

2019

M. ruber Mrub_3013 is Orthologous to *E. coli* b2755


Laura Butcher

Augustana College, Rock Island Illinois

Dr. Lori Scott

Augustana College, Rock Island Illinois

Follow this and additional works at: <https://digitalcommons.augustana.edu/biolmruber>

 Part of the [Bioinformatics Commons](#), [Biology Commons](#), [Computational Biology Commons](#), [Genomics Commons](#), and the [Molecular Genetics Commons](#)

Augustana Digital Commons Citation

Butcher, Laura and Scott, Dr. Lori. "M. ruber Mrub_3013 is Orthologous to E. coli b2755" (2019). *Meiothermus ruber Genome Analysis Project*.

<https://digitalcommons.augustana.edu/biolmruber/48>

This Student Paper is brought to you for free and open access by the Biology at Augustana Digital Commons. It has been accepted for inclusion in Meiothermus ruber Genome Analysis Project by an authorized administrator of Augustana Digital Commons. For more information, please contact digitalcommons@augustana.edu.

***M. ruber* Mrub_3013 is Orthologous to *E. coli* b2755**

Introduction

What is the *M. Ruber*?

M. ruber stands for *Meiothermus ruber*, a pink, thermophilic bacteria, thriving in warmer temperatures. Its optimal temperature is 60°C and contains a bright red intracellular carotenoid pigment. It is part of the Bacteria domain and Deinococcus - Thermus phylum (Tindall *et al.*, 2010). *M. ruber* was selected for sequencing by the Genomic Encyclopedia of Bacteria and Archaea project (Hugenholtz *et al.*, 2009), but the actual sequencing was performed by DOE Joint Genome Institute (JGI) and deposited in Gen-Bank. The majority of the protein coding genes were given a putative function but others were left as hypothetical genes with 3,052 protein-coding genes of 3,105 genes predicted (Tindall *et al.*, 2010).

What is the *M. Ruber* genome analysis project?

There is very little known about *M. ruber* because it does not cause disease. The purpose of this project is to begin to determine if our assumptions about bacteria, like *E. coli*, are consistent with *M. ruber*. This project studies one gene at a time, each *M. ruber* gene is compared to other bacteria, mostly *E. coli*. *E. coli* is used as the model organism because it is very well studied with a vast amount of online resources for its genome. Most of what we know about bacteria are based off of *E. coli* making it an important part of the project purpose. Earlier parts of this project looked at the ProC gene in *E. coli* and *M. ruber* comparing its function in both species, but now the project turns to look at the CRISPR-Cas system in *M. ruber* as compared to *E. coli*. Preliminary research using bioinformatics tools shows the *M. ruber* does have a CRISPR-Cas system allowing us to ask more specific questions about Cas proteins. These results will be discussed in the results of the paper. More specifically, this research focuses on the Mrub_3013 gene in *M. ruber* to determine if it is an ortholog with Cas1, locus tag b2755, gene in *E. coli*.

What do we know about the CRISPR-Cas system?

The CRISPR-Cas system is one of the ways bacteria uses its genome flexibility to fight against bacteriophages and plasmids by a sort of inventory of previous infections. CRISPR-Cas sequences first have a leader sequence for recognition, CRISPR array sequence, and genes that code for CRISPR associated proteins (Cas). In between the CRISPR array sequences there are spacer sequences which are the inventory sequences of previous bacteriophages and plasmids. There are several types of CRISPR-Cas systems that each have a different set of Cas proteins involved. (Darmon and Leach, 2014).

The three steps in the CRISPR-Cas process of defense are adaptation, CRISPR RNA (crRNA) expression, and interference. During the adaptation step, Cas proteins help recognize protospacers within the plasmid or bacteriophage and adds it to the CRISPR array. crRNA expression is when the spacer of the CRISPR array gets transcribed and eventually forms crRNA. In the final step the crRNA is able to guide Cas proteins in a complex to the protospacer of the invading DNA and inactivate the phage DNA by silencing or degradation (Darmon and Leach, 2014).

It is important to note that the CRISPR-Cas system is an operon so they are transcribed together. Operons are highly conserved throughout evolutionary history and are important for the evolutionary selection of this system (Nunez *et al.*, 2012). This why you can see conservation of this system as an operon throughout different species. Under this assumption we use the presence of an operon and what is known about the *E. coli* CRISPR-Cas system to compare to *M. ruber*. The CRISPR-Cas systems are divided up into two classes 1 and 2 each with different subtypes (Wright *et al.*, 2016). *E. coli* is a class 1, type 1E CRISPR-Cas system with an identifying protein, Cas3 followed by five other proteins that make up the Cascade complex and Cas1 and Cas2 involved in spacer acquisition through the formation of a heterocomplex. These nine Cas sequences are followed by a leader sequence and the CRISPR array (Jiang & Doudna, 2015).

What do we know about Cas1 in *E. coli*?

Cas1 has been crystalized in *E. coli* bound to Cas2 as they are bound to DNA (Wang *et al.*, 2015). Both Cas1 and Cas2 are capable of cleaving various types of DNA. Cas 1 contains a pair of dimers, Cas1a and Cas1b, that are on either side of one Cas2 dimer. The complex undergoes a conformational change upon binding with the protospacer. The structure of the complex is what allows for strict length requirements for acquiring new spacers in the CRISPR array. The asymmetry and different conformations of Cas1a and Cas1b indicate the two dimers are likely to have different functions. Each of the asymmetrical Cas1 homodimers possesses one catalytic subunit, Cas 1a and Cas 1a' generate the 3'-OH group following cleavage to recognize the complementary sequence in protospacer selection. Cas1b and Cas1b' are responsible for forming the Cas1-Cas2 complex. (Wang *et al.*, 2015)

There are different types of protospacer selection shown in Figure 1, naive and primed adaptation in *E. coli* but it is proposed that the structural understanding of Cas1-Cas2 is suitable for both types. Cas1 selects and cuts the foreign DNA to make the spacer using the Cas1-Cas2 complex to restrict the size for recognition of an appropriate protospacer. (Wang *et al.*, 2015)

Figure 1. (A) There are at least three mechanisms of protospacer acquisition at this point. In the Type 1 primed system the Cascade complex binds to a partially matched target with the crRNA

direction and Cas3 is recruited to and moves to the target site for protospacer selection. This target site is selected as protospacer and Cas1-Cas2 is somehow involved in the integration of this protospacer into the CRISPR array. Type I naive adaptation nuclease/helicase RecBCD degrades some of the invader DNA as substrates for Cas1-Cas2 and through an unknown process becomes integrated as a double stranded protospacer. Type II uses Cas9 to recognize PAM sites and recruit Cas1-Cas2 to acquire flanking sequence, but this does not happen in *E. coli* only systems with the Cas9 system. (B) This shows how Cas1-Cas2 act as an integrase to insert protospacers into the CRISPR array as new spacers. Through a couple of transesterification reactions the complex recognizes a leader repeat and leads to a gapped product that can be repaired for successful duplication of the first repeat (Wright *et al.*, 2016).

Purpose

The purpose of this project is to determine if Mrub_3013 is an ortholog for Cas1 in *E. coli*, locus tag, b2755 using bioinformatics tools.

Materials and Methods

I first used a bioinformatics site called Ecocyc (Keseler *et al.*, 2013) to learn more about the CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats-CRISPR associated proteins) system in my model system, *Escherichia coli* K12 MG16655, *E. coli*. This site is strictly a database for information about *E. coli* and the pathways in *E. coli*. I used this site to identify my gene of interest in the CRISPR-Cas system as *cas1* and learn more about *Cas1* in *E. coli*. Next I used another online bioinformatics tool called Kyoto Encyclopedia of Genes and Genomes, KEGG (Kanehisa *et al.*, 2000), a database resource to understand the functions and utilities of biological systems as they function in genes and genomes. I used KEGG to relate CRISPR-Cas systems across species of *E. coli* and *M. ruber* and find *cas1* genes that are possibly orthologous between the two species. *M. ruber*, Mrub_3013 was chosen as the gene of interest and is directly compared to *E. coli* b2755 in this paper to see if they are orthologous. I next used GenBank, a bioinformatics tool that contains the complete genome sequence of *M. ruber*, and Ecocyc (Keseler *et al.*, 2013) for *E. coli* gene to acquire the names, amino acid and nucleotide sequences, chromosome position, and locus tag of the suspected *Cas1* orthologous genes in each species. IMG/M (Integrated Microbial Genomes and Microbiomes), a bioinformatics tool that supports the annotation, analysis, and distribution of microbial genomes and microbiomes, was useful in finding coordinates of Mrub_3013 and visualize the chromosomal region of this gene of interest. The Basic Local Alignment Search Tool, BLAST, (Madden, 2003) was used to compare sequences of similar amino acid sequences to Mrub_3013 in both pair-wise alignment and a multiple sequence alignment. This allowed for finding conservation of a sequence in several different species and determining if the correct start codon was identified. The next step was to use a bioinformatic tool called BLAST again to obtain amino acid sequences and to compare two

sequences for similarities. This allowed the direct comparison of b2755 and Mrub_3013 amino acid sequences.

Three bioinformatics tools were used to predict the functional location of the proteins inside the cell. Transmembrane Helices; Hidden Markov Model or TMHMM (Krogh & Rapacki, 2016) predicts the presence of membrane-embedded transmembrane helices. Prediction of Transmembrane Beta-Barrels, PRED-TMBB (Bagos *et al.*, Beta-barrel), is a bioinformatic tool that identifies membrane-embedded beta-barrels within a protein. Like TMHMM, this information helps understand the location, structure and function of gene products. PSORT-B (Yu *et al.*, 2010) calculates five localization scores for a protein, which include: cytoplasm, cytoplasmic membrane, cell wall and extracellular space.

In the next phase of the project, four databases were used to predict the function of each protein by matching it to a domain consensus sequence or a family/superfamily consensus sequence. CDD, Conserved Domain Database (Marchler-Bauer *et al.*, 2016), determines domains within the protein of both *E. coli* Cas1 and Mrub_3013. TIGRFAM is another bioinformatics tool that is a collection of manually curated protein families designed to support manual and automated genome annotation. TIGRFAM (Haft *et al.*, 2001) allowed for identification of protein families associated with b2755 and Mrub_3013. The next bioinformatics tool I used was Pfam, Protein Families, (Finn *et al.*, 2016), which is a collection of protein domains and families. PDB, the Protein DataBank (Berman *et al.*, 2016), is a small but highly curated database of protein 3D structures.

The last set of bioinformatics tools were used to determine if Mrub_3013 is a component of an operon in the same way b2755 in *E. coli* is part of the CRISPR-Cas operon system. IMG/M, Integrate Microbial Genomes and Microbiomes (Markowitz *et al.*, 2012) Chromosome map and Gene Neighborhood were used to visualize the gene of interest and show flanking genes. It is also used to identify gene maps of the same region, CRISPR-Cas region of the genome in different species for comparison of the operon. And lastly, we returned to a protein BLAST using Mrub_3013 as the query against the *M. ruber* DSM1279 genome to check for paralogs, which is duplicate gene of the same or similar function.

Results

The comparison of protein name and information using Ecocyc show *E. coli* can be used as a model system for the CRISPR-Cas system when learning about *M. ruber*. The locus tag, b2755, is the *E. coli* CRISPR-Cas protein, Cas1. It is located in the cytosol, 305 amino acids long and part of an operon of Cas proteins. Figure 1 shows the order of b2755 in the operon and general function.

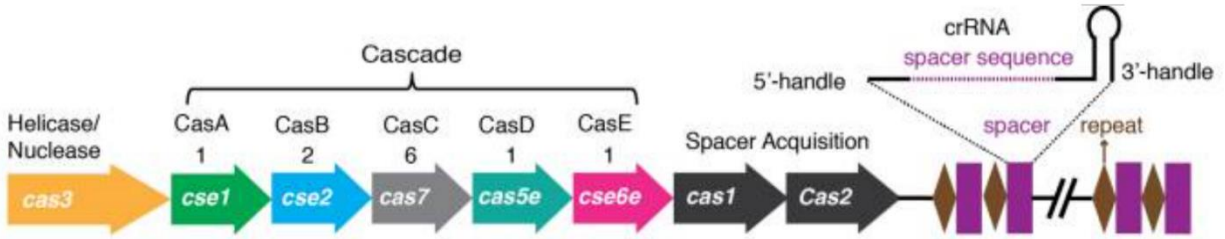


Figure 1. This image shows the *E. coli* CRISPR-Cas operon. *cas1*, b2755, is the gene of interest and it is part of an operon within *E. coli* of several other Cas proteins. This shows that *cas1* and *cas2* are part of gene acquisition. (Jiang&Doudna, 2015)

Three genes were identified as CRISPR-associated protein Cas1 in *M. ruber* as possible orthologs to b2755; all are located in the expected region of the suspected *M. ruber* CRISPR-Cas operon. Although the other two genes were options, Mrub_3013 is the focus of this research. The DNA coordinates of Mrub_3013 on the chromosome are found at 3,053,978-3,054,940 on the *M. ruber* genome with similar surrounding genes to b2755 which will be discussed further in the paper.

We propose that the correct start codon has been called for Mrub_3013. Figure 2 shows an NCBI BLAST multiple sequence alignment using the Cas1 amino acid sequence as the query, where the amino terminus of 14 different species are aligned. Six of the sequences start with the sequence MKY and even more of them have similar number of amino acids that line up fairly similarly. In addition, there is no other M (methionine) near the amino terminus that could be an alternative start codon. Although not everything is conserved throughout the species, this shows that the start codon called by automated annotation for Mrub_3013 is likely the correct start codon. This allows for further research to be pursued.

MULTISPECIES: type I-E CRISPR-associated endonuclease Cas1 [Meiothermus]	MK	YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVAYYSQEGVVAIPAAALGVLMLGPGTSTHAAIRQLANINGCS	76
type I-E CRISPR-associated endonuclease Cas1 [Meiothermus taiwanensis]	MK	YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVAYYSQEGVVAIPAAALGVLMLGPGTSTHAAIRQLANINGCS	76
type I-E CRISPR-associated endonuclease Cas1 [Meiothermus cateniformans]	MR	YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVAYYSQEGVVAIPAAALGVLMLGPGTSTHAAIRQLANINGCS	76
CRISPR-associated protein Cas1 [Meiothermus silvanus DSM 9946]	MA[17]	YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVAYYSQEGVVAIPAAALGVLMLGPGTSTHAAIRQLANINGCS	93
type I-E CRISPR-associated endonuclease Cas1 [Meiothermus silvanus]	-- [9]	YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVAYYSQEGVVAIPAAALGVLMLGPGTSTHAAIRQLANINGCS	83
CRISPR-associated endonuclease Cas1 [Meiothermus roseus]	MA[17]	YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVAYYGDGAVMIPAAALGVLMLGPGTVVTHAAMRQLANINGCS	93
type I-E CRISPR-associated endonuclease Cas1 [Meiothermus roseus]	-- [9]	YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVAYYGDGAVMIPAAALGVLMLGPGTVVTHAAMRQLANINGCS	83
type I-E CRISPR-associated endonuclease Cas1 [Meiothermus hypogaeus]	MK	YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVAYYSQEGVVAIPAAALGVLMLGPGTSTHAAIRQLANINGCS	76
type I-E CRISPR-associated endonuclease Cas1 [Meiothermus luteus]	MK	YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVAFYQEGVSIIPAAALGVLMLGPGTAVTHAAIRQLANINGCS	76
type I-E CRISPR-associated endonuclease Cas1 [Meiothermus timidus]	MK	YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVAFYQEGVSIIPAAALGVLMLGPGTAVTHAAIRQLANINGCS	76
type I-E CRISPR-associated endonuclease Cas1 [Meiothermus rufus]	MK	YETRNLQELPKFRDGLSYLYLEHGRLEQQDQAVAYYSQEGVVAIPAAALGVLMLGPGTSTHAAIRQLANINGCS	76
type I-E CRISPR-associated endonuclease Cas1 [Truepera radiovictrix]	MP	YTTQNLKELPKFRDGLSYLYLEHGRLEQQDQAVAYYSQEGVVAIPAAALGVLMLGPGTSTHAAIRQLANINGCS	76
type I-E CRISPR-associated endonuclease Cas1 [Thermus tenuipunicus]	MP [3]	-PARNLKEPKFRDGLSYLYVEHAFLEQEAQGGVYDREGLTLPVPAALGVLFLGPGTRITHAAIRLAGINGCT	78
type I-E CRISPR-associated endonuclease Cas1 [Thermus aquaticus]	MP [3]	-SARNLKEPKFRDGLSYLYVEHAVVEREAGGIYDQEGTLAPVAGLGVFLGPGTRITHAAIRLAENGTCT	78
type I-E CRISPR-associated endonuclease Cas1 [Thermus sp. L198]	MP [3]	-SARNLKEPKFRDGLSYLYVEHAVVEREAGGIYDQEGTLAPVPAALGVLFLGPGTRITHAAIRLAENGTCT	78

Figure 2. This image shows the alignment of the start codon for *Cas1* in different species. The name of the CRISPR-Cas endonuclease Cas1 from several different species is on the left in black, listed next to the species names in parentheses, followed by the amino acid sequence in

Score	Expect	Method	Identities	Positives	Gaps
222 bits(565)	2e-75	Compositional matrix adjust.	114/284(40%)	170/284(59%)	2/284(0%)
Query 8	LQELPKFRDGLSYLYLEHGRLEQQDQAVAYYSQEGV-VAIPAAALGVLMLGPGTSITHAA				66
Sbjct 6	L +P +D +S ++L++G+++ D A + G+ IP ++ +ML PGT ++HAA				64
Query 67	IRQLANNGCSVFWWGEEMVRFYASGMGETRSSANLMRQVRAWADPEAHLEVVKRLYLRF				126
Sbjct 65	+R A G + WVGE VR YASG S L+ Q + D + L+VV++++ LRF				124
Query 127	PEPLSPELSLEQIRGLEGVRVRETYARWSRETGVWVKGRNYQRGNWAAADPINRAISAGA				186
Sbjct 125	EP S+EQ+RG+EG RVR TYA +++ GV W GR Y +W D IN+ ISA				184
Query 187	ACLYGLAHAAILSAGYSPALGFIHTGKQLSFVYDVADIYKAETLIPTAFRVVAESDVGVE				246
Sbjct 185	+CLYG+ AAAIL+AGY+PA+GF+HTGK LSFVYD+ADI K +T++P AF + + +				244
Query 247	RRVRHTLREQLKEVKLLERIVSDLHSLFDALETDPDYAADPAAP			290	
Sbjct 245	R VR R+ + K L +++ + + A E P + A P				288

Figure 2. This suggests an orthologous relationship by amino acid alignment of *M. ruber Cas1*, Mrub_3013, “Query” and *E. coli Cas1*, b2755, “Sbjct”. The amino acid sequence is listed and in between the sequences are common amino acids or similar amino acids (+). The important parts of this image are that there are only two gaps in the sequence, a high positive and identity percentage, a very low expected value and high bit score. All these things suggest an orthologous relationship between the two sequences. Analysis done by BLAST at <http://www.ncbi.nlm.nih.gov/blast>.

Another piece of information used to support the orthologous relationship between b2755 and Mrub_3013 is cellular localization data. Both *E. coli* and *M. ruber* are Gram negative Eubacteria and using TMHMM for membrane-embedded alpha helices, PRED for membrane-embedded beta-barrels and PSORT-B localization scores it is clear that b2755 and Mrub_3013 protein products are located in the cytosol. Figure 3 shows two graphs, one of b2755 and the other Mrub_3013 both comparing the amino acid number on the x-axis and probability on the y-axis. The thicker, pink line near the top of the graph shows the minimum probability it would take for the amino acid region to form transmembrane helices, although *M. ruber* has a few peaks they are not statistically likely to be part of transmembrane helices.

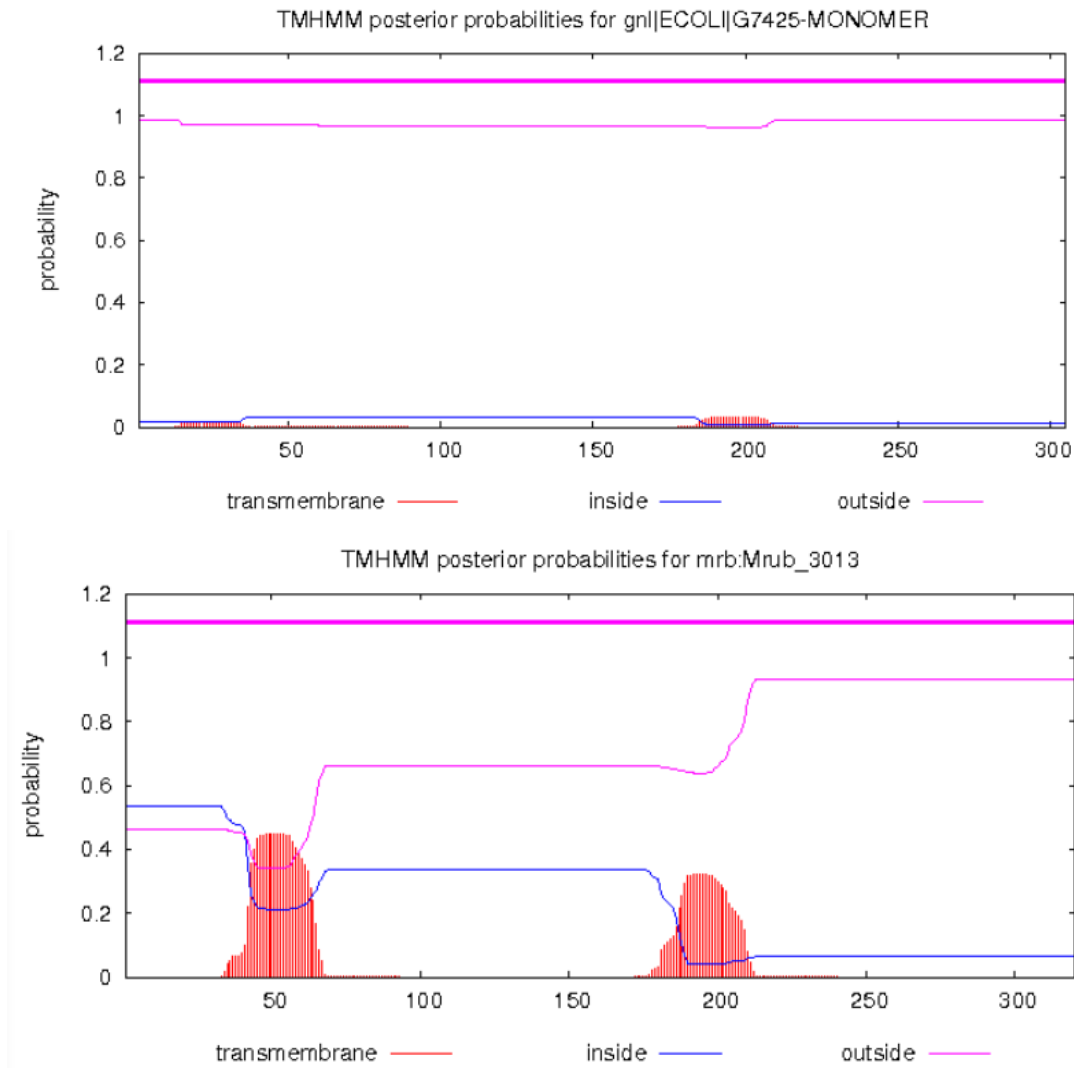


Figure 3. This figure shows TMHMM probability, topical graphs of amino acids in transmembrane helices is shown for *E. coli* b2755 and *M. ruber* Mrub_3013 respectively. There is no statistical probability that either gene product contain membrane embedded alpha helices. Analysis done using TMHMM at <http://www.cbs.dtu.dk/services/TMHMM/>.

Figure 4 shows the posterior probability plots of *E. coli* b2755 and *M. ruber* Mrub_3013 respectively of membrane-embedded beta-barrels. Both plots have strong peaks that seem to be significant but do not reach the number of transmembrane regions needed for an embedded beta barrel. This information shows that there are no embedded barrels encoded in either gene and likely supporting the conclusion of cytosolic function of both proteins.

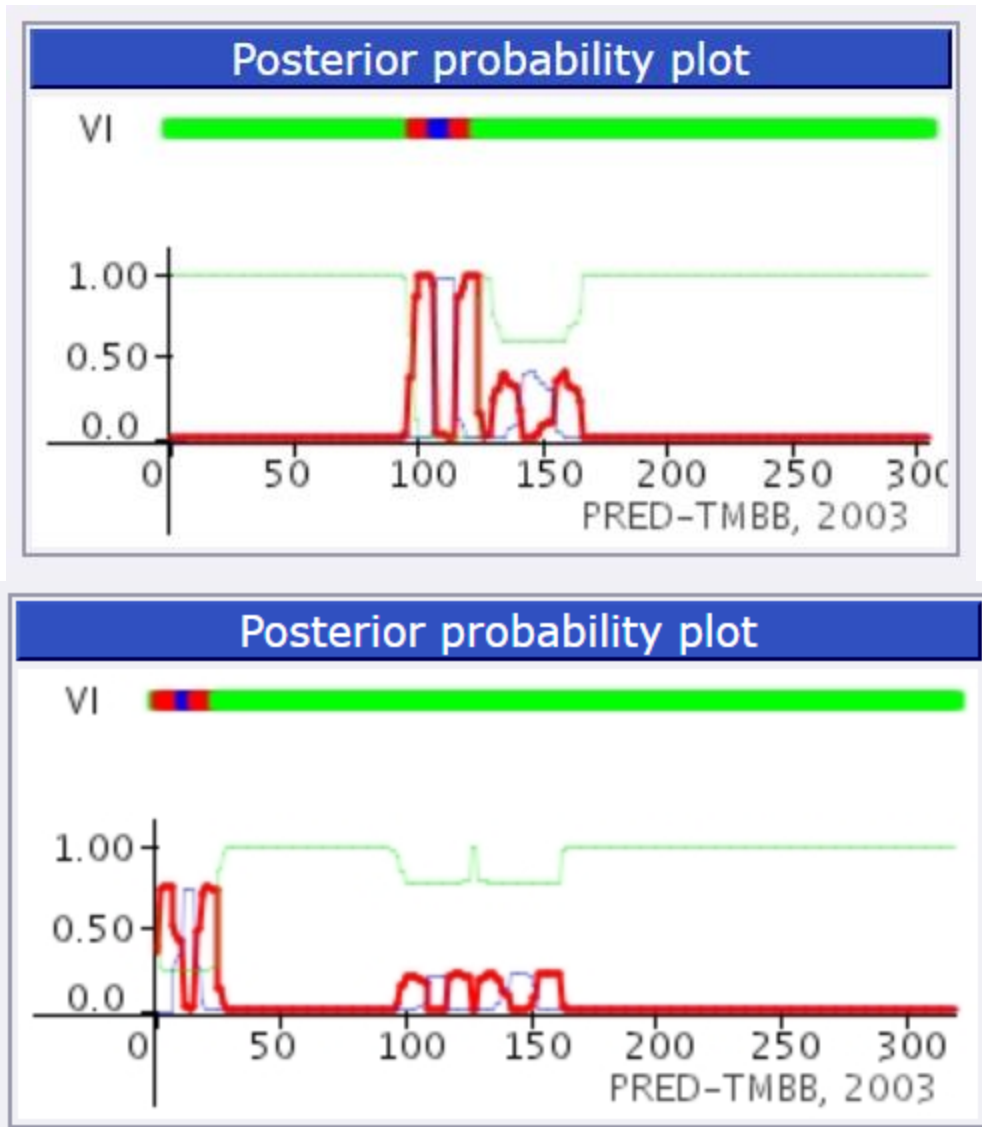


Figure 4. These plots show posterior probability plots of *E. coli* b2755 and *M. ruber* Mrub_3013 respectively of membrane-embedded beta-barrels to be insignificant. Both plots have peaks that do not reach the threshold of embedded beta barrels. Analysis done with PRED-TMBB at <http://bioinformatics.biol.uoa.gr/PRED-TMBB/input.jsp>.

The PSORT-B data assigns scores out of 10 for the probable cellular location of a protein based on its sequence. b2755 and Mrub_3013 both received the same scores suggesting that the protein products are located in the cytoplasm. The scores for all other locations are too low in comparison to the cytoplasmic score of 8.96 indicating the others have no significance. Table 2 compares the scores of *M. ruber* and *E. coli* Cas1 proteins.

Table 2. PSORT-B cellular location scores of both Mrub_3013 and b2755

Cellular location	Scores	
	<i>M. ruber</i> , Mrub_3013	<i>E. coli</i> , b2755

Cytoplasmic score	8.96	8.96
Cytoplasmic Membrane score	0.51	0.51
Periplasmic score	0.26	0.26
Outer Membrane score	0.01	0.01
Extracellular score	0.26	0.26

Analysis done with PSORT-B at <http://www.psort.org/psortb/>.

By using bioinformatics tools listed in Table 3 there is structure-based evidence that b2755 and Mrub_3013 share the same physical features. It is notable that Pfam, TIGRFAM, and CDD all come up with the same results for both genes for the number and name of the top hit. Pfam is an exception because the score and E-value could not be found, but TIGRFAM and CDD show high scores and very low E-values that show significance that they are not aligned to the consensus sequence by chance. PDB does not come up with the same results for number and name as listed in the table. However, the same result that came up as the top hit for Mrub_3013 also came up as a hit for b2755 but not the top hit. Pfam is looking for similar protein families or domains, which we see identified here. TIGRFAM is also looking for protein families and we see the same families between these genes. CDD assigns domains to different Clusters of Orthologous Genes (*aka* COG) groups, with COG hits being defined as similarity to a specific protein domain. PDB found significant similarity to a crystalized Cas1 (or Cas1-Cas2 complex, but from different organisms). We propose that all these data support the hypothesis of structural similarity between Mrub_3013 and b2755, which also indicates similar function.

Table 3. Pfam, TIGRFAM, CDD, and PDB analysis of Mrub_3013 and b2755 hits and E-values

Bioinformatics tool	Number	Name	Score	E-value (top hit)
Pfam	b2755:PF01867 Mrub_3013: PF01867	b2755: CRISPR associated protein Cas1 Mrub_3013: CRISPR associated protein Cas1	b2755: Mrub_3013:	b2755: Mrub_3013:
TIGRFAM	b2755: TIGR03638 TIGR00287 Mrub_3013: TIGR03638 TIGR00287	b2755: cas1_ECOLI: CRISPR-associated endonuclease cas1: CRISPR-associated endonuclease Cas1 Mrub_3013: cas1_ECOLI: CRISPR-associated endonuclease cas1: CRISPR-associated endonuclease Cas1	B2755: 582.8 293.5 Mrub_3013: 579.4 126.9	B2755: 4.8e-172 5.7e-85 Mrub_3013: 5.1e-171 8.5e-35
CDD/COG hits	b2755: cI00656 Mrub_3013: cI00656	b2755: cas1_ECOLI Mrub_3013: cas1_ECOLI	B2755: 455.20 Mrub_3013: 416.68	b2755: 4.41e-148 Mrub_3013: 1.53e-163
PDB	b2755: 5VVK Mrub_3013: 3NKD	b2755: 5VVK: Entity 1 containing chain A, B, C, D Cas1-Cas2 bound to full-site mimic Mrub_3013: 3NKD: Entity 1 containing Chain A, B Structure of CRISP-associated protein Cas1 from Escherichia Coli str. K-12	B2755: 590.497 bits Mrub_3013: 220.32 bits	b2755:1.284e-168 Mrub_3013: 3.54314e-57

Analysis done using Pfam at <http://pfam.sanger.ac.uk/search> TIGRFAM at <http://tigrblast.tigr.org/web-hmm/>, CDD at <http://www.ncbi.nlm.nih.gov/blast>, PDB at <http://www.rcsb.org/pdb/home/home.do>

Figure 5 shows the 3D ribbon structure of the best hit for Mrub_3013, 3NKD. This structure also came up in the hits for b2755; however it was not the best hit. The E-value was still very low showing that the 3NKD is also an accurate depiction of the structure of b2755.

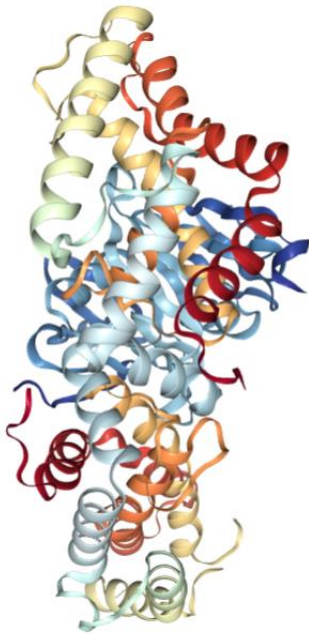


Figure 5. 3D structure of a similar domain found in Mrub_3013 and in b2755 found as a top hit for Mrub_3013 and a significant hit for b2755 in the PDB database. Analysis done using PDB at <http://www.rcsb.org/pdb/home/home.do>.

Lastly operons are highly conserved regions of DNA and are important for identifying orthologs. Mrub_3013 and b2755 have nearly identical gene organization in their likely CRISPR-Cas operons, as shown in Figure 6. Bacteria are highly diverse population and are constantly changing on an evolutionary scale but to see that the CRISPR-Cas operon maintains its integrity through different species then it is strong evidence that the *Cas1* in *M. ruber* and *E. coli* are orthologous. (Nunez *et al.*, 2013)

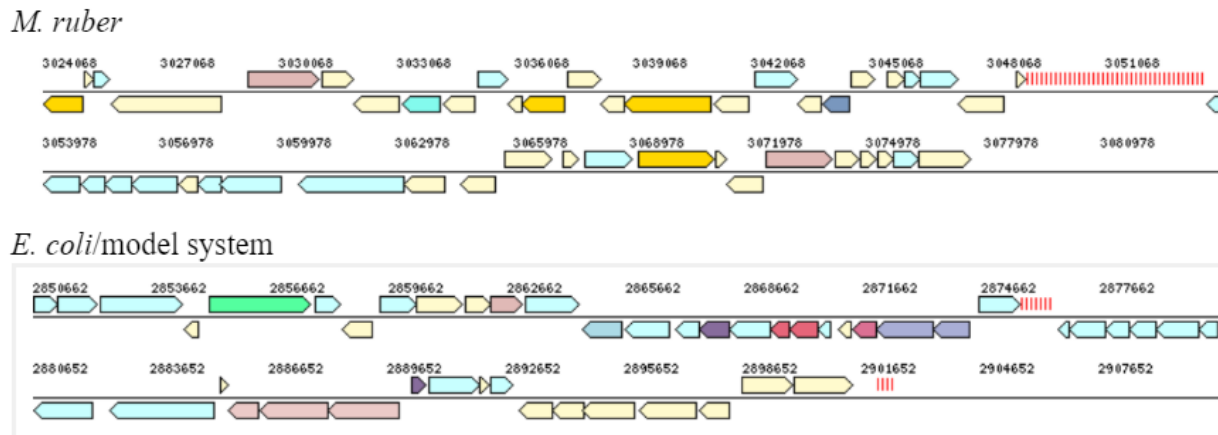


Figure 6. This figure shows the similar CRISPR-Cas operon in both systems regardless of variable surroundings of those operons. Following first set of red vertical lines, the CRISPR Array, in both genomes follows a series of eight light blue genes. In both *M. ruber* and *E. coli* those genes are as follows: *Cas2*, *Cas1*, *Cse3 family*, *Cas5e family*, *Cse4 family*, *Cse2 family*, *CasA/Cse1 family*, *Cse3 family*. The *M. ruber* sequence has one light yellow gene following *Cse2 family* gene which is identified as a hypothetical protein with no known purpose. This shows high levels of conservation of this system throughout evolutionary time. Analysis done with IMG/M at <https://img.jgi.doe.gov/cgi-bin/m/main.cgi>.

This same order of operons was found to be conserved throughout various species as well. Also, there seem to be no paralogs of Mrub_3013 or b2755 in their respective genomes.

Discussion

Overall this data supports the conclusion that Mrub_3013 and b2755 are orthologous to one another. The strongest evidence is the NCBI BLAST sequence alignment with a high identity and low E-value, the same cellular localization in the cytosol, structural basis of the same family and domains, and conservation in the CRISPR-Cas operon. It may be interesting to do further study about Cas1 interactions with Cas2 in a complex to see if there are similarities and differences in that interaction from *M. ruber* to *E. coli*.

Literature Cited

Bagos PG, Liakopoulos TD, Spyropoulos IC and Hamodrakas SJ.

PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins.

Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E.. [Internet]. 2000. The Protein Data Bank. [2016 Dec 6]. Available from: <http://www.rcsb.org/>.

Darmon, E., & Leach, D. R. (2014). Bacterial Genome Instability. *Microbiology and Molecular Biology Reviews*, 78(1), 1-39. doi:10.1128/mmbr.00035-13

Finn, R.D., Coghill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A. 2016. The Pfam protein families database: towards a more sustainable future: *Nucleic Acids Res.*, 44:D279-D285; [2016, Dec. 6]. Available from: <http://pfam.xfam.org/>

Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O. 2001. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res* 29(1):41-3.

Jiang, F., & Doudna, J. A. (2015). The structural biology of CRISPR-Cas systems. *Current Opinion in Structural Biology*, 30, 100-111. doi:10.1016/j.sbi.2015.02.002

Kanehisa, M. and Goto, S.; KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27-30 (2000). Available from: <https://www.kegg.jp/kegg/>

Keseler, I.M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martinez, C., Fulcher, C., Huerta, A.M., Kothari, A., Krummenacker, M., Latendresse, M., Muniz-Rascado, L., Ong, Q., Paley, S., Schroder, I., Shearer, A., Subhraveti, P., Travers, M., Weerasinghe, D., Weiss, V., Collado-Vides, J., Gunsalus, R.P., Paulsen, I., and Karp, P.D. 2013. EcoCyc: fusing model organism databases with systems biology *Nucleic Acids Research* 41:D605-612.

Krogh A, Rapacki K. TMHMM Server, v. 2.0. Cbs.dtu.dk. 2016 [accessed 2016 Dec 6]. <http://www.cbs.dtu.dk/services/TMHMM/>

Madden T. The BLAST Sequence Analysis Tool. 2002 Oct 9 [Updated 2003 Aug 13]. In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2002-. Chapter 16. Available from: <http://www.ncbi.nlm.nih.gov/books/NBK21097/>

Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. CDD: NCBI's conserved domain database. *Nucleic Acids Res.* 28(43): D222-2: [2016 Dec 6]. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/25414356?dopt=AbstractPlus>

Markowitz VM, Chen IA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, et al. 2012. IMG: The integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research* 40(D1):D115-22. Available from: <http://nar.oxfordjournals.org/content/40/D1/D115.full>

Nucleic Acids Res, 2004 Jul 1;32(Web Server issue):W400-4.

- Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E.. [Internet]. 2000. The Protein Data Bank. [2016 Dec 6]. Available from: <http://www.rcsb.org/>.
- Nuñez, P. A., Romero, H., Farber, M. D., & Rocha, E. P. (2013). Natural Selection for Operons Depends on Genome Size. *Genome Biology and Evolution*,5(11), 2242-2254. doi:10.1093/gbe/evt174
- N.Y. Yu, J.R. Wagner, M.R. Laird, G. Melli, S. Rey, R. Lo, P. Dao, S.C. Sahinalp, M. Ester, L.J. Foster, F.S.L. Brinkman (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, *Bioinformatics* 26(13):1608-1615
- Tindall, B. J., Sikorski, J., Lucas, S., Goltsman, E., Copeland, A., Rio, T. G., . . . Lapidus, A. (2010). Complete genome sequence of *Meiothermus ruber* type strain (21T). *Standards in Genomic Sciences*,3(1), 26-36. doi:10.4056/sigs.1032748
- Wang, J., Li, J., Zhao, H., Sheng, G., Wang, M., Yin, M., & Wang, Y. (2015). Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. *Cell*,163(4), 840-853. doi:10.1016/j.cell.2015.10.008
- Wright, A., Nuñez, J., & Doudna, J. (2016). Biology and Applications of CRISPR Systems: Harnessing Nature's Toolbox for Genome Engineering. *Cell*,164(1-2), 29-44. doi:10.1016/j.cell.2015.12.035