EBI) ClustalW tool for multiple sequence alignment and generating phylogenetic trees based on amino acid sequences.

*CRISPR analysis*. We analyzed the direct repeat (DR) sequences, spacer sequences, and CRISPR-associated sequence (CAS) proteins using the finished *H. mukohataei* genome from JGI, and the most finished versions of the eight halophile genomes from San Diego State University. We used the CRISPRFinder web tool to search the *H. mukohataei* genome for CRISPR sequences therein.

I used the consensus direct repeat (DR) sequences given by CRISPRFinder for the direct repeat analysis. I used NCBI BLASTn to search for similar sequences in other organisms, and EBI’s ClustalW for multiple sequence alignment and phylogenetic tree generation based on nucleotide sequences. For halophiles having numerous scaffolds, redundant DR sequences were omitted and treated as a single unit. A total of 31 direct repeat sequences from 12 halophiles were submitted to ClustalW for sequence alignment and phylogram mapping. I used WebLogo (weblogo.berkeley.edu) to generate direct repeat consensus sequence logos. I compared our sequences with those analyzed by Kunin V et al (2007), the sequences and data available online as additional data files.

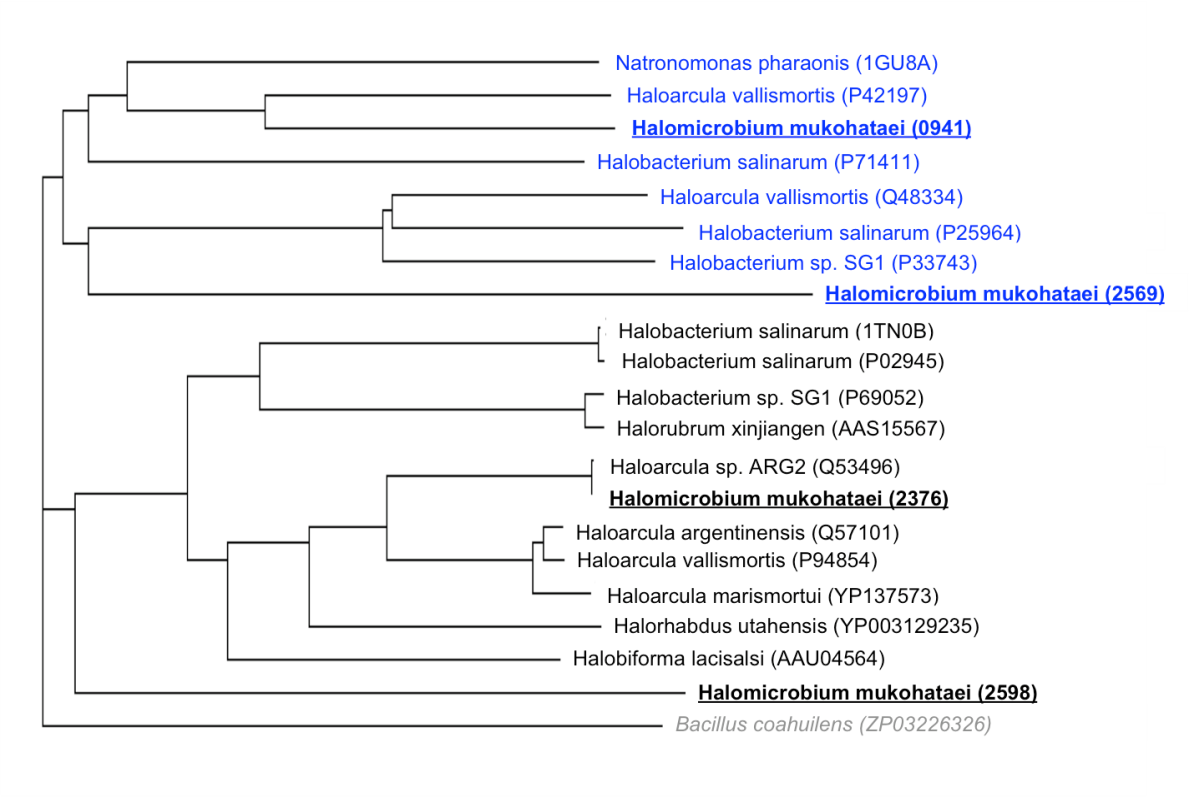
In our analysis of CAS proteins, Sarah found all the CAS proteins that JGI annotated in the *H.* mukohataei genome. Rapid Annotation using Subsystem Technology (RAST) provided the proteomes of the eight halophiles sequenced by Roche Diagnostics and of *H utahensis*. I generated a Perl script to compare each of the proteomes to *H. mukohataei*. Sarah utilized these results to identify which CAS genes are conserved in each species. In our CRISPR spacer analysis, Katie used NCBIBLAST to search for viral sequences similar to *H. mukohataei* spacer sequences.

**Results**

***Rhodopsins*** ***in* H. mukohataei**

JGI automated annotation found 4 coding regions predicted to encode bacteriorhodopsin (bR) in the *H. mukohataei* genome. The 4 sequences predicted are numbered 644030941 (0941; 243 aa), 644032376 (2376; 231 aa), 644032569 (2569; 237 aa), and 644032598 (2598; 273 aa). I compared these four sequences found in *H. mukohataei* to seventeen other halophilic rhodoposin genes (See Materials and Method).

When I used ClustalW to compare *H. mukohataei* bR to the other halophile sequences, the multiple sequence alignment did not cluster together the 4 *H. mukohataei* genes as one might have expected. The resulting tree bisects into two large clusters (Figure 3), each cluster containing two of the *H. mukohataei* bR sequences. In an attempt to characterize the difference between the two clusters, I found that NCBI Database describes the genes contained in the smaller cluster as sensory rhodopsins II. This observation suggests that two of the coding sequences in *H. mukohataei* named bacteriorhodopsin may be more structurally and functionally related to sensory rhodopsins than to bR proton pumps.



**Figure 3. Phylogenetic tree of halophile rhodopsins.** A multiple sequence alignment was performed on amino acid sequences from 16 haloarchaea rhodopsins, one halophilic bacteria rhodopsin (grey), and 4 rhodopsins from *H. mukohataei* (underlined). The two main clusters are separated by colour, the blue sequences denoting sensory rhodopsins, and the black sequences denoting proton pump bacteriorhodopsins. Gene accession numbers are in parentheses.

In order to better determine the specific roles of the predicted *H. mukohataei* rhodopsin, I used BLASTp to compare the sequences to more extensively studied rhodopsins. *Halobacterium salinarum* has been studied as a model species for endogenous bR, hR and sensory rhodopsins (Mukohata et al, 1998). *H. mukohataei* sequences were aligned to bacteriorhodopsin (P02945), halorhodopsin (Q48315) and sensory rhodopsin-2 (P71411) of *H. salinarum* (Table 1). Curiously, *H. mukohataei* sequence 2569 did not align significantly to any of the *H. salinarum* rhodopsin types. Sequence 0941 best aligned with the *H. salinarum* sensory rhodopsin-2. Sequence 2376 best aligned to bacteriorhodopsin, and Sequence 2598 surprisingly aligned best to halorhodopsin, although *H. mukohataei* is reported to be a bR+, hR– species (Sugiyama et al, 1994).

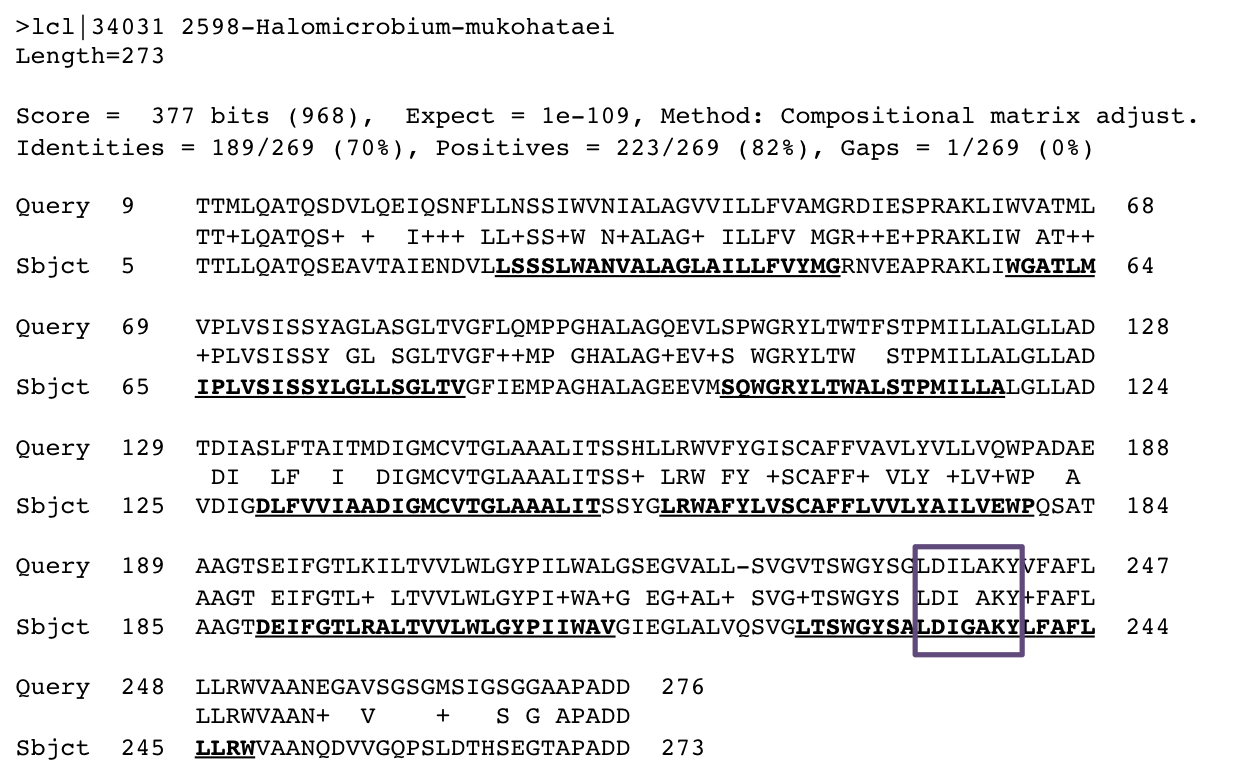
The sequence alignment for Sequence 2598 to halorhodopsin (Figure 4) shows 82% similarity with few gaps and an extremely small E-value (1E–109). Figure 4 highlights the helical stretches that compose the transmembrane portion of the hR protein (Figure 1; Blanck and Oesterhelt, 1987), which between *H. mukohataei* and *H. salinarum* appear to be well conserved. An area of particular conservation in comparing hRs is the retinal-binding region (Amino acids 236 – 242 in *H. salinarum*), which consists of three conserved amino acids, followed by a possible mismatch, and then three more conserved amino acids including the lysine needed to bind the retinal molecule. (Blanck and Oesterhelt, 1987)

Sugiyama and colleagues named the bR that they studied in *H. mukohataei* cruxrhodopsin-2 (Sugiyama et al, 1994). In order to identify which coding region was designated cruxrhodopsin in *H. mukohataei*, the sequences were then aligned using BLASTp to cruxrhodopsin-2 (S76743). *H. mukohataei* sequence 2376 aligned with 100% identity. Sequences 0941 and 2598 had 52% and 49% positives. However, sequence 2569 again did not align significantly.

**Table 1. BLASTp Alignments of *H. mukohataei* sequences to *H. salinarum* genes. †**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***H. mukohataei* Sequence No.** | ***H. salinarum* Bacteriorhodopsin** | | ***H. salinarum* Halorhodopsin** | | ***H. salinarum* Sensory Rhodopsin-2** | |
|  | **Bit Score** | **E-value** | **Bit Score** | **E-value** | **Bit Score** | **E-value** |
| 0941 | 102 | 6E–27 | 64.3 | 3E–15 | 134 | 2E–36 |
| 2376 | 189 | 7E–53 | 74.7 | 2E–18 | 73.2 | 4E–18 |
| 2569 | 15 | 2.0 | 16.5 | 0.63 | 21.9 | 0.015 |
| 2598 | 111 | 1E–29 | 377 | 1E–109 | 65.9 | 8E–16 |

**†** Underlined values highlight the lowest significant (< 0.01) E-value hit for each *H. mukohataei* sequence.



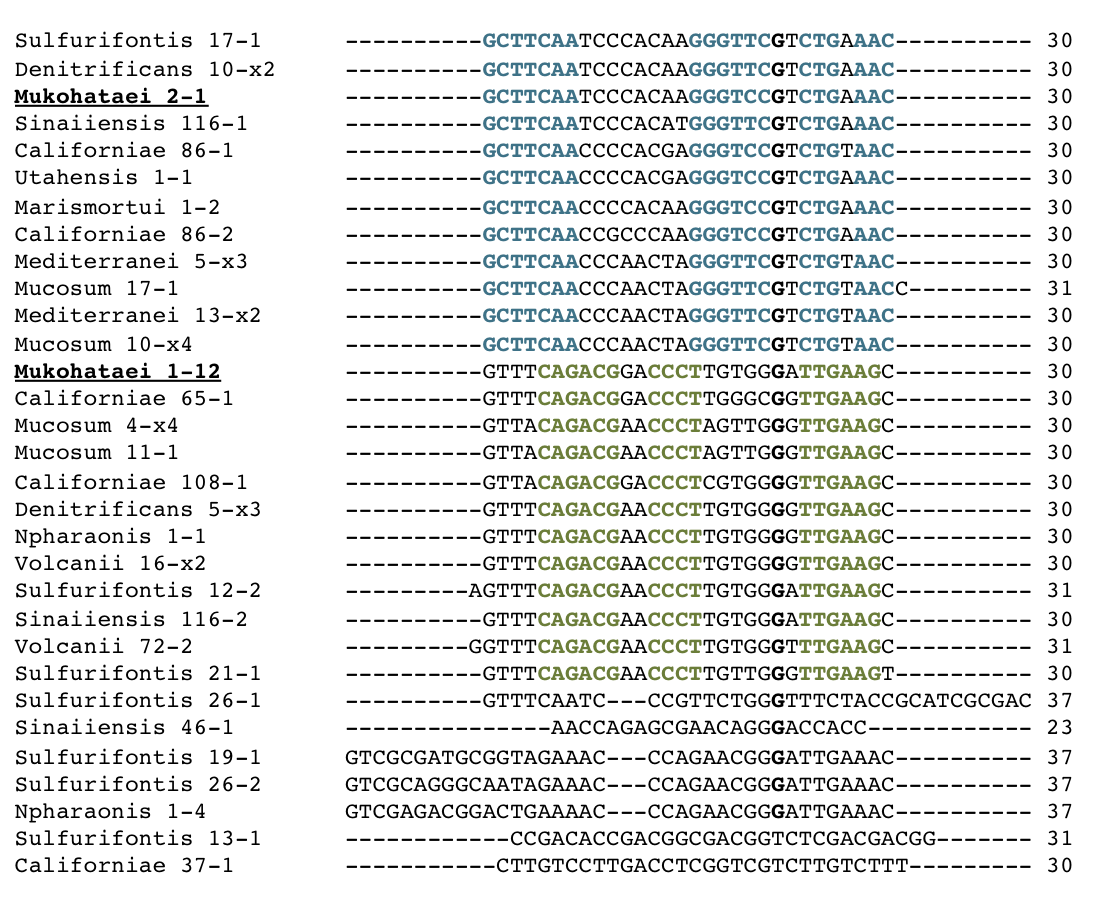
**Figure 4.** **BLASTp alignment of *H. mukohataei* Sequence 2598 to *H. salinarum* halorhodopsin.** *H. mukohataei* (lower sequence) aligned with 82% identity to *H. salinarum* halorhodopsin sequence (upper sequence). Amino acids composing the seven helical regions of hR are underlined and bolded. The retinal binding region in halorhodopsin is boxed.

***CRISPR Direct Repeats in Halophiles***

JGI automated annotation detected two CRISPR sequences, one found on the first DNA scaffold and the other on the second of the unfinished genome sequence. When both scaffolds of the genome were submitted to CRISPRFinder, two confirmed CRISPR sequences were also found, one on each scaffold. In CRISPRFinder, the two CRISPRs on the first and second scaffolds are reported to be 2,197 and 2,929 bp, respectively. These CRISPR lengths do not match the lengths that JGI has currently annotated (2134 and 2667 bp), this discrepancy probably due to the differences between the finished and unfinished genomes. The JGI positions for CRISPRs correspond with the unfinished genomes, not the finished sequence we used. Nonetheless, the two sequences that CRISPRFinder found in the finished genome exhibit the defined structure of CRISPRs (Figure 5), so we used these sequences for our CRISPR analysis. The direct repeats (DRs) in both CRISPRs are 30 bp in length, separated by spacers of similar length. The CRISPR on the first and second scaffold have 33 and 44 spacers, respectively.

I submitted the two *H. mukohataei* consensus direct repeat sequences from each CRISPR to BLAST in order to see if similar sequences were found in other species. For both DRs, I found three species with hits of significant E-values (< 0.01): *Haloarcula marismortui* (E-value: 7E–07), *Halorhadbus utahensis* (1E–05), and *Natronomonas pharaonis* (2E–04). I obtained the genome sequences for these three halophiles and used CRISPRFinder to find these confirmed DR sequences also. When we submitted the genomes of the eight other halophiles (See Materials and Methods) to CRISPRFinder, it did not find any CRISPRs in *Haloarcula vallismortis*. Neither did we find any of the four CRISPR-associated (CAS) protein families that are always found in conjunction with CRISPR sequences. In all other halophiles surveyed, we found *cas1, cas2, cas3* and *cas4* genes adjacent to the CRISPR sequence.

Although direct repeats have been thought to be unique and highly variable between different species (Jansen et al, 2002), the ClustalW alignment of the halophile DRs surprisingly showed high similarity between the surveyed halophiles (Figure 5). The alignment can be divided into three main clusters based on conserved regions of nucleotides, as indicated by the blue and green clusters in Figure 5. The third cluster encompasses primarily DRs from *Haloarcula* *sulfurifontis*, indicating *H. sulfurifontis* has more unique CRISPR sequences than the other species, but is still similar enough to align with some identity to other halophiles.



**Figure 5. ClustalW multiple sequence alignment of halophile direct repeat sequences.** Consensus direct repeats were found using CRISPRFinder for *H. mukohataei* and 11 other halophiles. Direct repeats are named by the trivial species name followed by the scaffold number on which the CRISPR was found, and finally the arbitrary number assigned by CRISPRFinder. Redundant direct repeats have the copy number (such as x2) in place of the arbitrary number. Nucleotides strictly conserved within a cluster of the alignment are coloured in blue for the top cluster and green for the middle cluster. The length of each direct repeat is printed to the right of the aligned sequence.

In order to determine if the direct repeats found in halophiles could indicate phylogeny, I created a phylogenetic tree based on the direct repeat sequences. This tree was compared to a phylogenetic tree based on 16S rRNA sequences found in each halophile, because 16S rRNA sequence comparison is a widely used method of inferring phylogeny (Olsen et al, 1986). The resulting DR phylogram (Figure 6a) created three large clusters, which did not correspond directly with the three clusters from the multiple sequence alignment (Figure 5). Direct repeats were not clustered together by species, as one would expect for a tree in order to indicate phylogeny. The 16S rRNA sequences however are clustered together by species (Figure 6b), offering a more interpretable map of phylogenetic relationships.

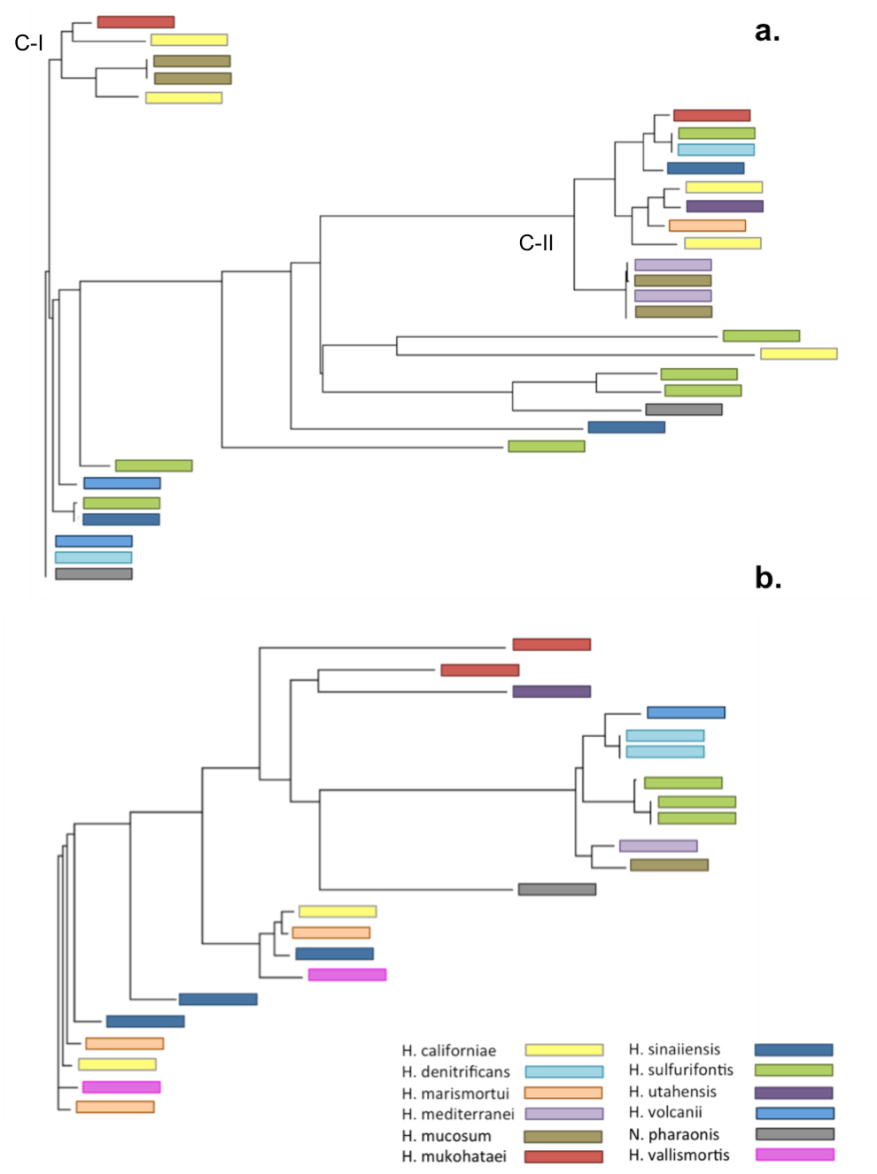
This extensive mixing of species might suggest widespread lateral gene transfer, or a conservation of function for direct repeats. Although the entire geographic range of these species has not been reported, geographic mapping of the point of species isolation might show physical proximity of similar species. However, the locations of isolation of the archaea analyzed here cover a wide geographical area that does not strongly suggest lateral gene transfer between halophiles whose direct repeats are very similar.

When the spacers were analyzed using BLASTn for halophiles with similar DRs, we found that spacers did not significantly match between CRISPRs in different species. CRISPRs with similar DRs varied widely in length (500 – 2500 bp) between species, and primarily direct repeats produced short significant BLAST alignments. For *H. mukohataei* CRISPRs, we did find three significant matches to three viral hits. Portions of these three viral genomes were found in other halophiles surveyed, but these portions were not within the CRISPR sequences of other halophiles. Altogether, the direct repeat sequences of halophiles showed strong similarities with each other, but the unique spacers between direct repeats are not strongly conserved between species, and random segments of phage DNA appear to be inserted into random portions of the host genome.

To explore a possible function based on the DNA structure of direct repeats, I used both ClustalW and Weblogo sequence logo generator to find the consensus direct repeat sequence for the 12 halophiles surveyed (Figure 7). The consensus sequence shows a high level of dyad symmetry, which enables transcribed RNA elements to fold onto itself as a part of primarily RNA interference mechanisms in the cell. Particularly interesting about the dyad symmetry of the direct repeats is the recurrence of triplet nucleotides, such as TTT or GGG. Since other prokaryotic sequences having dyad symmetry do not mimic this sort of triplet motif (König et al, 1989), this characteristic does not appear to be a normal artifact found in dyad-symmetrical regions. When non-halophilic bacterial genomes were also search for CRISPRs and aligned with ClustalW, I found that this triplet motif is recurrent even in more distant species as well like non-halophilic bacteria (Figure 8). These data could suggest that there is a functional role for direct repeats in prokaryotes dependent on the secondary structure of the RNA or binding domains in the DNA sequence.

A recent study investigating direct repeats in bacterial and archaeal CRISPRs predicted the secondary RNA structure of DRs in 195 species (Kunin et al, 2007). Kunin and colleagues defined 12 separate clusters of direct repeat sequences based on their sequence similarity and predicted secondary structure. Six of the twelve clusters defined form very stable stem-loop RNA structure. The other clusters have sequences with dyad symmetry, but do not appear to form stable stem-loop structures. In order to see if the direct repeat phylogram (Figure 6a) may be indicative of the functional RNA structure instead of indicating phylogeny, I compared the consensus sequences of two clusters of the phylogram that contain *H. mukohataei* DRs to the consensus sequences for the 12 DR clusters described by Kunin.

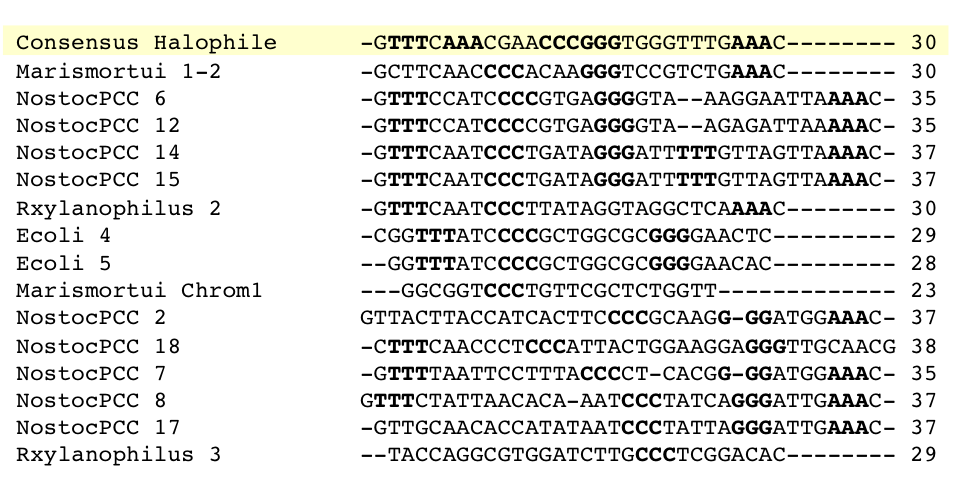
I found that clusters I and II (C-I and C-II, Figure 6a) have similar consensus sequences to three clusters defined by Kunin (Figure 9). The C-I halophile cluster had 53% identity with Kunin cluster 6, a cluster of unfolded archaeal DRs. Cluster C-II had 56% identity with Kunin clusters 3 and 8, both described as folded bacterial DRs (Kunin et al, 2007). These two latter clusters both have predicted stable secondary-structures of RNA, suggesting at least some of the halophile direct repeats including those found in *H. mukohataei* may form stem-loop structures in the cell. In the C-II alignment, the areas with triplet motifs are particularly conserved (Figure 9b), which could indicate these triplet motifs are necessary for the stem-loop structure. Other DRs such as those in cluster C-I show weaker dyad symmetry and fewer triplet motifs (Figure 9a). These DRs may not be capable of forming a stable stem-loop structure, but may function by some other cellular mechanism or structure. Moreover, these observations suggest that the clustering of the phylogram for direct repeats, like the rhodopsin phylogram (Figure 3), could be a good indicator of the function of the genetic element.



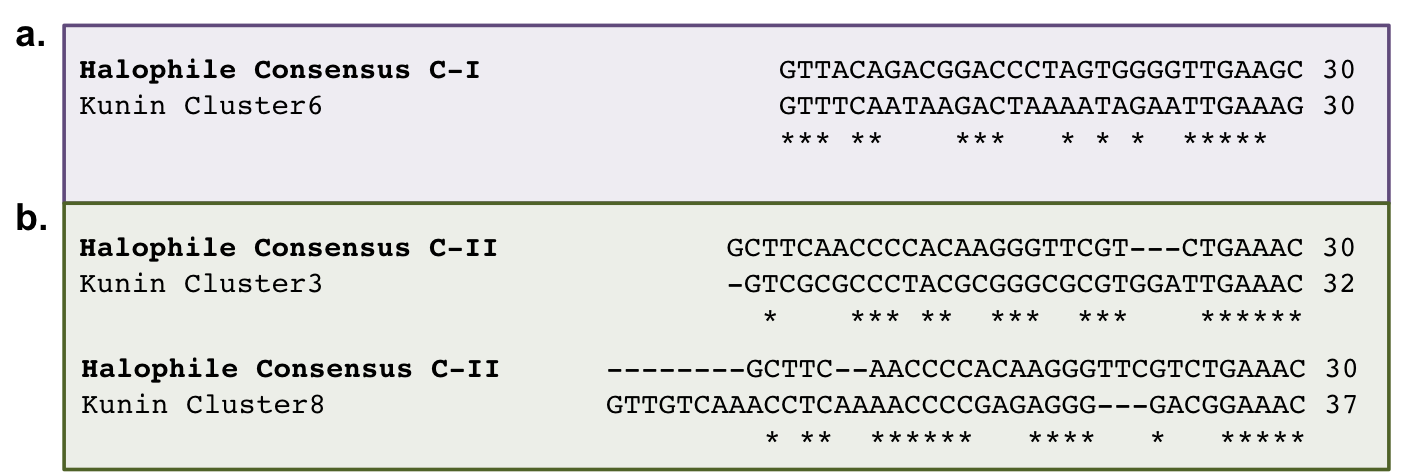
**Figure 6. Phylogenetic trees of halophile direct repeats and 16S rRNA.** (a.) Phylogram based on 31 consensus direct repeats obtained by CRISPRFinder, and (b.) is based on 16S rRNA sequences (1100 – 1300 bp) for each species found on the NCBI Entrez database. Each branch is represented by the species, colour-coded according to the bottom legend.



**Figure 7. Sequence logo of consensus halophile direct repeats.** Consensus sequence and logo were generated from the 31 direct repeat sequences from the 12 surveyed halophiles including *H. mukohataei*. The consensus sequence was generated by ClustalW and validated with the corresponding sequence logo. Occurrences of the triplet nucleotide motif are boxed in the corresponding nucleotide colour.



**Figure 8. ClustalW multiple sequence alignment of non-halophilic bacterial CRISPR direct repeats.** The consensus halophile sequence (yellow) was aligned with direct repeats from one halophile (*H. marismortui)* and 3 bacteria: *Escherichia coli, Nostoc* sp. PCC 7120*,* and *Rubrobacter xylanophilus*. Occurrences of the triplet nucleotide motif are bold.



**Figure 9. ClustalW alignment of halophile cluster with Kunin cluster consensus sequences.** Consensus sequences for halophile clusters I and II (C-I and C-II) from the direct repeat phylogenetic tree were aligned with the consensus sequences for all clusters from Kunin et al 2007. Individual pair-wise alignments with the highest scores are shown; asterisks indicate a match at the aligned base. (a) Cluster C-I had the highest pair-wise alignment with Kunin cluster 6, and (b) Cluster C-II had the highest pair-wise alignment with Kunin clusters 3 and 8.

**Discussion**

A phylogram of rhodopsin genes indicated that archaeal rhodopsins found in haloarchaea may be more functionally similar than they are species-specific. Similarly, the phylogram of direct repeats may also suggest functional similarities by the predicted folding structure of transcribed RNA. This would explain the lack of clustering of rhodopsins and direct repeats found within the same species, particularly *H. mukohataei* sequences. Alternatively, *H. mukohataei* sequences may be distributed throughout the phylogenetic clusters due to lateral gene transfer, if its genome received some rhodopsins or CRISPR sequences from other microbes. Although we found no direct support for lateral gene transfer, there is evidence for gene transfer based on methods used in previous studies of rhodopsins (Sharma et al, 2007) and CRISPR sequences (Godde and Bickerton, 2006).

The four rhodopsin genes annotated by JGI may have very different functional roles. Phylogenetic trees and BLAST alignments done in this study suggest that at least one of the rhodopsin sequences, Sequence 0941, has a high probability to be a sensory rhodopsin. While the role of archaeal rhodopsins as energy producing proton pumps is a critical part of the microbial energy production, it is logical that *H. mukohataei* also requires sensory rhodopsin for phototaxis to orient the cell toward light sources. JGI also failed to specifically name the cruxrhodopsin protein – Sequence 2376 –, a bacteriorhodopsin analogue unique to *H. mukohataei* genome (Sugiyama et al 1994). This example with cruxrhodopsin illustrates how automated annotation can sometimes inaccurately or vaguely designate gene functions due to the inherent limitations of completely automated analysis. Predicted coding sequences should be analyzed in more detail and compared with other automated genome annotations to further validate genetic elements.

Another interesting finding is rhodopsin gene Sequence 2598, which shows very strong homology to the halorhodopsin ion pump. Sugiyama and colleagues previously reported that *H. mukohataei*, or *Haloarcula* sp. arg-2, completely lacked any halorhodopsin-like proteins (Sugiyama et al, 1994). They validated this conclusion by Southern hybridization using hR DNA fragments as probes, and by polymerase chain reaction (PCR). Based on the genomic analysis done here, however, *H. mukohataei* does appear to have a halorhodopsin gene. Although bacteriorhodopsin and halorhodopsin may have similar tertiary structure, they have only 36% similarity in their primary sequence structure (Blanck and Oesterhelt, 1987), so Sequence 2598 is unlikely to be merely a mutated bacteriorhodopsin gene. Sugiyama’s previous study may have overlooked the halorhodopsin because of their methods; genomic annotation studied here may be a more thorough method of species analysis. If *H. mukohataei* does indeed have an hR-like protein, it may *in vivo* no longer be functional in the cell, which would essentially give *H. mukohataei* a bR+, hR– phenotype. Future laboratory research can characterize the function of these rhodopsin genes in *H. mukohataei*.

In our investigation of *H. mukohataei* CRISPR sequences, we found that the current annotated position for the two CRISPRs on JGI are not up to date with the finished genome sequence that is available for public access. Since JGI does not provide the user with the CRISPR sequence separate from the entire genome sequence, this outdated position listing was an obstacle in validating the JGI annotation for CRISPRs. Nonetheless, both JGI and CRISPRFinder found two large CRISPR structures in *H. mukohataei*.

Although the mechanism of immunity is not fully understood, studies have shown that CRISPRs do provide antiviral immunity to microbial hosts. Barrangou and colleagues found that when bacteria are infected with phage DNA, some DNA is incorporated into the bacterial genome. When these sequences were removed from the bacterial genome, phage resistance weakened. (Barrangou et al, 2007) It was proposed that CRISPRs might accomplish this mechanism of “adaptive” immunity by a system in which the CRISPR is transcribed from the promoter-like leader sequence (Figure 2). Unique spacers are spliced from the RNA by a prokaryotic analogue of the enzyme Dicer to make small interfering RNA (siRNA) or microRNA (Makarova et al, 2006). Small interfering RNA is around 20 bp of double-stranded RNA with a two to three-bp overhang on each end. MicroRNA is around 20 bp of single-stranded RNA, formed when Dicer processes a 70-bp stem-loop structure. Both RNAs function by base-pairing with their complement mRNA in order to down-regulate gene expression. (Lee et al, 2002) *Cas3* encodes a helicase that is thought to be the prokaryotic analogue of the Dicer enzyme (Makarova et al, 2006) supporting the idea that CRISPRs are a prokaryotic equivalent of RNA interference.

Kunin and colleagues predicted the secondary RNA structures for direct repeats from 195 microbes. They clustered these DRs by sequence and structure similarity into 12 groups. When I compared the DR sequences clustered with *H. mukohataei* DRs in the ClustalW alignment (Figure 6) to the sequences of these 12 clusters, I found that the halophile clusters aligned with some similarity to the Kunin clusters. Clusters I and II (Figure 6a) had 53% and 56% identity to a cluster of unfolded archaeal DRs and folded bacterial DRs, respectively. Although these identity scores might seem too low to be significant for a gene sequence, direct repeat sequences analyzed were around 30 bp of non-coding DNA. This is significantly shorter than a normal coding gene, dramatically influencing the scoring if there are gaps or mismatches. Moreover, the consensus sequences used are representative but may not be completely equivalent to the alignments of each individual sequence.

The triplet motifs in the C-II cluster sequence aligned with folded bacterial clusters, suggesting that these areas of the direct repeat sequence may be necessary for appropriate folding of the stem-loop secondary RNA structure. For other halophile direct repeats like those found in C-I may be non-folding RNAs similar to eukaryotic siRNA. It would be interesting to compare these direct repeat sequences with microRNA and siRNA sequences from eukaryotic genomes to see if this potential prokaryotic RNA is indeed structurally similar to a known mechanism of RNA interference.

In the examination of CRISPR spacers, we found that of the 77 spacers in *H. mukohataei* CRISPRs, three spacers had significant similarity to viral genomes within the NCBI database. Previous studies of spacers observed that only about 10% of unique spacers in CRISPR units are homologous to viral fragments or plasmid genomes (Makarova et al, 2006). Moreover, the current database of viral genomes is only a fraction of the total diversity of phage and other mobile elements. Therefore, with our current knowledge of viral genomes, we can expect to find very few significant viral matches within a single species, as we observed. Since CRISPRs do not code any known protein and probably function as an adaptive immune mechanism against viruses, it would make sense that random mutations within the unique spacers could be tolerated and even facilitated by the cell in order to increase the diversity of mobile elements CRISPRs can resist upon infection. This would be interesting to address in future studies of archaeal phage resistance in the presence of CRISPRs.

CRISPR sequences and rhodopsins have become interesting focal points of study with many questions still unanswered. *In silico* investigations into the theoretical secondary RNA structure of CRISPRs provide feasible hypotheses that can be tested with further laboratory research. A better understanding of the mechanisms of CRISPR structures will give valuable insight into microbial mechanisms of viral immunity and interference RNA. A greater understanding of the origins and structural adaptations of bacteriorhodopsin may lead to significant advancements in future renewable energy sources.

**References**

Blanck A and Oesterhelt D. (1987). The halo-opsin gene. II. Sequence, primary structure of halorhodopsin and comparison with bacteriorhodopsin. *EMBO J* 6:265–73.

Godde JS, Bickerton A. (2006). The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. *J Mol Evol* 62:718–29.

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.

Ihara K, Watanabe S, Tamura T. (1997). Haloarcula argentinensis sp. nov. and Haloarcula mukohataei sp. nov., two new extremely halophilic archaea collected in Argentina. *Int J Syst Bacteriol* 47:73–77.

Jansen R, Embden JD, Gaastra W, Schouls LM. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol* 43: 1565–75.

König H, Ponta H, Rahmsdorf U, Büscher M, Schönthal A, Rahmsdorf HJ, Herrlich P. (1989). Auto-regulation of fos: the dyad symmetry element as the major target of repression. *EMBO J* 8:2559 – 66.

Kunin V, Sorek R, Hugenholtz P. (2007). Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol* 8:R61.

Lanyi JK. (2004). Bacteriorhodopsin. *Annu Rev Physiol* 66:665–88.

Lee Y, Jeon K, Lee JT, Kim S, Kim VN. (2002). MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J* 21:4663–70.

Makarova KS, Grishin NV, Shabalina SA, Wold 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* 1:7.

Mukohata Y, Ihara K, Tamura T, Sugiyama Y. (1998). Halobacterial rhodopsins. *J Biochem* 125:649–57.

Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA. (1986). Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol* 40:337–65.

Sharma AK, Walsh DA, Bapteste E, Rodriguez-Valera F, Ford Doolittle W, Papke RT.(2007). Evolution of rhodopsin ion pumps in haloarchaea. *BMC Evol Biol* 7:79.

Sugiyama Y, Yamada N, Mukohata Y. (1994). The light-driven proton pump, cruxrhodopsin-2 in *Haloarcula* sp. arg-2 (bR+, hR–), and its coupled ATP formation. *Biochim Biophys Acta* 1188:287–92.

**Acknowledgements**

I would like to thank Dr. AM Campbell and our entire Genomics Lab Methods class for their collaboration and suggestions. This genome annotation project is facilitated by the Joint Genome Institute and the US Department of Energy.