Winter 2-13-2019

Mrub_3019 casA gene is an ortholog to *E. coli* b2760

Kelsey Heiland
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

*Mrub_3019 cas*A gene is orthologous to *E. coli* b2760 gene.

Kelsey Heiland and Lori Scott

ABSTRACT. This research is part of the *Meiothermus ruber* genome annotation project which aims to predict gene function with various bioinformatics tools. We investigated the function of Mrub_3019, which encodes the CasA protein involved in the multi-subunit effector complex for the CRISPR-Cas immunity system and predicted it to be an ortholog of *E. coli* K12 MG1655 b2760 (*cas*A). We predicted that Mrub_3019 encodes the protein CasA, which is involved in PAM recognition of CRISPR interference pathway. Foreign DNA will bind to CasA, which signals Cas3 for helicase-mediated DNA degradation. Our hypothesis is supported by low E-values for pairwise alignment in NCBI BLAST, Pfam, and TIGRfam. Both proteins were predicted to be non-transmembrane-embedded, and in close proximity to Cas3 in the Type I-E CRISPR-Cas complex. Finally, both revealed several amino acids that were identical in the Pfam domain hit (PF0481) for the CRISPR_Cse1 family.

Key words: *Meiothermus ruber,* genome, bioinformatics, annotation, CRISPR-Cas, prokaryotic defense, Mrub_3019, b2760, ortholog, *cas*A

INTRODUCTION

What is the M. ruber genome analysis project?

Deriving from Greek words, ‘meion’ and ‘thermos’, the gram-negative organism *Meiothermus ruber* is characterized as a rod-shaped eubacteria that thrive in warm climates fluctuating around 60°C, and its second name derives from its red pigmentation. In the Thermales order, *M. ruber* is characterized by highly dependent thermostability and was sequenced for the *Genomic Encyclopedia of Bacteria and Archaea* (GEBA) because of its phylogenetic relationship (Tindall *et al.,* 2010). The number of publications and research on *M. ruber* is significantly under developed leaving gaps between evolutionarily-related species. The limited research leaves functions of genes within the *M. ruber* genome unstudied which may provide novel relationships and genetic variations. Annotations of *M. ruber* genome were completed by the DOE Joint Genome Institute (JGI) to help examine and provide better understanding of poorly studied bacteria and defining biochemical and the underlying evolutionary processes.

E. coli as the model organism

Biogenesis in under-studied bacterial species must be cross analyzed against well-known model organisms like *Escherichia coli*, to determine if their mechanisms are comparable. *E. coli* is one of the one of the most well-studied bacterial organisms and has its own website, EcoCyc, to provide vast and reliable information on biochemical processes and structures. Most of the information on EcoCyc and been found through extensive experimentation. Bioinformatic tools, which are typically computer programs that aid in collecting data of biological complexes like genetic information, were used for comparing the model organism to the gene of interest in *M. ruber* (Keseler *et al.,* 2013). Of these tools, NCBI BLASTs revealed location-specific genes, Mrub_3019 from *M. ruber* and b2760 from *E. coli* indicating a possible orthologous relationship of the CRISPR-Cas system (Madden, 2002). In this thesis, I will examine the *cas*A gene function and structure in the type I-E CRISPR-Cas systems of *Escherichia coli*, and *Meiothermus ruber*.

What is the CRISPR-Cas system?

Bacteria and archaea harbor defense mechanisms that range in target specificity against foreign invaders. Viruses (bacteriophage) will infect a bacteria by attaching to proteins in the membrane and injecting its own DNA through the cell wall. Successful infection will lead to replication of the viral DNA by first connecting and hijacking the host cell's DNA and ribosomes. Once replication is complete, the bacteriophage initiates cell lysis in the host, resulting in the new DNA being released into the environment to find a new host. To combat this, bacteria will attempt to defend themselves using non-specific biochemical mechanisms, like preventing and blocking foreign DNA injection, or abortive infection, where a bacteriophage enters the host cell, but then fail replicate. Highly specific and more successful defense mechanisms in bacteria include restriction modification and sugar non-specific nucleases that degrade foreign protein components, and the CRISPR-Cas (clustered regularly interspaced short palindromic repeats-CRISPR associated proteins) system whose adaptive immunity derives from the genes ability to archive information from past invasions (Horvath & Barrangou, 2010).

The CRISPR array is a collection of location and function-specific genes contains a leader sequence, a 5’-AT-3’ rich region, repeat regions, and spacer regions which carry unique DNA from previous invaders, called protospacers. The first stage of CRISPR called acquisition, protospacers of bacteriophage from past infections using are detected using protospacer adjacent motif (PAM) sequences (Wright, Nunez, & Doudna, 2016). Detection of bacteriophage DNA in the cell will induce transcription of the CRISPR region and excising the spacer regions to make mature crRNA. Together, the *cas* genes and mature crRNA form an effector complex which bind and degrade complementary regions of foreign DNA. *Cas* genes assist by encoding proteins that possess analogous functions in the immunity response among subtypes, like foreign DNA detection and degradation.

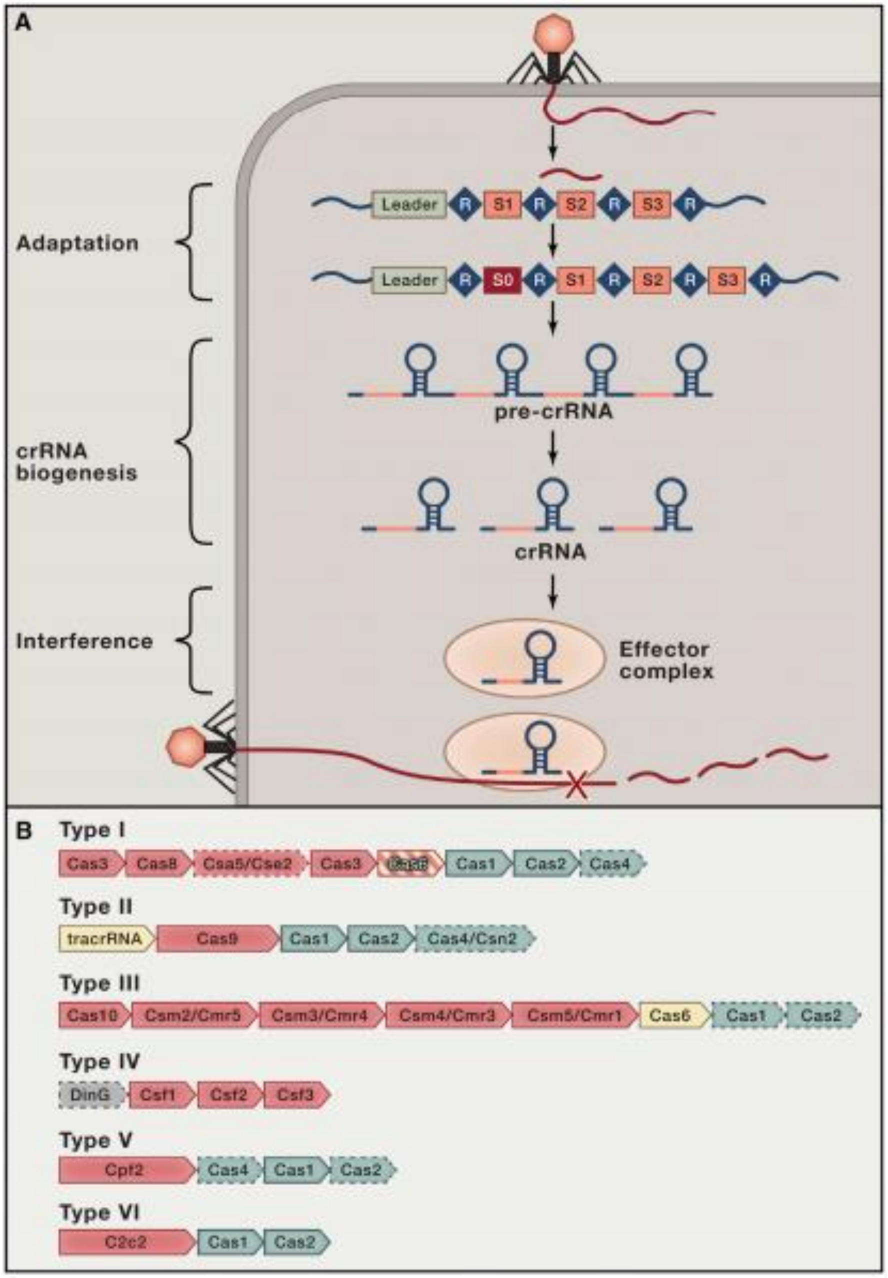


Figure 1. Model mechanism and types if CRISPR-Cas systems. **(A).** Three stages of CRISPR-Cas system immunity response. Protospacers of foreign DNA are selected and inserted between repeats sequences near the leader sequence. CRISPR array is transcribed to produce mature crRNA guiding Cas proteins to form the effector complex necessary for degradation in the last stage of interference. ‘R’ for repeat sequence; ‘S’ for spacer sequence. **(B)**. Six main types of CRISPR-Cas systems characterized by unique operon components. Variable gene within subtypes are dashed. Adaptation genes are coded with blue, crRNA biogenesis genes are coded with yellow, and interference genes are coded with red (Wright *et al.,* 2016).

CRISPR system in E. coli K12

The CRISPR-Cas system is an operon that coordinately expresses *cas* genes and CRISPR array. The response of the CRISPR-Cas system relies on *cas* genes encoding proteins that serve multiple functions for defense, which are used to categorize six types and two classes. (Wright, Nunez, & Doudna, 2016). Types I, III, and IV are characterized as Class 1 systems for their multi-subunit effector (Cascade), in comparison to types II, V, and VI who only possess single-subunit effectors. The type I-E CRISPR-Cas multisubunit-effector complex is comprised of several smaller single-subunits, including CasA, the protein of interest for this study, and almost all of them directly are involved in crRNA biogenesis (Jackson *et al.,* 2014). Previous research examining the RNA-binding proteins involved in this system revealed the configuration and functions of the Cascade complex (Amlinger, 2016). Cas subunits come together to produce a 3' stem-loop of mature crRNA and a 5' handle involved in binding the Cascade and crRNA. Studies have found CasA protein to be weakly associated with the Cascade and mediates detection via PAM recognition (Sashital, Wiedenheft, & Doudna, 2012). While the cell is not under invasion of foreign DNA, PAM sequences regulate and prevent plasmid transformation within the cell's own genes (Westra *et al.,* 2013). The loose link between CasA and the rest of the protein complex, allows for it to cover more surface area to potentially detect a foreign PAM sequence.

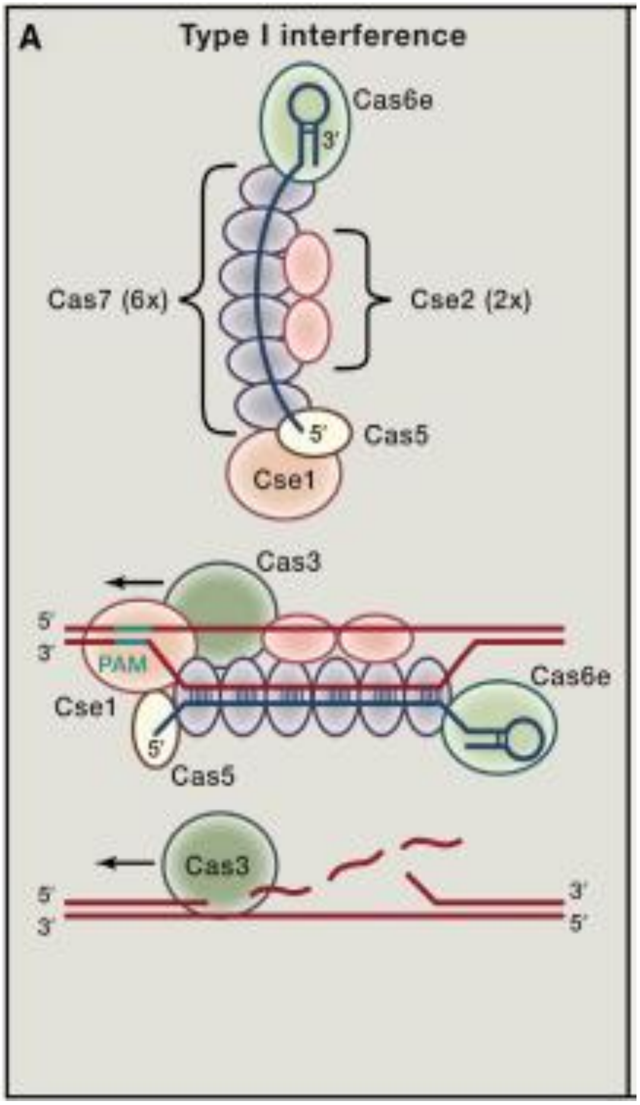


Figure 2. Cascade and Cas3 interference mechanism in type I. The Cascade complex is formed by crRNA held together by a unique configuration of *cas* genes for detecting foreign DNA. Recognition of foreign PAM sequences by CasA, induce the large complex to unwind the foreign DNA and binding crRNA. Cas3 is signaled by CasA (depicted as Cse1) which travels along the strand while simultaneously degrading it (Wright *et al.,* 2015).

PAM sequences of *E. coli* type I-E were found at the 3' end near the target protospacer to prevent self-targeting and recognized by CasA resulting in apoptosis (Amlinger, 2016). Once the foreign PAM sequences are recognized by the CasA protein and binds to the target protospacer, it causes the first nucleotides in the sequence to flip outward, preparing it for crRNA-protospacer base pairing (Shashital *et al.,* 2012). Evidence has suggested that the first bases of a unique sequence in a protospacers are heavily influenced by foreign target recognition from PAM sequences (Amlinger, 2016). The Cascade complex is seahorse-shaped, where CasA covers a large corner, of it and allowing PAM sequences to be easily detected. Once CasA has detected a foreign PAM, the Cascade will undergo reconfiguration, in which CasA will rotate around to open space for Cas3, a protein known for its involvement in nuclease and helicase-mediated degradation of foreign DNA.

Previous research has shown that CasA has a critical role in the Cascade system in target-validation to initiate degradation of foreign DNA via Cas3 (Hochstrasser *et al.,* 2014). This was found by examining the direct contact CasA has with the PAM sequence of foreign DNA. CasA was found to aid in the system's detection mechanism by being intentionally located where unmutated foreign PAM sequences easily bind to aL1 loop structure, unique to the PAM sequence. Mutated PAM sequences impair Cas3 from cleaving the DNA because they failed to be detected by CasA. This feature suggests that CasA governs DNA-binding-specific, and homologous DNA joint formation, generally by catalytic reaction of a DNA recombination protein called RecA, which is present in *E. coli* (Shinohara *et al.,* 2015).

Within the L1 loop, Asp-161 was found to be a highly conserved amino acid and involved in recognition via RecA protein. The Asp-161 in the *cas*A gene also showed high favorability in selecting single-stranded DNA for binding. An alanine substitution mutation (D161A) to Asp-161 showed to decrease binding and joint formation which suggests RecA's preference is dependent on the negative charge at Asp-16. In the presence of ATP, and researchers proposed that RecA's specific selectivity is driven by steric barriers and electrostatic repulsion of double-stranded DNA. The function of the *cas*A gene is characterized in *E. coli* as being a part of the type I-E CRISPR-Cas subsystem which signals Cas3-mediated degradation and governs RecA DNA-specific binding proteins.

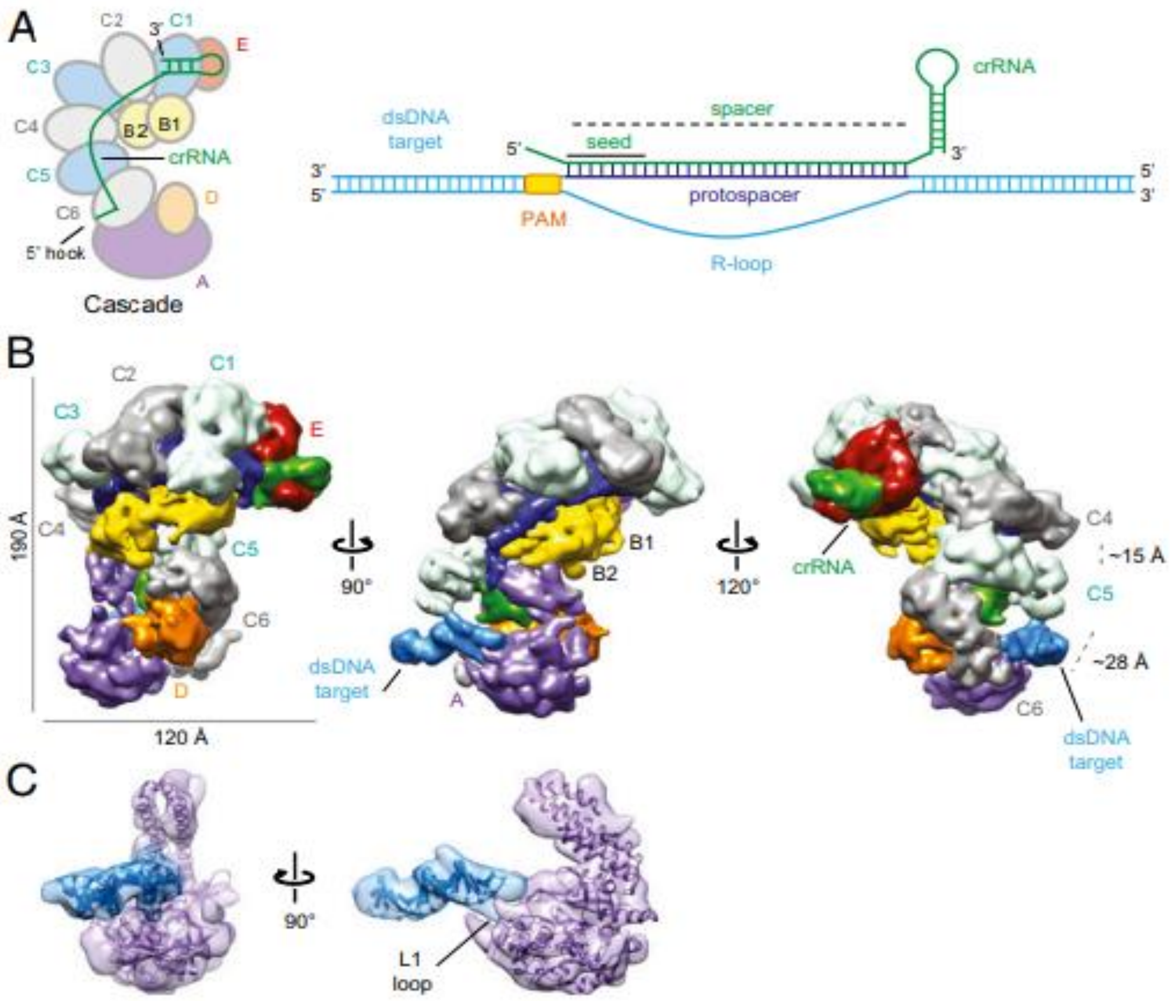


Figure 3. Structure of Cascade complex binding foreign DNA and positioning the PAM sequence near CasA. **(A)**. Cascade subunit, CasA depicted as large light purple region near 5' hook, CasB (yellow); CasC (light blue and gray); CasD (orange); CasE (red), and crRNA configuration (green) and foreign DNA (blue) shown to the right of the Cascade. **(B)**. Crystal structure representation of DNA binding to the Cascade complex and rotating for Cas3 recruitment. **(C).** Foreign DNA docking into the L1 loop of CasA. (Hochstrasser *et al.,* 2014).

Purpose of the study

Using *E. coli* as a model organism, the goal of this annotation is to provide insight into the CRISPR-Cas system, particularly type I-E *cas*A gene in *M. ruber*, to fill gaps in the literature. By using bioinformatic tools, orthology between the *cas*A gene of Mrub_3019 and *E. coli* b2760 can be studied to help provide for future research in adaptive functions of the CRISPR-Cas system immunity response in *Meiothermus ruber*, and among diverse bacteria.

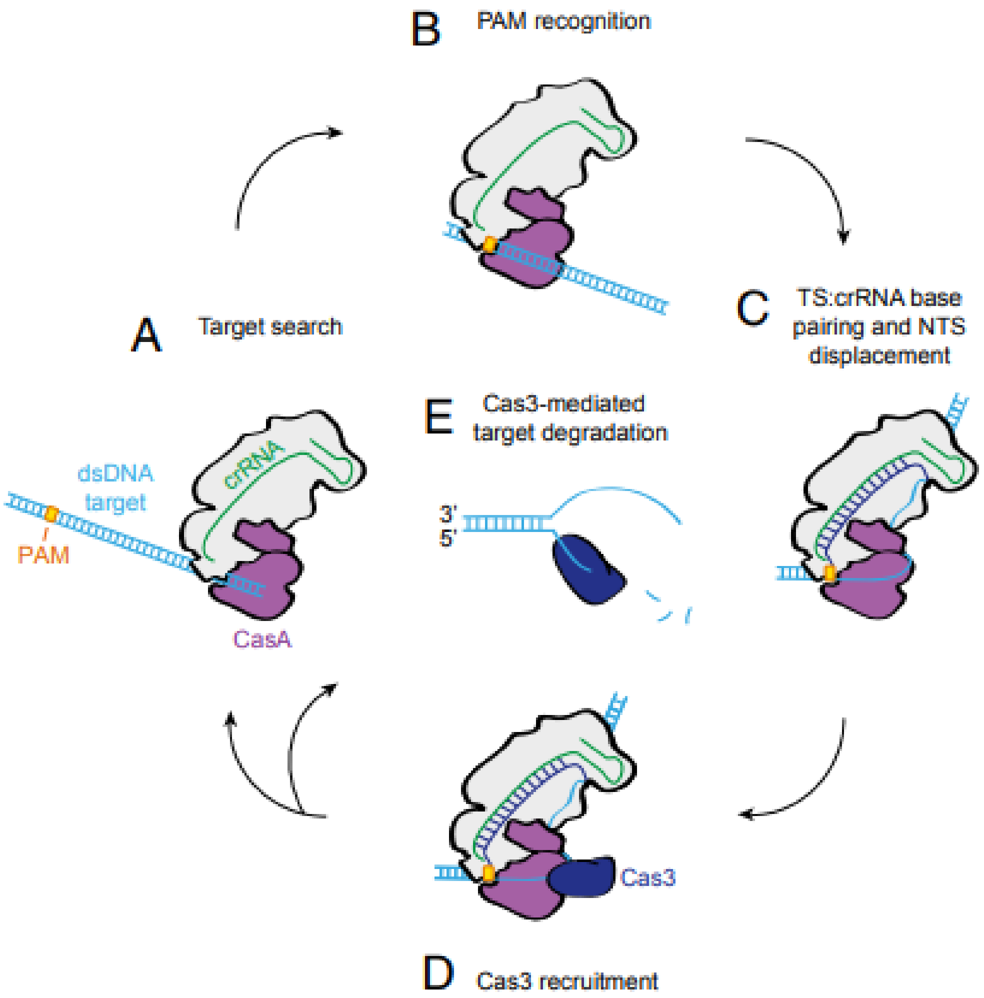


Figure 4. CasA-mediated mechanism for Cas3 signaling. (A). Cascade searching for PAM sites of foreign DNA. (B). L1 loop of CasA detects PAM site and (C), induces structural changes of the Cascade complex to (D), signal Cas3 for (E), degradation of foreign DNA (Hochstrasser *et al.,* 2014).

METHODS

Information on the CRISPR-Cas system in *Escherichia coli* K12 MG1655 was first collected using the database, EcoCyc, which carries extensive background information on many biological processes of this well-studied prokaryote (Keseler *et al.,* 2013). The KEGG and GenBank databases were used to obtain FASTA-formatted nucleotide and amino acid sequences of the genes of interest, as well as identifying the components of *E. coli's* CRISPR-Cas system (Kanehisa *et al.,* 2019; GenBank). The IMG/M database (Markowitz *et al.,* 2012) was used to determine if *M. ruber* was predicted to have one or more CRISPR-Cas systems and the presence of CRISPR-Cas operon(s). Once the CRISPR-Cas system in *Meiothermus ruber* DSM1279 was identified, Mrub_3019, a putative *casA* gene, was chosen for this project. A protein BLAST (Juncker *et al.,* 2017) was performed once the start codon of Mrub_3019 was confirmed, to determine how well Mrub_3019 and *E. coli* CasA (locus tag b2760) aligned.

Three tools were used to predict the cellular location of Mrub_3019. The presence of transmembrane alpha helices was predicted by TMHMM (Juncker *et al.,* 2017), and the presence of outer membrane beta-barrels was predicted by PREDD (Bagos, Liakopoulos, Spyropoulos, & Hamodrakas, 2004). PSORTb (Yu *et al.,* 2010) generates prediction results for five major localizations for Gram-negative bacteria (cytoplasmic, inner membrane, periplasmic, outer membrane and extracellular) and four localizations for Gram-positive bacteria (cytoplasmic, cytoplasmic membrane, cell wall and extracellular).

In addition to an NCBI BLAST alignment to quantify their sequence similarity, functional characteristics of Mrub_3019 and *E. coli* CasA were compared. For example, we used the CDD tool (Marchler-Bauer *et al.,* 2016) and Pfam tool (Finn *et al.,* 2016) to identify possible protein domains. Assigning Mrub_3019 and *E. coli* CasA to a particular protein family was achieved using TIGRFam (Haft *et al,* 2001). We mined the PDB database (Berman *et al.,* 2000) for putative orthologs to Mrub_3019 and E. coli CasA. The PDB archive contains information about experimentally-determined structures of proteins, as well as containing a wealth of curated functional information. Chromosome maps from IMG/M (Markowitz *et al.,* 2012). were used identify the roles of CRISPR-Cas proteins and CRISPR-Cas operon organization, and well as assessing commonality between related species and the gene of interest's (GOI) position in relation to other genes

RESULTS

Table 1 compares *E. coli* b2760 gene and *M. ruber* Mrub_3019 gene using outputs from various bioinformatics tools. The table begins with the protein BLAST analysis to compare the sequences. The BLAST revealed some resemblance between the two genes.

The low bit score of 35.0 can be attributed to the significant difference in sequence length, as well as the percent identity of 37%. The E-value 4e-06 provides evidence of significant similarity between the two sequences, which is well below the cot-off of 0. Consequently, we are confident these two proteins do not align just by chance but have are similar due to their functional similarities. The BLAST alignment reveals that these two sequences may share a common ancestor.

PSORT-B was used to predict the likelihood of protein cellular location (Yu *et al.,* 2010). Mrub_3019 was not determined and remained unknown. *E. coli* b2760 was predicted to be 89.6% likely to be in the cytoplasm. By eliminating the possibility that either gene was located within the membrane, and defining the most probable location of b2760, can indicate that there is a higher probability that Mrub_3019 is in the same place as b2760.

Pfam was used to determine similar protein domain by aligning the protein against their respective conserved sequences (Finn *et al.,* 2016). Alignment of *E. coli* b2760 against its conserved sequence produced the same domain (PF09481), CRISPR-associated protein Cse1 that Mrub_3019 was determined to have. Determining a homologous protein domain in the two genes, provides significant indication of an orthologous relationship. Figures for the pairwise alignments were not obtained due processing errors on the Pfam database.

Table 1. Comparison and characterization of the gene Mrub_3019 and *E. coli* b2760 using measurements from various bioinformatics tools and databases.

Database/ Bioinformatics tool			
KEGG	Locus tag	Mrub_3019	b2760
BLAST *E. coli* b2760 against Mrub_3019	Score	35.0	
	E-value	4e-06	
	Identities	58/244 (37%)	
TMHMM	Predicted α-helices	0	
PRED	Predicted β-barrels	0	
PSORT-B	Predicted cellular location	Unknown	Cytoplasmic location
CDD	COG number	pfam09481	PRK09693
	COG name	CRISPR_Cse1	Cascade antiviral complex
	E-value	1.60e-109	0e+00
	Score	333.56	892.21
TIGRFAM	TIGRFAM number	TIGR02547	
	TIGRFAM name	CRISPR system CASCADE complex	
	E-value	1.3e-34	2.2e-217
	Score	124.6	731.7
PFam	PFam number	PF09481	
	PFam name	CRISPR-associated protein Cse1	
	E-value	1.4e-104	4.8e-140
PDB	PDB code	4F3E	4QYZ
	PDB name	CasA crystallized in *Thermus thermophilus*	Crystal structure of CRISPR RNA-guided surveillance complex
	E-value	1.45358e-47	0e+0.0
	Score	189.119	1014.22

TMHMM and PRED predicted zero α-helices and β-barrels, respectively, of both proteins indicating neither are embedded or pass through the cell membrane. This was consistent with PSORT-B, which determined *E. coli* b2760 was predicted to be in a cytoplasmic location, while Mrub_3019's location could not be determined. The cellular location of Mrub_3019 was narrowed down, providing some evidence of an ortholog to b2760, but was still ultimately undefined.

The CDD pulled different COG numbers and the PDB revealed different PDB codes for Mrub_3019 and *E. coli* CasA, but the associated protein names suggested similar function. The same TIGRfam (TIGR02547) and Pfam numbers (PF09481) were pulled from the respective databases, thereby indicating similar protein families and domains. The high E-values from both TIGRfam and Pfam bioinformatics tools further suggest an orthologous relationship between the two genes. Figure 5 depicts quaternary protein structures representations Mrub_3019 and *E. coli* b2760.

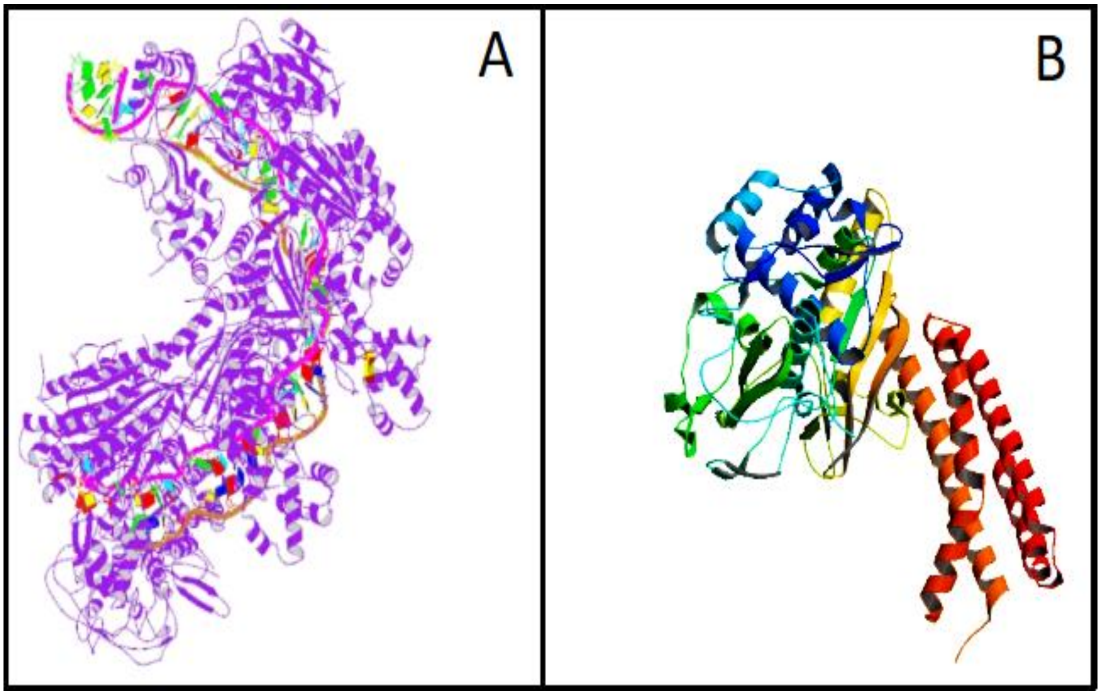
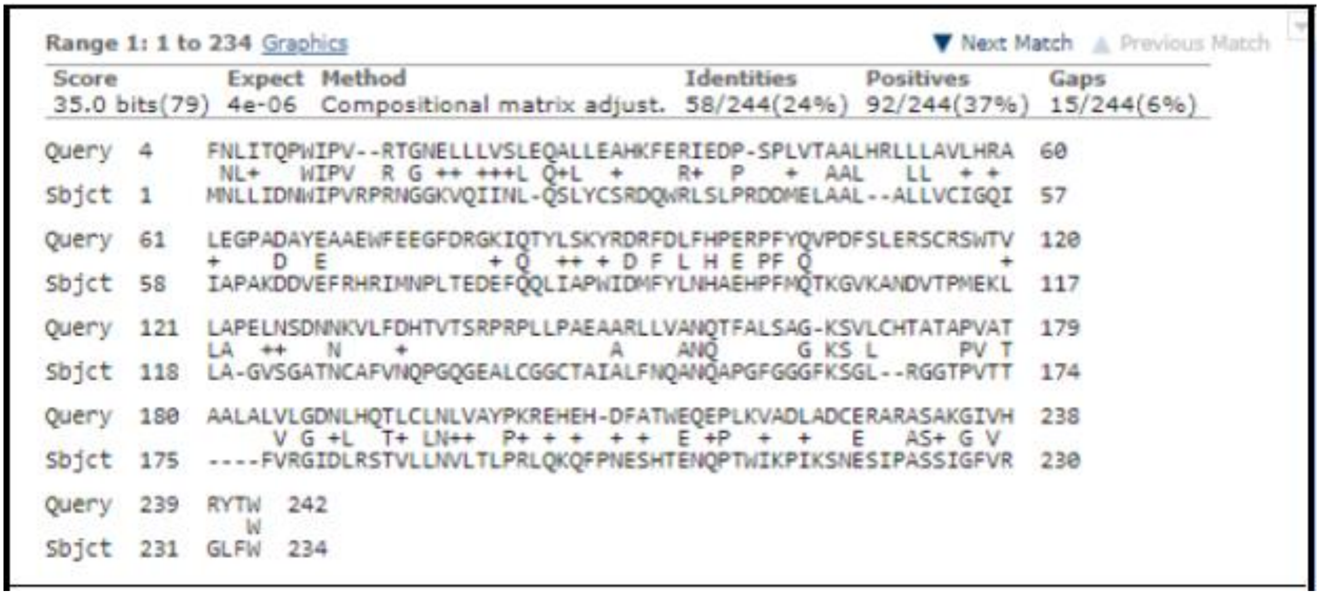


Figure 5. Crystal structures of the top hits for *E. coli* b2760 and Mrub_3019 against the PDB database. **(A).** Crystal structure of the top hit for *E. coli* b2760 with PDB code 4QYZ (CRISPR RNA-guided surveillance complex protein). **(B).** Crystal structure of the top hit for Mrub_3019 with PDB code 4F3E (CasA protein) (Markowitz et al., 2012). Analysis was performed using the Protein Data Bank (PDB) database from http://www.rcsb.org/.

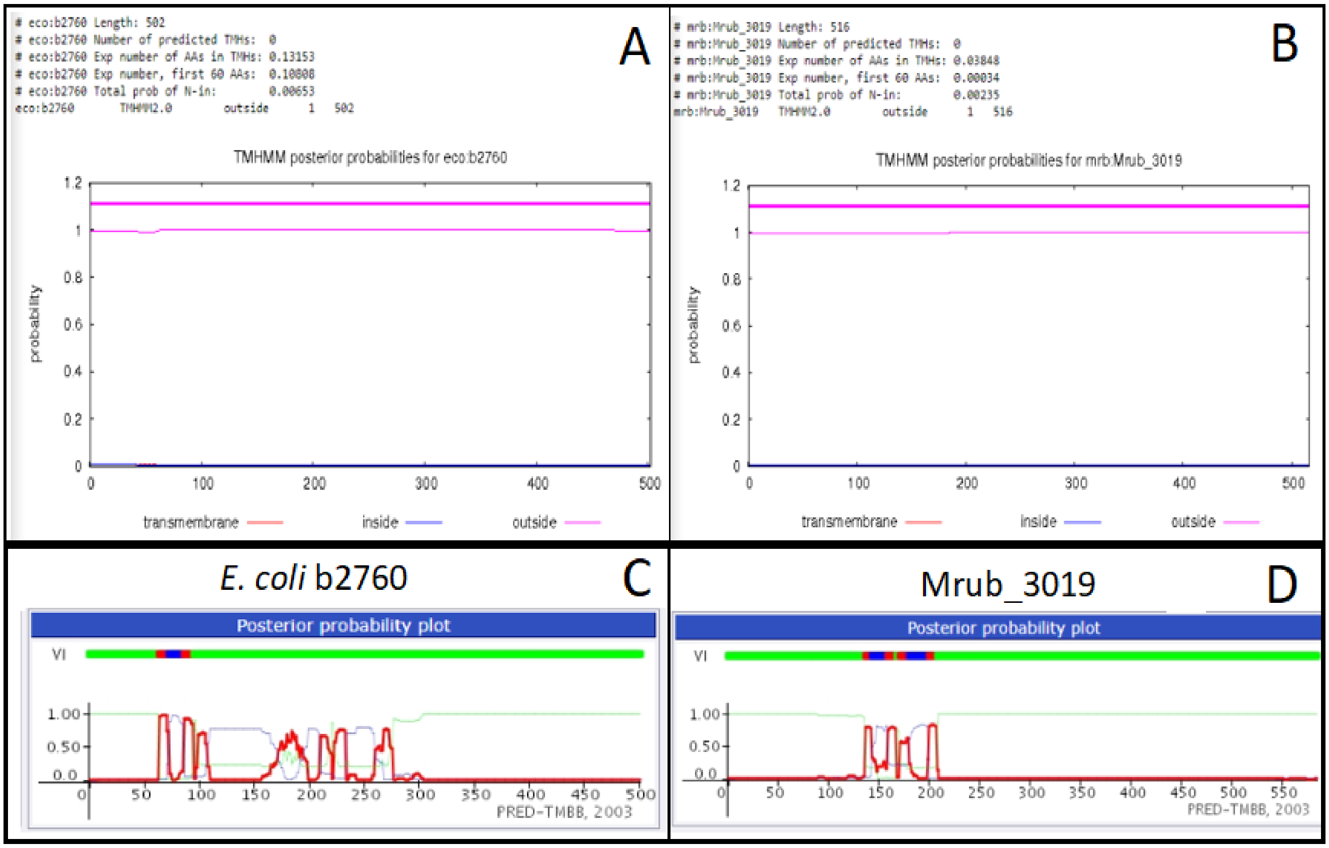
Figure 6 shows three BLAST alignments between Mrub_3019 and *E. coli* b2760 (Madden, 2002). The last two alignments in the figure are very short in sequence length, which causes the alignment to have a high and misrepresentative percent identity. Along with this evidence, the E-values were significantly high (7.7) which suggests that the last two alignments were most likely aligned by chance and not evolutionarily related. The top hit showed a 38% identity, with a lower E-value of 4e-06. The top alignment was also significantly longer in amino acids that the other two, which provides strong evidence that Mrub_3019 is an ortholog of *E. coli* b2760.



```
Range 1: 1 to 234 Graphics                                       ▼ Next Match  ▲ Previous Match
Score           Expect  Method                          Identities     Positives      Gaps
35.0 bits(79)   4e-06   Compositional matrix adjust.    58/244(24%)    92/244(37%)    15/244(6%)

Query  4      FNLITQPWIPV--RTGNELLLVSLEQALLEAHKFERIEDP-SPLVTAALHRLLLAVLHRA  60
               NL+   WIPV  R G ++ +++L Q+L  +    R+  P   +  AAL   LL  + +
Sbjct  1      MNLLIDNWIPVRPRNGGKVQIINL-QSLYCSRDQWRLSLPRDDMELAAL--ALLVCIGQI  57

Query  61     LEGPADAYEAAEWFEEGFDRGKIQTYLSKYRDRFDLFHPERPFYQVPDFSLERSCRSWTV  120
              +    D  E            + Q  ++ + D F L H E PF Q              +
Sbjct  58     IAPAKDDVEFRHRIMNPLTEDEFQQLIAPWIDMFYLNHAEHPFMQTKGVKANDVTPMEKL  117

Query  121    LAPELNSDNNKVLFDHTVTSRPRPLLPAEAARLLVANQTFALSAG-KSVLCHTATAPVAT  179
              LA  ++   N    +               A    ANQ      G KS L      PV T
Sbjct  118    LA-GVSGATNCAFVNQPGQGEALCGGCTAIALFNQANQAPGFGGGFKSGL--RGGTPVTT  174

Query  180    AALALVLGDNLHQTLCLNLVAYPKREHEH-DFATWEQEPLKVADLADCERARASAKGIVH  238
                   V G +L  T+ LN++  P+ + +  + +  E +P  +  +   E   AS+ G V
Sbjct  175    ----FVRGIDLRSTVLLNVLTLPRLQKQFPNESHTENQPTWIKPIKSNESIPASSIGFVR  230

Query  239    RYTW  242
                 W
Sbjct  231    GLFW  234
```

Figure 6. Protein BLAST alignment (Madden, 2002) of the Mrub_3019 and *E. coli* b2760 gene sequences. Protein BLASTs were performed using NCBI BLAST bioinformatics tool at http://www.ncbi.nlm.nih.gov/blast.

Figure 7 is the TMHMM and PRED topology graphs of Mrub_3019 and *E. coli* CasA with zero transmembrane helices (Krogh & Rapacki, 2016) and no beta-barrel (Bagos *et al.,* 2002). Consequently, it is unlikely these proteins leave the cytoplasm.



```
# eco:b2760 Length: 502
# eco:b2760 Number of predicted TMHs:  0
# eco:b2760 Exp number of AAs in TMHs: 0.13153
# eco:b2760 Exp number, first 60 AAs:  0.10808
# eco:b2760 Total prob of N-in:        0.00653
eco:b2760       TMHMM2.0        outside      1   502
```

```
# mrb:Mrub_3019 Length: 516
# mrb:Mrub_3019 Number of predicted TMHs:  0
# mrb:Mrub_3019 Exp number of AAs in TMHs: 0.03848
# mrb:Mrub_3019 Exp number, first 60 AAs:  0.00034
# mrb:Mrub_3019 Total prob of N-in:        0.00235
mrb:Mrub_3019   TMHMM2.0        outside      1   516
```

Figure 7. TMHMM and PRED Topology graphs for E. coli CasA and Mrub_3019 suggest a cytoplasmic location. **(A)**. Topology graph of *E. coli* b2760 for the prediction of the number of alpha-helices (0). **(B)**. Topology graph of Mrub_3019 predicted no zero alpha-helices. The TMHMM bioinformatics tools was used to gather this data at http://www.cbs.dtu.dk/services/TMHMM/ (Krogh & Rapacki, 2016). **(C).** Topology graph of *E. coli* b2760 indicates a few peaks but it is not significant enough to be a beta-barrel (0). **(D).** Topology graph of Mrub_3019 gene also predicting zero beta-barrels. Beta-barrel structures of the proteins were collected from PRED bioinformatics tool (Bagos *et al.,* 2004) at http://bioinformatics.biol.uoa.gr/PRED-TMBB/input.jsp.

Data collected from the IMG/M database depicted the chromosome maps of the CRISPR-Cas system in both genes seen in Figure 4 (Markowitz *et al.,* 2012). In both organisms, the *cas*A gene was found second in the operon sequence. These maps provide a visual representation of the similarities of type I-E CRISPR-Cas system in both organisms, helping to define an orthologous relationship.

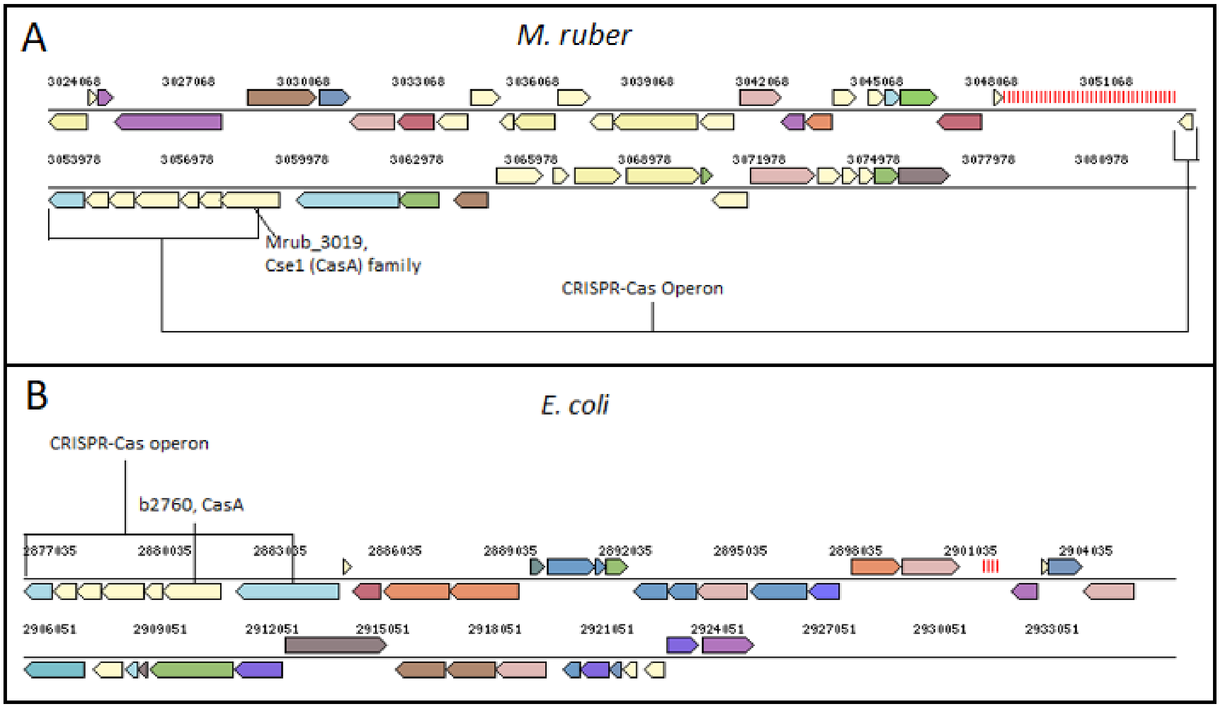


Figure 8. Chromosome maps of the CRISPR-Cas system in *M. ruber* and *E. coli* with labeled genes of interest collected on IMG/M (Markowitz *et al.,* 2012). **(A).** The chromosome map with Mrub_3019 gene tag and its protein named indicates this gene is involved in the CRISPR-Cas operon. **(B).** *E. coli* b2760 is tagged on the chromosome map and is also found within the CRISPR-Cas operon. IMG/M bioinformatics tools were available at https://img.jgi.doe.gov/cgi-bin/m/main.cgi.

CONCLUSION

To further examine the CRISPR-Cas system in bacteria, the protein components must be isolated for study. The *cas*A gene is one of the Cascade components involved in PAM recognition and signals for Cas3-mediated protein degradation. To understand each individual component allows for better understanding of the complex as a whole. *E. coli* was used as the model organism to compare its casA gene to *M. ruber's.* An orthologous relationship of the two genes would signify that the two organisms are evolutionarily related and fills in gaps of the phylogenetic tree.

After comparing two amino acid sequences using various bioinformatics tools and collecting database information, we are confident that Mrub_3019 is an ortholog of b2760. The BLAST alignment between the two proteins produced an E-value well below the cot-off, thereby suggesting that Mrub_3019 and *E. coli casA* have a common ancestor. Bioinformatics data was collected from different bioinformatics programs like TMHMM, PRED, PSORT-B, Pfam, PDB and databases like KEGG, IMG/M, GenBank, and EcoCyc. Chromosome maps from IMG/M confirmed that both genes were part of a CRISPR-Cas operon which are important indicators of gene function or genes encoding proteins. Conserved genes and gene order in an operon among organisms is a strong indication of functional gene (Nunez *et al.,* 2013) .

TMHMM, PRED, and PSORT-B provided strong evidence that neither proteins are found in the cell membrane. Predicting protein motifs and cellular location provides powerful information on structure-specific functions of the gene. The CDD pulled different COG numbers but they had similar COG names and low E-values which provides evidence for the claim of an orthologous relationship between Mrub_3019 and b2760. This indicates that they may have once been closely related but structural changes over time lead to the organisms traveling genetically further apart. Both TIGRFAM and Pfam pulled the same top hits for both proteins, suggesting that Mrub_3019 and CasA have the same domain and have sequence similarity to the same protein family, respectively. TIGR02547 is a CRISPR system Cascade complex protein that is found in both organisms, which supports the hypothesis that Mrub_3019 gene is an b2760 ortholog. Crystallized quaternary protein structures of the genes were compared by examining the closest hit to the genes on PDB. The 3D structures revealed variations in size and complexity, but still maintained an overall common orientation.

There is strong evidence to suggest that Mrub_3019 is an ortholog of b2760 by their similar functions and structures as shown by various bioinformatics tools. This study provides information of the casA gene of the CRISPR-Cas system in *M. ruber* in hopes to fill the gaps that may form a clear phylogenetic relationship between bacteria.

References

Amlinger, L. (2017). The type I-E CRISPR-Cas system: Biology and application of an adaptive immune system in bacteria. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 1466.

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

Bagos PG, Liakopoulos TD, Spyropoulos IC & Hamodrakas SJ. (2004) A Hidden Markov

Model method, capable of predicting and discriminating beta-barrel outer membrane. *Proteins. BMC Bioinformatics*, 15;5(29). Available at http://bioinformatics.biol.uoa.gr//PRED-TMBB/process.jsp

Finn, R.D., Coggill, 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.

Hochstrasser, M.L., Taylor, D.W., Bhat, P., Guegler, C.K., Sternber, S.H., Nogales, E., & Doudna, J.A. (2014). CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. *PNAS, 111,* 6618-6623.

Horvath, P. & Barrangou, R. (2010). CRISPR/Cas, the immune system of bacteria and archaea. *Science, 237*, 167-170.

Jackson, R.N., Golden, S.M., van Erp, P.B., Carter, J., Westra, E.R., Brouns, S.J., van der Oost, J., Terwilliger, T.C., Read, R.J., & Wiedenheft, B. (2014) Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. Science 345: 1473-1479.

Juncker, A., H. Willenbrock, G. von Heijne, H. Nielsen, S. Brunak, & A. Krogh (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci*. 12(8):1652-62 [2016 Dec 6]. Available at: http://www.cbs.dtu.dk/services/LipoP/

Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K., &Tanabe, M. (2019) New approach for understanding genome variations in KEGG. *Nucleic Acids Res*. 47, D590-D595

Kanehisa, Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res*. 45, D353-D361.

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

Keseler, I.M et al., (2013). EcoCyc: fusing model organism databases with systems biology \ *Nucleic Acids Research* 41:D605-612.

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

Madden T.(2002), The BLAST Sequence Analysis Tool. [Updated 2003 Aug 13]. In: McEntyre J, Ostell J, editors. The NCBI Handbook. Bethesda (MD): National Center for Biotechnology Information (US); 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 (2016) . 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

Sashital, D.G., Wiedenheft, B., & Doudna, J.A. (2012). Mechanism of foreign DNA selection in a bacterial adaptive immune system. *Mol Cell* 46: 606-615.

Shinohara, T., Ikawa, S., Iwasaki, W., Hiraki, T., Hikima, T., Mikawa, T., … & Shibata, T. (2015). Loop L1 governs the DNA-binding specificity and order for RecA-catalyzed \ Reaction In homologous recombination and DNA repair. *Nucleic Acids Research, 43*(2), 973-986.

Tindell, B.J., Sikorski, J., Lucas, S., Goltsman, E., Copeland, A., Galvina Del Rio, T., … &

Lapidus, A. (2010). Complete genome sequence of *Meiothermus ruber* type strain (21^T). *The Genomic Standards Consortium, 3,* 26-36.

Westra, E.R., Semenova, E., Datsenko, K.A., Jackson, R.N., Wiedenheft, B., Severinov, K., & Brouns, S.J. (2013) Type I-E CRISPR-cas systems discriminate target from non-target DNA through base pairing-independent PAM recognition. *PLoS Genet* 9: e1003742.

Wright, A.V., Nunez, J.K., & Doudna, J.A. (2016). Biology and application of CRISPR systems: Harnessing nature's toolbox for genome engineering. *Cell, 164*, 29-44.

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.