2019

Mrub_3014 is orthologous to b2756

Samir Abdelkarim
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Samir Abdelkarim and Dr. Lori Scott

Mrub_3014 is orthologous to b2756

Abstract:

This project is part of the *Meiothermus ruber* genome analysis project, which uses a collection of online bioinformatics tools to predict gene function. We investigated the biological function of the gene Mrub_3014, which we hypothesize is a component of the CRISPR-Cas prokaryotic defense system. We predict that Mrub_3014 (DNA coordinates 3054943..3055575) encodes CRISPR-associated protein Cse3/case which function as endonuclease. Our hypothesis is supported by identical hits for Mrub_3014 and b2756 to the KEGG, Pfam, TIGRfam, CDD and PDB databases, as well as a low E-value for a pairwise NCBI BLAST comparison. Both protein products are predicted to be localized to the cytoplasm. Finally, both proteins share numerous highly conserved amino acids, when compared to the consensus sequence of Pfam domains.

Key words for everyone: *Meiothermus ruber*, genome, bioinformatics, annotation, CRISPR-Cas, prokaryotic defense, Mrub_3014, b2756, casE

Introduction:

Little is known about thermophilic organisms and how they adept to extreme temperature (Scott, personal communication). The *Meiothermus Rube*r (*M. ruber*) genome analysis project is centered on analyzing the biological processes and temperature adaptations of the thermophile e *M. ruber*, in part by comparing the *M. ruber* genome to a well-studied microbe as *E. coli.* We to focus our study on *Meiothermus genus because* little is known about the organisms that belongs *Meiothermus genus*. Among all the organisms in the *Meiothermus genus* we specially chose *Meiothermus ruber* because it is one of the organisms that had its genome completely sequenced (Tindall et al, 2010). Major database search engines such as PubMed has very few published articles about *M. ruber*. Infect, a PubMed survey in October of 2014 identified only 41 published articles about *M. ruber* while there are about 366688 entries about *E. coli*. For that reason, *M. ruber* is an excellent candidate to investigate its genome and the functionality of these genes. Other organisms such as *E. coli* and Salmonella are great organisms to use as a model-systems because of the extensive research done on these organisms. In this project, *Escherichia coli* K-12 MG1655 (*E. coli*) was used as a model organism to compare its genome to *M. ruber*. Comparison of the DNA sequences between organisms can determine if there is a close relationship between them (Liu et al., 2015). To assist with this project, we utilized resources provided by the U.S Joint Genome Institute Genomic Encyclopedia of the Bacteria and Archaea (GEBA) project, which provided a final draft of the M. *ruber'* sequenced genome. The goal of the GEBA project is to fill in the gaps in our knowledge of bacteria and archaea (Wu *et al.,* 2009; Tindall *et al.,* 2010). It is hoped that this project will provide novel gene discoveries and biochemical processes (Wu et al., 2009; Tindall et al., 2010).

Study organism:

M. ruber is one of the eight known organisms within *Meiothermus genus* (Euzeby, 1997). *Meiothermus Rube* is a thermophilic organism that grows and thrives in a high-temperature environment such as hot spring. According to Loginova and colleagues*, M. ruber* was isolated from hot-spring in the Kamchatka Peninsula (Loginova *et al.,* 1975). The organism's optimal growth temperature is 60°- Celsius and growth range is 35°-70°Celsius. *M. ruber* is red pigmented, rod-shaped, gram-negative eubacteria (Tindall *et al.,* 2010). The organism is named *Meiothermus ruber* because the word "meion" means "lesser," the word "thermus" meaning "hot," and the "ruber" which means "red" describes the color of the pigment the organism produces. The eubacteria belong to the bacteria phylum Peinococcus and part of the Meiothermus genus (Tindall *et al.,* 2010).

CRISPR-Cas system:

In this project, a mechanism of adaptive immunity in bacteria known as the Clustered Regulatory Interspaced Short Palindromic Repeats-CRISPR-associated (CRISPR-Cas) system was studied by investigating the similarities between *M. ruber* and *E. coli*. A survey of prokaryotes found that 50 percent of bacteria, which includes *E. coli, contains* the CRISPR-Cas system and 90% of archaeal species contains the CRISPR-Cas system (Makarova *et al.,* 2015). The CRISPR-Cas systems are defense mechanisms for bacteria against invading nucleic acid from bacteriophage and plasmid (Jiang *et al.,* 2015).

In brief, the CRISPR-Cas System produces small CRISPR RNAs (crRNA) consisting of repetitive sequences known as palindromic repeats separated by unique sequences known as spacer sequences. The spacer sequences are remnants of DNA from previous invaders that serve as complementary sequences to regions of genomic DNA within an infective agent. Upstream of the spacer sequences and the palindromic repeats are CRISPR associated genes that code for cas proteins. The proteins normally function as helicase and nuclease which unwind and cleave to degrade foreign DNA (Barrangou et al., 2007; Medina-Aparicio et al., 2011). The CRISPR-cas system genes are transcribed in an operon (Medina-*Aparicio et al., 2011).* An operon is a unit of genes that are transcribed or controlled by a single promoter.

The CRISPR-Cas system consists of three stages; spacer acquisition or adaptation, CRISPR RNA (crRNA) biogenesis, and interference. Figure 1 summarizes the different stages. In the spacer acquisition also called the adaptation stage, invading bacteriophage genome enters the bacteria and a short protospacer sequence from the bacteriophage genome is recognized and incorporated into the CRISPR array as new spacer serving (as a genetic memory). The protospacer sequence from the bacteriophage genome is recognized by a specific recognition spacer know as protospacer adjacent motifs (PAMs) that is present on invading phage genome (Jiang *et al.,* 2015; Wright *et al.;* 2016). The next stage is the CRISPR RNA (crRNA) biogenesis in which the CRISPR array is transcribed with the newly incorporated spacer into a single pre-crRNA. The pre-crRNA gets processed and cleave into a mature crRNAs with a single spacer. In the final stage which is called interference, crRNA recognizes complementary sequences in the invading DNA and bind to it. The recognition of complementary sequences on target invading

DNA recruits cas proteins transcribed by the cas genes to cleave the invading DNA and essentially degrade the DNA (Wright *et al.;* 2016).

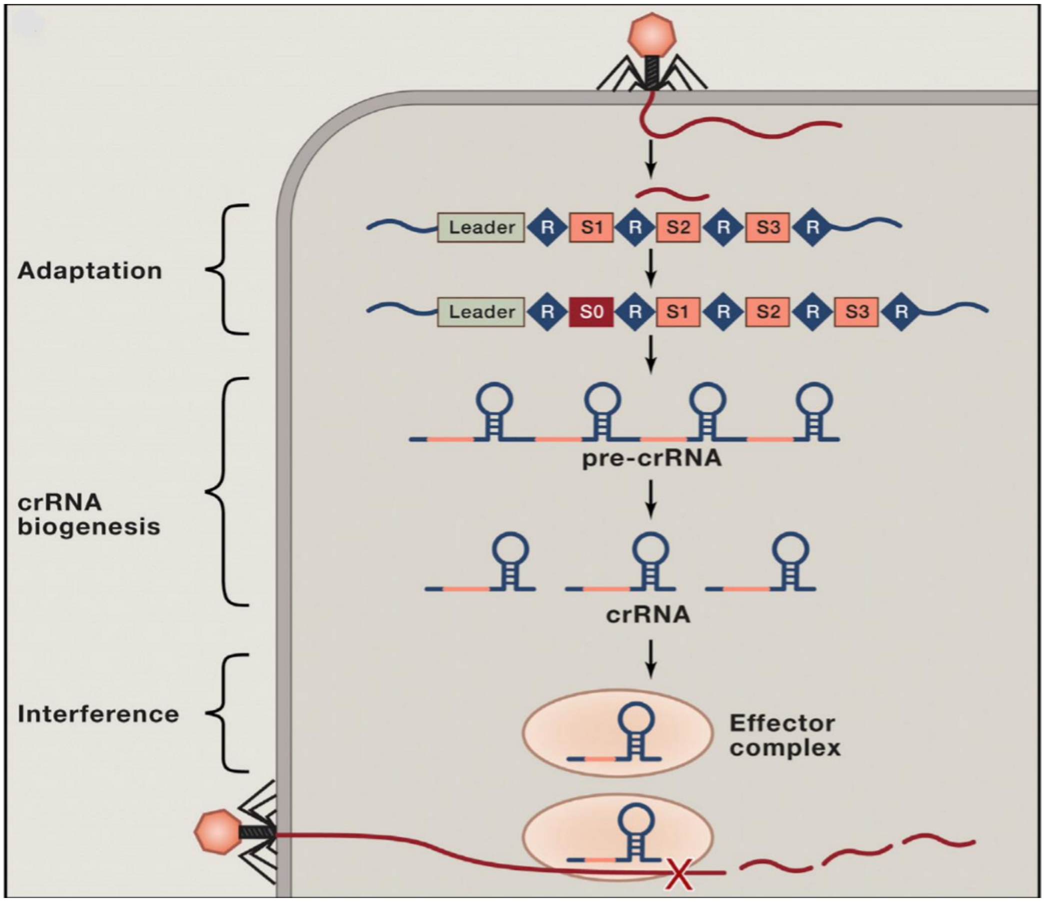


Figure 1: Stages of the CRISPR/Cas System

CRISPR/Cas system contains three stages: acquisition/adaptation, crRNA biogenesis, and interference. In the acquisition/adaptation stage, the bacteriophage's protospacer sequence is recognized by its protospacer adjacent motif (PAM) and incorporated into the CRISPR array region of the CRISPR/Cas as a new spacer (R are palindromic repeats, S1,2, and 3 are spacers from previous invaders, and S0 is the spacer from the new bacteriophage).In the crRNA biogenesis stage, the CRISPR/Cas locus with the new spacer gets transcribed, processed and cleaved into single crRNA. In the interference stage, the crRNA targets foreign DNA with complementarity spacer sequences and presents it to effector complexes (which are the cas proteins) to degrade the foreign DNA (Makarova *et al.,* 2015; Shmakov *et al.,* 2015)

The cas genes of the CRISPR-Cas system are classified into 6 types based on "signature genes". Types I-III are the most well-studied, while types IV-VI are still been farther investigated. These 6 types are farther categorized into two-classes system based on the type of effector complex: class 1, which includes types I, III and IV are known to have multi-subunit effector complex while the class two system, which includes type II, V and VI are known for having a single subunit. The type I system (which also contains subtypes 1A-E) contains the signature *cas3* gene, which encodes a protein with nuclease and helicase, participates in foreign DNA degradation. The type II system is known for its signature *cas9* genes which encode a protein that interferences with foreign DNA. Type III is recognized by its signature gene *cas10*. The *cas10* encode for proteins that search for foreign DNA and destroy it (Makarova *et al.,* 2011b; Makarova *et al.,* 2015; Shmakov *et al.,* 2015). Type IV system which is one of the less studied types contains *csf1* (colony stimulating factor 1) gene (Makarova and Koonin, 2015). Another less studied type is the type V system which contains several signature genes (either *cpf1*, *c2c1*, or *c2c3*) all of whom are similar to cas9 (Shmakov *et al*., 2015; Zetsche *et al*., 2015a). Finally, type VI system is known for its signature gene, *csc2* gene, which is predicted to encode for a protein with two HEPN (higher eukaryotes and prokaryotes nucleotide-binding) RNase domains (Shmakov *et al.,* 2015).

Purpose/Hypothesis:

In this project, we learned that the model organism *E. coli* contains CRISPR-Cas type 1 system. Using Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al*, 2019) which is a database that provides genes and proteins of many different organisms as well as biochemical processes and cellular structures, we discovered that both *E. coli* contains signature *cas3* gene which meant that their CRISPR-Cas system is type I. Within the *E. coli* CRISPR-Cas system genes, we chose to study on *casE* (b2756) gene.

According to KEGG, *casE* which is also named *cas6e*, *cse3*, and *ygcH,* is a CRISPR associated gene that encode for a protein that is involved in endonuclease. The recommended name of *casE*'s protein is CRISPR system Cascade subunit CasE, CasE endoRNase or simply just casE. According to EcoCyc (Keseler *et al,* 2013), one of the bioinformatics websites used, *casE* is necessary for cleave and processing of pre-cRNA into single repeat-spacer unit (palindromic repeats). Besides having endonuclease activity, *casE* holds the CRISPR-Cas cascade complex by connecting to *casB* and *casC.* The *E. coli* CRISPR associated genes such as casE, casB, and casC are all part of an operon control by the promoter casAP. Figure 2 provides an image of the operon; the dark purple highlighted gene is E. coli's gene of interest (GOI) *casE.* Furthermore, E. coli's *casE* (b2756) must be part of an operon because all the neighboring genes and itself are lined in the same direction. Based on the provided information, we proposed that *Mrub_3014* is orthologous to *E. coli b2756 (*CasE) gene. An ortholog is two or more genes from different organisms that have evolved due to common ancestral gene.

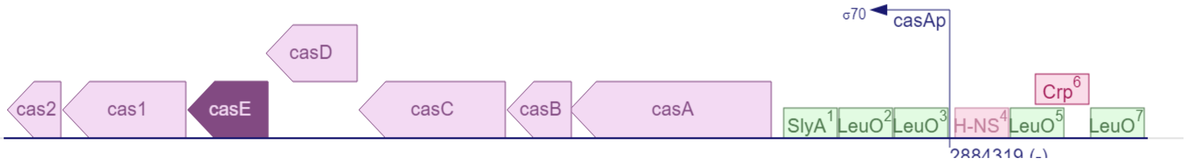


Figure 2. EcoCyc image depicts that CasE (b2756) is part of an operon system**.** This conclusion was reached because all gene are lined in the same direction and they are all control by a single promoter which is the casAP. https://ecocyc.org/gene?orgid=ECOLI&id=G7430-MONOMER

Materials and Methods:

To confirm or disprove our hypothesis, we used a variety of bioinformatics databases to collect comparison data on putative orthologs of *casE* in *M. ruber* and *E. coli.* We first used the bioinformatics tool EcoCyc (Keseler *et al,* 2013) to learn more about the CRISPR-Cas system in our model organism *Escherichia coli* K-12 MG1655. EcoCyc is scientific database dedicated to the prokaryote *Escherichia coli* K-12 MG1655, which includes extensive literature-based curation of the many processes known to occur in this bacterial strain. Next, we used the KEGG database (Kanehisa et al., 2017; and Kanehisa et al., 2000) and the IMG/M database (Markowitz *et al.,* 2012) to confirm that *E. coli* K-12 MG1655 had the CRISPR-Cas system, and to determine if *M. ruber* was predicted to have any of the known components of the CRISPR-Cas system. The *M. ruber* gene Mrub_3014 was chosen for this project, which we will compare the putative ortholog in E. coli *cse3* (*aka casE*).

The NCBI BLAST tool (Madden T *et al*, 2002) was used to perform a protein BLAST of using the *E. coli* casE amino acid sequence as the query against the putative Mrub_3014 as the subject sequence. BLAST is a database that aligns base sequences or amino acid sequences to determine their degree of similarity by producing a quantifiable value called the E-value. An E value indicates if two sequences are similar due to evolutionary relatedness (which is represented by lower E-value) or due to random chances (which is represented by higher E-value). The E-value cut-off of 0.001 is used in this study; E-values below 0.001 are interpreted as demonstrating a significant level of sequence similarity, and therefore, functional similarity. From BLAST database, we properly identified the start codon for mrub_3014, which was GTG. The start codon GTG was different from the typical start codon ATG which is the case E. coli casE. Both the start codon encodes for methionine. This is further explained in the results section.

Three bioinformatics tools were then used to determine where Mrub_3014 functions in the cell. TMHMM database (Krogh *et al,* 016) predicts if proteins have transmembrane alpha helices, which suggests that a protein might be membrane-embedded or pass through the membrane. PRED-B database (Bagos *et al.,* 2005) predicts if the protein creates a membrane-embedded beta-barrel. PSORT-B (Yu *et al,*2010) provides a prediction of five subcellular localizations for Gram-negative bacteria by combining several methods, including homology analysis, identification of sorting signals and other motifs.

In the third phase of the project, Mrub_3014 and *E. coli* CasE (locus tag b2756) were analyzed for functional similarities using the tools CDD, TIGRFAM, PFAM and PDB. Conserved Domain Database (CDD) analyzed both sequences for conserved domains, which are independently functioning units within a protein sequence. TIGRFAM (Haft *et al,* 2001) assigns proteins to proteins families, groups of proteins with similar function, based on similarity to a family consensus sequence. PFAM (Finn et al, 2016) analyzes an amino acid sequence for protein domains, families or clans within its database. Finally, we used PDB database (Berman et al, 2000) to see if Mrub_3014 and CasE/b2756 have significant sequence similarity to proteins within this highly curated database of crystalized proteins.

Results:

We started our experiment by first using KEGG database to compare *M. ruber* Mrub_3014 with *E. coli* b2756. Table 1 shows the KEGG results for Mrub_3014 and b2756. The data displays that both Mrub_3014 and b2756 are assigned the same gene name which are *casE, cse3*, and *cas6e* and protein name CasE. The two genes contain a similar sequence and their sequence length is also close to each other.

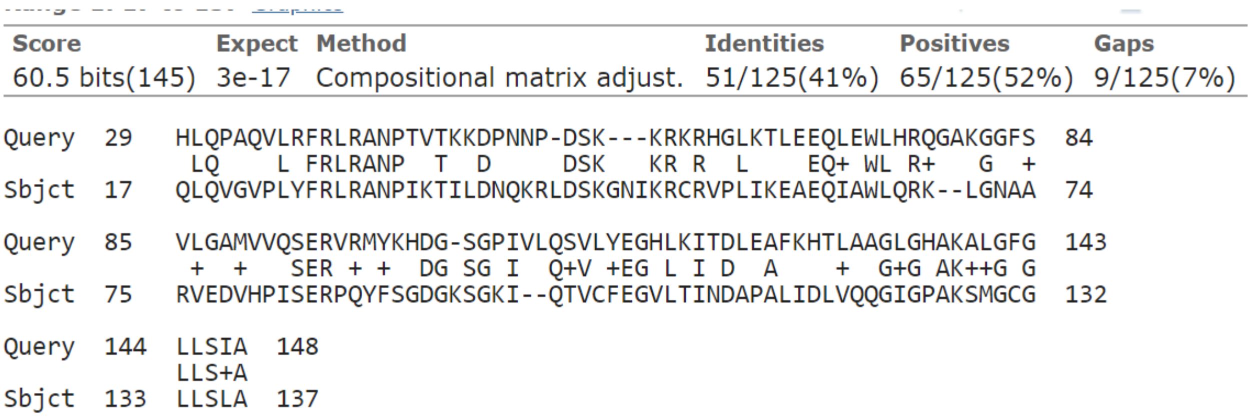
Table 1. KEGG data for Mrub_3014 and *E. coli* b2756. The data in the table depicts similarities between the two genes.

DNA Nucleotide sequences	>mrb:Mrub_3014 K19126 CRISPR system Cascade subunit CasE gtgtacctgagccgactccagcttgatccccgctctaagcaggcccgcaccgacctggcc agcccctatgagctgcacgccaccctgtgccatgcctttgccgggcccaatcagacccca gcgcgctttttgtggcgggctgaggtaggaaaaatccccattgtgctggtgcagagtgcc gggatgccggactgggaaaaattggtccagcgtttccccggctactttgcccagcccccca gcctccaaacccatccccctcgagcacctccagcctgcccaggtgctgcgctttcgccta cgtgctaaccctactgtgaccaaaaaagatcccaacaatcctgatagcaaaaagcgcaag cgccacggtttgaaaaccctcgaagagcagctcgagtggctgcatcgccagggagccaaa gggggcttctcggtgctgggcgcgatggtggttcagagcgagcgggtgcgcatgtacaaa cacgacggctccggcccgattgtgcttcagtcggtgctgtacgaggggcatctgaagatc accgacctcgaggctttcaaacacaccctggctgctggcctgggccacgccaaagccctg ggttttggcctgctttccatcgcaaaggtgtag	>eco:b2756 K19126 CRISPR system Cascade subunit CasE atgtatctcagtaaagtcatcattgccagggcctggagcagggatctttaccaacttcac cagggattatggcatttatttccaaacagaccggatgctgctcgtgattttctttttcat gttgagaagcgaaacacaccagaaggctgtcatgttttattgcagtcagcgcaaatgcct gtttcaactgccgttgcgacagtcattaaaactaaacaggttgaatttcaacttcaggtt ggtgttccactctattttcggcttcgggcaaatccgatcaaaactattctcgacaatcaa aagcgcctggacagtaaagggaatattaaacgctgtcgggttccgttaataaaagaagca gaacaaatcgcgtggttgcaacgtaaattgggcaatgcggcgcgcgttgaagatgtgcat cccatatcggaacggccacagtatttttctggtgatggtaaaagtggaaagatccaaacg gtttgctttgaaggtgtgctcaccatcaacgacgcgccagcgttaatagatcttgtacag caaggtattgggccagctaaatcgatgggatgtggcttgctatctttggctccactgtga
Nucleotide Sequence Length	633 nt	600 nt
Amino Acid Sequence	**>mrb:Mrub_3014 K19126 CRISPR system Cascade subunit CasE** **MYLSRLQLDPRSKQARTDLASPYELHATLCHAFAGPNQTPARFLWRAEVGKIPIVLVQSA** **GMPDWEKLVQRFPGYFAQPPASKPIPLEHLQPAQVLRFRLRANPTVTKKDPNNPDSKKRK** **RHGLKTLEEQLEWLHRQGAKGGFSVLGAMVVQSERVRMYKHDGSGPIVLQSVLYEGHLKI** **TDLEAFKHTLAAGLGHAKALGFGLLSIAKV**	**>eco:b2756 K19126 CRISPR system Cascade subunit CasE** **MYLSKVIIARAWSRDLYQLHQGLWHLFPNRPDAARDFLFHVEKRNTPEGCHVLLQSAQMP** **VSTAVATVIKTKQVEFQLQVGVPLYFRLRANPIKTILDNQKRLDSKGNIKRCRVPLIKEA** **EQIAWLQRKLGNAARVEDVHPISERPQYFSGDGKSGKIQTVCFEGVLTINDAPALIDLVQ** **QGIGPAKSMGCGLLSLAPL**
Amino Acid Sequence Length	210 aa	199 aa

https://www.kegg.jp/dbget-bin/www_bget?eco:b2756,

https://www.kegg.jp/dbget-bin/www_bget?mrb:Mrub_3014

Using BLAST, we compared the amino acid sequences between Mrub_3014 and *E. coli* b2756. The protein BLAST comparison between Mrub_3014 and *E. coli* b2756 is shown in figure 3. The BLAST alignment has an E. value of 3e-17, which suggests that Mrub_3014 is orthologous *to E. coli* b2756. A low E-value indicates that the likelihood that the Mrub_3014 sequence similarity to b2756 is not by chance, but due to their similar function derived from a common ancestor.



Score	Expect	Method	Identities	Positives	Gaps
60.5 bits(145)	3e-17	Compositional matrix adjust.	51/125(41%)	65/125(52%)	9/125(7%)

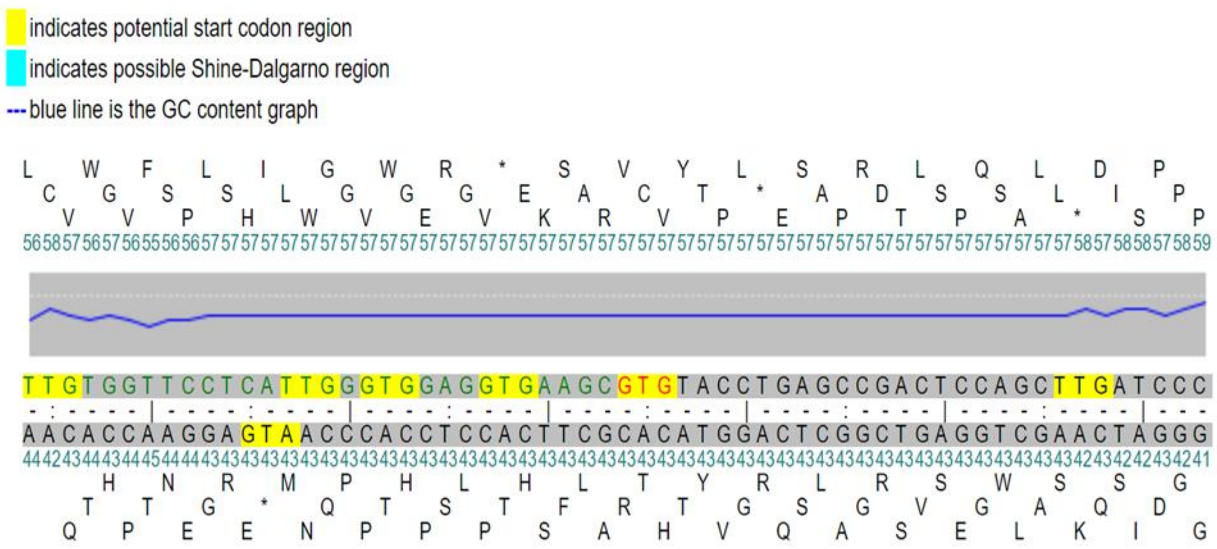
```
Query  29   HLQPAQVLRFRLRANPTVTKKDPNNP-DSK---KRKRHGLKTLEEQLEWLHRQGAKGGFS  84
             LQ    L FRLRANP  T  D     DSK   KR R  L    EQ+ WL R+   G  +
Sbjct  17   QLQVGVPLYFRLRANPIKTILDNQKRLDSKGNIKRCRVPLIKEAEQIAWLQRK--LGNAA  74

Query  85   VLGAMVVQSERVRMYKHDG-SGPIVLQSVLYEGHLKITDLEAFKHTLAAGLGHAKALGFG  143
             +  +   SER + +  DG SG I  Q+V +EG L I D  A    +  G+G AK++G G
Sbjct  75   RVEDVHPISERPQYFSGDGKSGKI--QTVCFEGVLTINDAPALIDLVQQGIGPAKSMGCG  132

Query  144  LLSIA  148
            LLS+A
Sbjct  133  LLSLA  137
```

Figure 3. Protein BLAST: Protein BLAST of Mrub_3014 against *E. coli* gene b2756. The query sequence is Mrub_3014 and subject sequence is *E. coli* gene b2756. Amino acids that are identical are shown the middle between the two sequences using it's abbreviation, "+" sign means that the amino acids are chemically similar, and the dashed lines "-"represents missing amino acids. The E-value for the sequences is 3e-17 suggesting Mrub_3014 is orthologous *to E. coli* b2756. http://www.ncbi.nlm.nih.gov/blast

To confirm that we are using the correct protein sequence in our analysis of Mrub_3014, we used IMG/M database to analyze the start codon of Mrub_3014. The open reading frame (ORF) from IMG/M is shown is figure 4. The ORF displayed that the start codon of mrub_3014 and a potential Shino-Dalgarno sequences (SDs). The figure displays the 5' upstream region of Mrub_3014 and its 6 reading frames. The nucleotide sequences highlighted in yellow are potential start codons; the start codon highlighted in red is the one predicted for mrub_3014. No Shine-Delgarno sequence is predicted for this region. In this case, GTG is the proposed start codon. None of the other potential start codons are in the same reading frame as the GTG. Consequently, there is no better alternative start codon in this region.



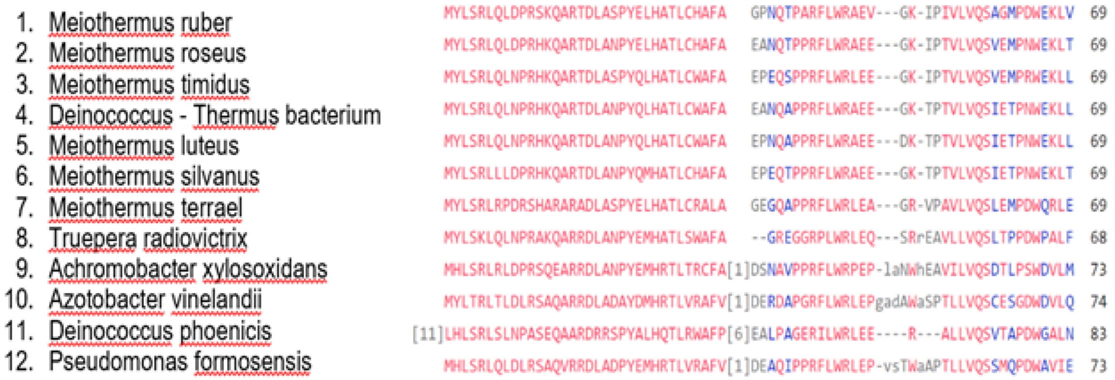
```
▮ indicates potential start codon region
▮ indicates possible Shine-Dalgarno region
--- blue line is the GC content graph

L  W  F  L  I  G  W  R  *  S  V  Y  L  S  R  L  Q  L  D  P
 C  G  S  S  L  G  G  G  E  A  C  T  *  A  D  S  S  L  I  P
  V  V  P  H  W  V  E  V  K  R  V  P  E  P  T  P  A  *  S  P
565857565756555656575757575757575757575757575757575757575757575757575757575757575757575757575757575757575757575758575858575859

TTGTGGTTCCTCATTGGGTGGAGGTGAAGCGTGTACCTGAGCCGACTCCAGCTTGATCCC
-:----|----:----|----:----|----:----|----:----|----:----|---
AACACCAAGGAGTAACCCACCTCCACTTCGCACATGGACTCGGCTGAGGTCGAACTAGGG
444243444344454444434343434343434343434343434343434343434343434343434343434343434343434343434343434343434343434342434242434241
    H  N  R  M  P  H  L  H  L  T  Y  R  L  R  S  W  S  S  G
   T  T  G  *  Q  T  S  T  F  R  T  G  S  G  V  G  A  Q  D
  Q  P  E  E  N  P  P  P  S  A  H  V  Q  A  S  E  L  K  I  G
```

Figure 4. 5' upstream region of Mrub_3014 with all 6 reading frames translated into the single-letter amino acid abbreviations. https://img.jgi.doe.gov/cgi-bin/m/main.cgi

A multiple sequence alignment created by NCBI BLAST using Mrub_3014 as the query sequence was also performed to confirm that we were using the correct start codon. Figure 5 shows the results of this alignment. All twelve proteins show strong sequence alignment to the amino end of the protein. Mrub_3014 does not extend in either direction beyond the other sequences. This suggests, as does the previously described IMG/M output (Figure 4) that we are using the correct start codon for mrub_3014.

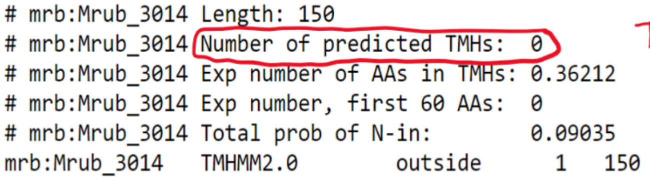


```
1.  Meiothermus ruber                      MYLSRLQLDPRSKQARTDLASPYELHATLCHAFA   GPNQTPARFLWRAEV---GK-IPIVLVQSAGMPDWEKLV  69
2.  Meiothermus roseus                     MYLSRLQLDPRHKQARTDLANPYELHATLCHAFA   EANQTPPRFLWRAEE---GK-IPTVLVQSVEMPNWEKLT  69
3.  Meiothermus timidus                    MYLSRLQLNPRHKQARTDLASPYQLHATLCWAFA   EPEQSPPRFLWRLEE---GK-IPTVLVQSVEMPRWEKLL  69
4.  Deinococcus - Thermus bacterium        MYLSRLQLNPRHKQARTDLANPYQLHATLCWAFA   EANQAPPRFLWRAEE---GK-TPTVLVQSIETPNWEKLL  69
5.  Meiothermus luteus                     MYLSRLQLNPRHKQARTDLANPYQLHATLCWAFA   EPNQAPPRFLWRAEE---DK-TPTVLVQSIETPNWEKLL  69
6.  Meiothermus silvanus                   MYLSRLLLDPRHKQARTDLANPYQMHATLCHAFA   EPEQTPPRFLWRAEE---GK-TPTVLVQSIETPNWEKLT  69
7.  Meiothermus terrael                    MYLSRLRPDRSHARARADLASPYELHATLCRALA   GEGQAPPRFLWRLEA---GR-VPAVLVQSLEMPDWQRLE  69
8.  Truepera radiovictrix                  MYLSKLQLNPRAKQARRDLANPYEMHATLSWAFA   --GREGGRPLWRLEQ---SRrEAVLLVQSLTPPDWPALF  68
9.  Achromobacter xylosoxidans             MHLSRLRLDPRSQEARRDLANPYEMHRTLTRCFA[1]DSNAVPPRFLWRPEP-laNWhEAVILVQSDTLPSWDVLM  73
10. Azotobacter vinelandii                 MYLTRLTLDLRSAQARRDLADAYDMHRTLVRAFV[1]DERDAPGRFLWRLEPgadAWaSPTLLVQSCESGDWDVLQ  74
11. Deinococcus phoenicis             [11]LHLSRLSLNPASEQAARDRRSPYALHQTLRWAFP[6]EALPAGERILWRLEE----R---ALLVQSVTAPDWGALN  83
12. Pseudomonas formosensis                MHLSRLQLDLRSAQVRRDLADPYEMHRTLVRAFV[1]DEAQIPPRFLWRLEP-vsTWaAPTLLVQSSMQPDWAVIE  73
```

Figure 5. A multiple Sequence Alignment using mrub_3014 as the query sequence (sequence 1) shows good alignment at the amino terminus, suggesting that the correct start codon was called. Twelve sequences with strong sequence similarity to mrub_3014 (1, *Meiothermus ruber)* were pulled from the GenBank using NCBI BLAST and aligned together. The single-letter abbreviations of each amino acid sequence are shown. http://www.ncbi.nlm.nih.gov/blast

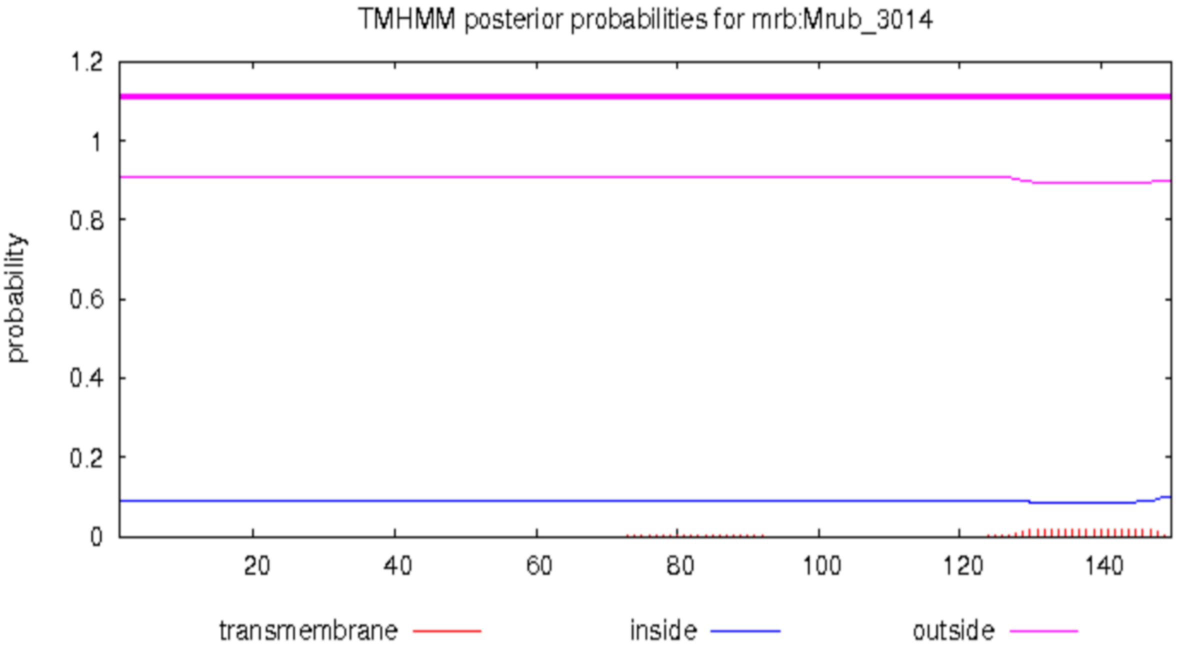
In the next phase of this project, we used 3 programs to predict the cellular localization for the mrub_3014 and b2756. Figure 6 provides results of the TMHMM analysis. Using the amino acid sequences for both Mrub_3014 (panel A) and b2756 (panel B) the database predicted the number of transmembrane helices in the proteins encoded by Mrub_3014 and b2756. In the topology graph, the dark pink line indicates that the protein is a transmembrane protein, the light pink line indicates that the protein is outside the cell, and blue line indicates that the protein is inside the cell. For both Mrub_3014 (panel A) and b2756 (panel B), the database calculated that the number of transmembrane helices is zero. This indicates that the protein for Mrub_3014 and b2756 must be localized in the cytoplasm.

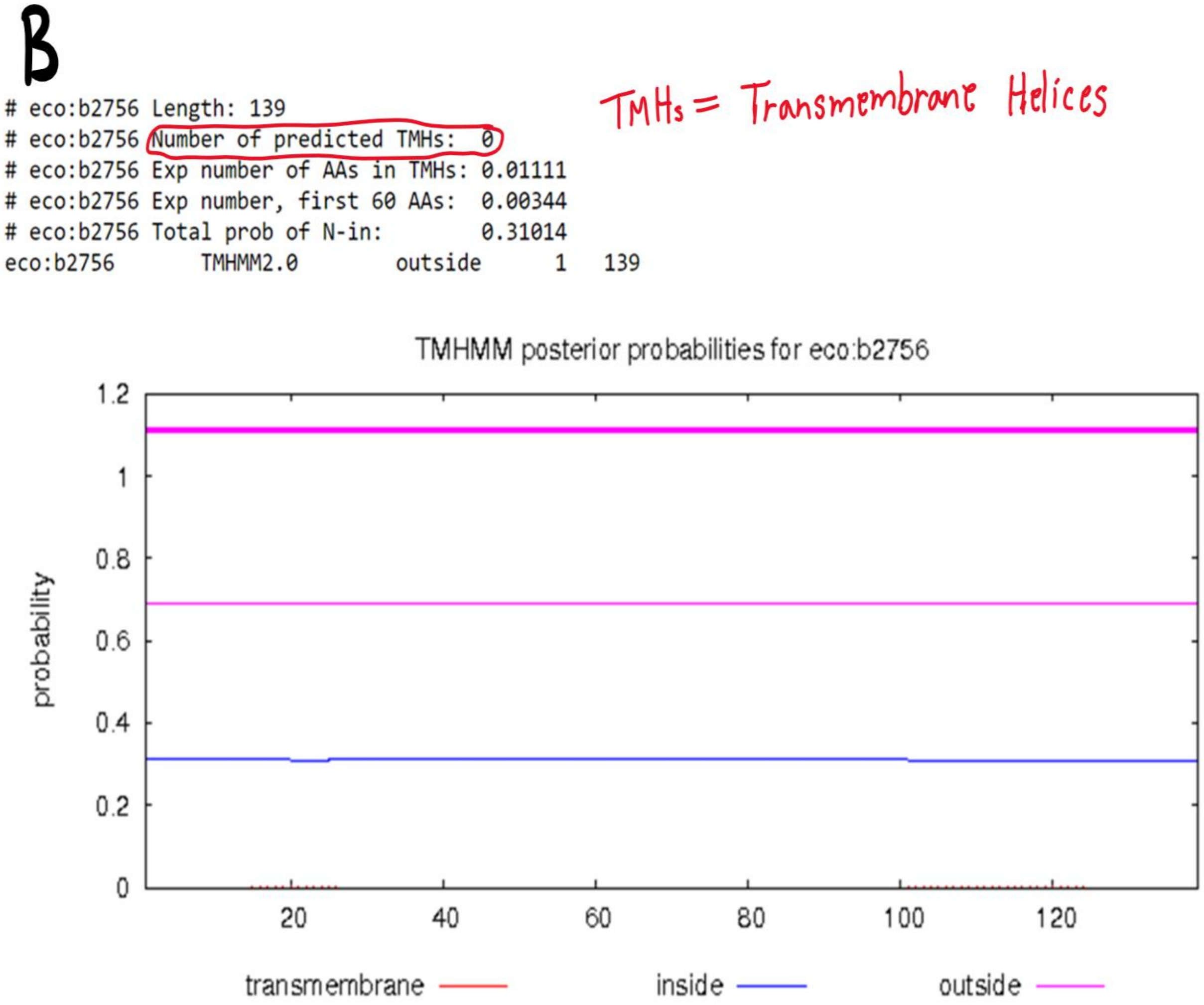
A



```
# mrb:Mrub_3014 Length: 150
# mrb:Mrub_3014 Number of predicted TMHs:  0
# mrb:Mrub_3014 Exp number of AAs in TMHs: 0.36212
# mrb:Mrub_3014 Exp number, first 60 AAs:  0
# mrb:Mrub_3014 Total prob of N-in:        0.09035
mrb:Mrub_3014   TMHMM2.0        outside      1   150
```

TMHs = Transmembrane Helices



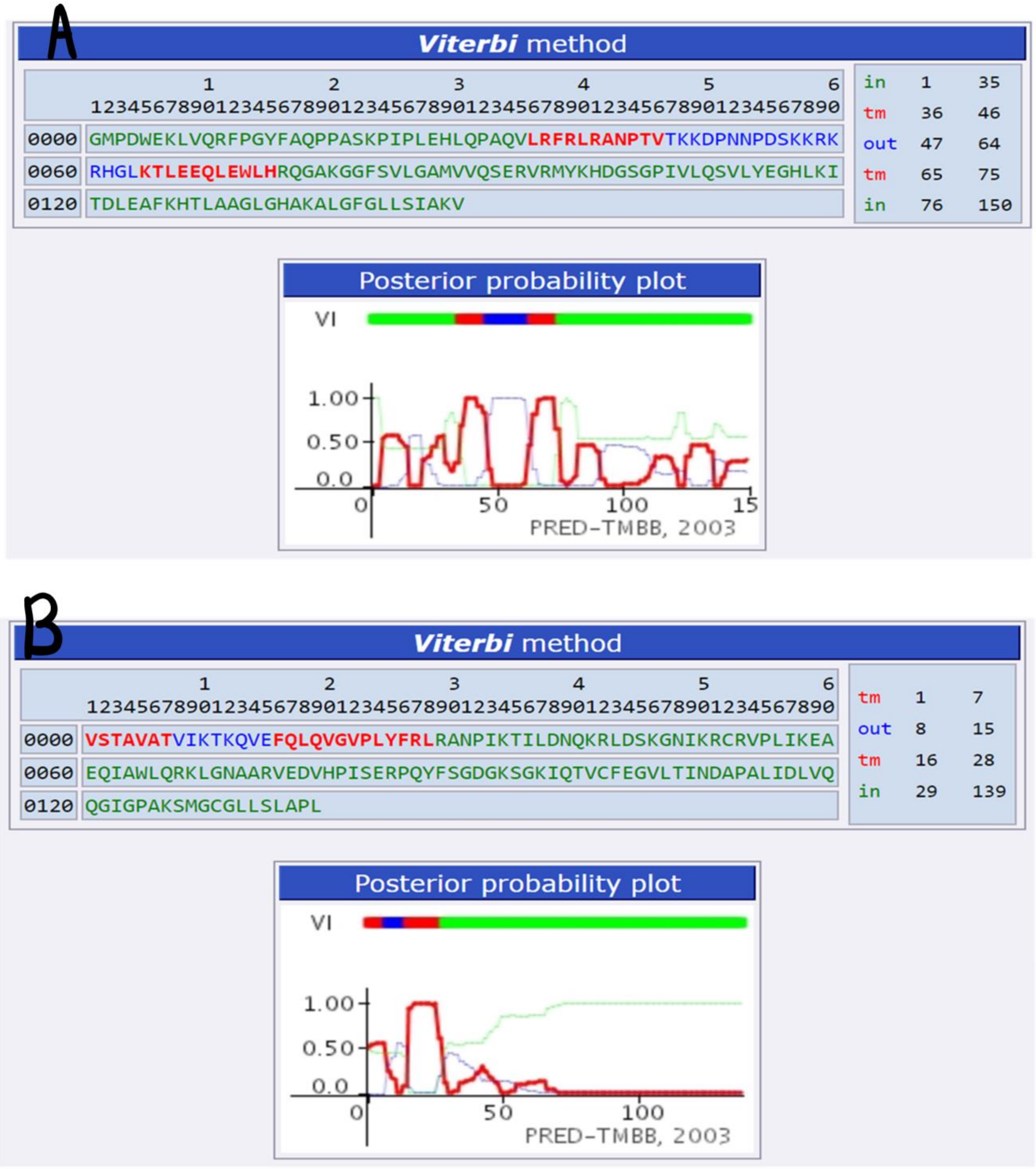


TMHs = Transmembrane Helices

```
# eco:b2756 Length: 139
# eco:b2756 Number of predicted TMHs:  0
# eco:b2756 Exp number of AAs in TMHs: 0.01111
# eco:b2756 Exp number, first 60 AAs:  0.00344
# eco:b2756 Total prob of N-in:        0.31014
eco:b2756       TMHMM2.0        outside      1   139
```

Figure 6. TMHMM results: Panel A is the results for Mrub_3014 and panel B is the results for b2756. The figure displays the probabilities of Mrub_3014 (*casE*) and b2756 (*casE*) are membrane-embedded alpha helices. The number of membrane-embedded alpha helices for both Mrub_3014 and b2756 is zero, which means that the protein encoded by the two do not participate in transmembrane helices meaning that the protein is located in the cell transmembrane. http://www.cbs.dtu.dk/services/TMHMM/

The PRED analysis determines if a protein folds into a beta-barrel outer membrane protein. On figure 7 (panel A and B), the Viterbi method above the probability plot displays that the red amino acid sequences suggest that the protein is a beta-barrel transmembrane protein, the blue amino acid sequences suggest that the protein is a beta-barrel outer membrane protein, and the green amino acid sequences suggest that the protein is not beta-barrel membrane protein and it is an intracellular protein. Base on the high percentage of the green amino acid sequences, the protein is not a beta-barrel membrane protein it is an intracellular protein.



Viterbi method

```
               1         2         3         4         5         6
      123456789012345678901234567890123456789012345678901234567890
0000  GMPDWEKLVQRFPGYFAQPPASKPIPLEHLQPAQVLRFRLRANPTVTKKDPNNPDSKKRK
0060  RHGLKTLEEQLEWLHRQGAKGGFSVLGAMVVQSERVRMYKHDGSGPIVLQSVLYEGHLKI
0120  TDLEAFKHTLAAGLGHAKALGFGLLSIAKV
```

```
in    1    35
tm    36   46
out   47   64
tm    65   75
in    76   150
```

Posterior probability plot

Viterbi method

```
               1         2         3         4         5         6
      123456789012345678901234567890123456789012345678901234567890
0000  VSTAVATVIKTKQVEFQLQVGVPLYFRLRANPIKTILDNQKRLDSKGNIKRCRVPLIKEA
0060  EQIAWLQRKLGNAARVEDVHPISERPQYFSGDGKSGKIQTVCFEGVLTINDAPALIDLVQ
0120  QGIGPAKSMGCGLLSLAPL
```

```
tm    1    7
out   8    15
tm    16   28
in    29   139
```

Posterior probability plot

Figure 7. PRED results: Panel A is the result for Mrub_3014 and panel B is the results of b2756. The figure predicts if Mrub_3014 and b2756 protein are beta-barrel membrane protein, beta-barrel outer membrane protein or not beta-barrel membrane protein. On the Viterbi method, the red amino acid sequence suggest that the protein is a beta-barrel transmembrane protein, the blue amino acid sequences suggest that the protein is a beta-barrel outer membrane protein, and the green amino acid sequences suggest that the protein is not beta-barrel membrane protein and it is an intracellular protein. Base on the high percentage of the green amino acid sequences, the protein is not a beta-barrel membrane protein it is an intracellular protein. http://bioinformatics.biol.uoa.gr/PRED-TMBB/input.jsp

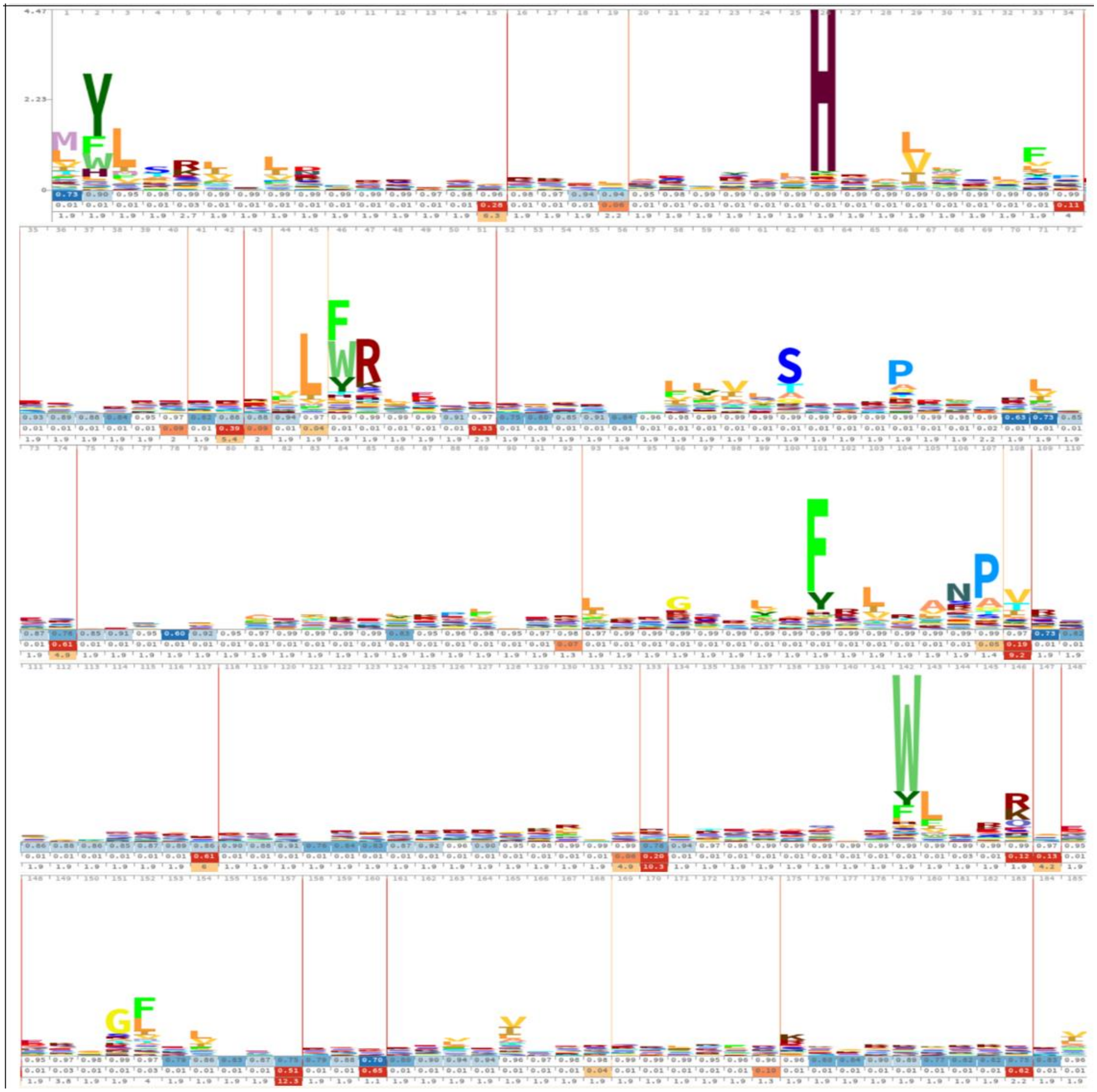
Table 2 compares the outcomes of the various bioinformatics tools used in this study between Mrub_3014 (*casE)* and its putative ortholog *E. coli b2756 (casE).* The first, is the comparison results from the protein BLAST. The second row comes from the CDD database. According to CDD database mrub_3014 and *E. coli* b2756 code for a protein known as CRISPR/Cas system-associated RAMP superfamily protein Cas6e, and both belong to the same superfamily, which is identified by the clan number c108497. An alignment to the consensus sequence of this superfamily produced low E-values, which indicates that the two sequences are similar base on evolutionarily and functional relatedness. The next row provides data from TIGRFAM database. TIGRFAM indicates that Mrub_3014 and *E. coli* b2756 have the same protein family (TIGRFAM number TIGR01907 and name casE_Cse3: CRISPR system CASCADE complex protein) and a low E-value suggesting that there is a strong sequence similarity. From the PDB database, Mrub_3014 and *E. coli* b2756 pulled two different PDB codes, but on further analyze the difference is the organism from which the protein was crystalized. The 3D protein matched to both *M. ruber* and *E. coli* sequences: 1) b2756 - PDB code 4DZD is the crystal structure of a CRISPR-associated protein Cas6e from *Escherichia coli* str. K-12; and 2) mrub_3014 - PDB code 3QRPPDB is the crystal tructure of *Thermus thermophilus* Cse3 bound to an RNA representing a product mimic complex. Both proteins belong to Chain A. However, the most crucial data were the low E-value indicating that the genes are indeed similar. Finally, the last row is the results from Pfam. Mrub_3014 and *E. coli* b2756 showed strong sequence similarity to the same PFAM domain (Pfam number (PF08798) and name CRISPR-associated protein Cse3) and low E-values indicating further similarity. Also, Pfam provided the same HMM Logo for both *M.ruber's* and *E. coli's* casE which indicates the two proteins having the same conserved amino acids (Figure 8). HMM logo is graphical representation of all amino acids in a given protein depicting which is amino acid is highly conserve and which are not. The HMM logo is obtained from collection of aligned sequences (Thomas D. Schneider).

Table 2: Summary of the major evidence. According tothe information provided from variety of bioinformatics database, M. ruber's mrub_3014 is orthologous to E. coli's b2756.

Bioinformatics tools	*E. coli* b2756	*M. ruber* Mrub_3014
BLAST	E. coli against M. ruber Score: 60.5 bits, E-value 3e-17	
CDD	CDD or COG Number: cl08497 CRISPR/Cas system-associated RAMP superfamily protein Cas6e	
	E-Value: 1.48e-69	E-Value: 5.18e-44
Cellular Localization	Cytoplasm of the cell	
TIGRfam	TIGR01907	

	Cse3: CRISPR system CASCADE complex protein	
	E-value: 3.6e-51	E-value: 7.1e-09
PDB	Crystal structure of the CRISPR-associated protein Cas6e from Escherichia coli str. K-12 (Uncharacterized protein casE) Chain A	Structure of Thermus Thermophilus Cse3 bound to an RNA representing a product mimic complex (Putative uncharacterized protein TTHB192) Chain A
	E-value: 1.72397E-77	E. value: 3.91322E-7
Pfam	PF08798 CRISPR-associated protein Cse3	
	E-value: 4.4e-28	E-value: 4.5e-34

http://www.ncbi.nlm.nih.gov/blast, http://bioinformatics.biol.uoa.gr/PRED-TMBB/input.jsp, http://www.ncbi.nlm.nih.gov/blast, http://tigrblast.tigr.org/web-hmm/, http://pfam.sanger.ac.uk/search.



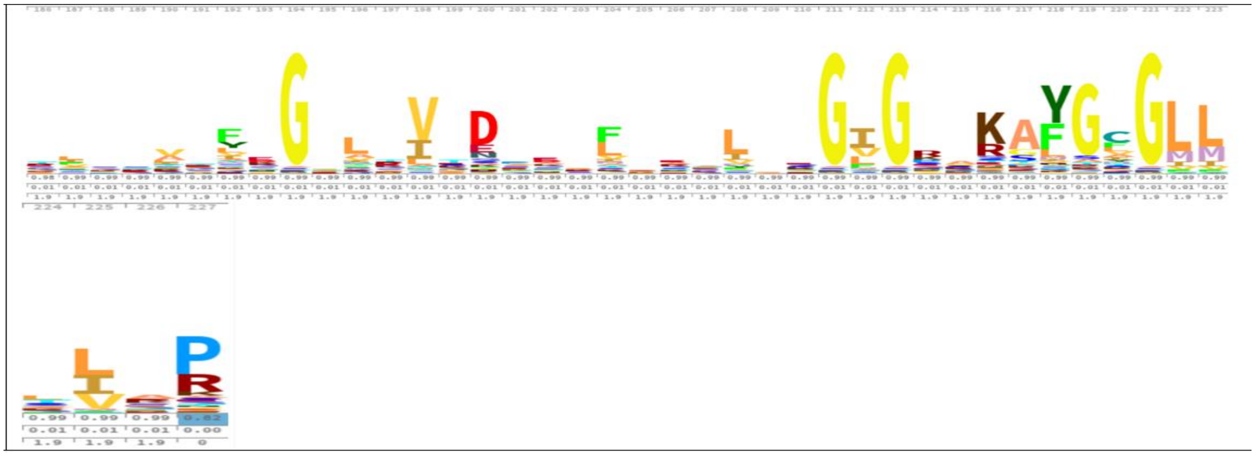
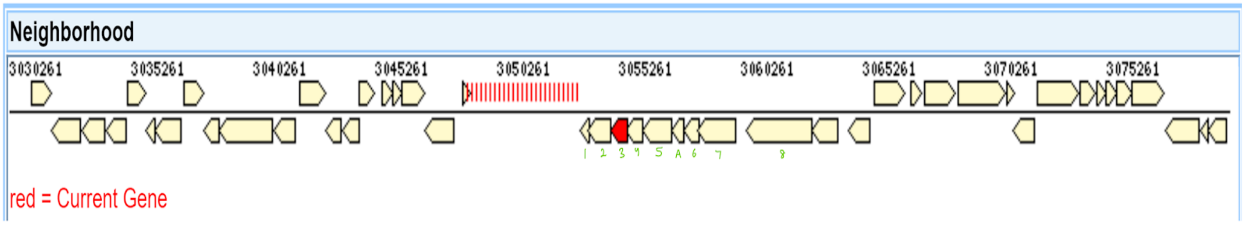


Figure 8: The following is HMM Logo provided by PFAM database.

http://pfam.sanger.ac.uk/search

Next, using data from IMG/M, we provided evidence that the mrub_3014 is part of an operon system and very similar to b2756. Figure 9 shows the chromosome map from IMG/M for the regions containing the Mrub-3014 and *E. coli* b2756 genes. The red highlighted gene *cse3* (Mrub_3014) on the top image is the gene of interest. Genes upstream of *cse3* are *cas1* and *cas2*. Genes downstream of *cse3* are *cas5*, *cse4*, *cse2*, *cse1*, and *cas3*. These genes that are flanking *cse3* are also found flanking *cse3* in *E. coli*. This indicates that mrub_3014 is part of an operon because all the neighboring genes are lined in the same direction. The bottom image, the red highlighted gene Cse3 (b2756) is the gene of interest. Genes upstream of *cse3* are *cas1* and *cas2*. Genes downstream of *cse3* are *cas5e*, *cse4, cse4*, *cse2*, *cse1* and *cas3*. These genes that are flanking Cse3 are also found flanking cse3 in *M. ruber* which indicates the similarities mrub_3014 and b2756 and provides evidence b2756 is part of an operon system.

M. ruber

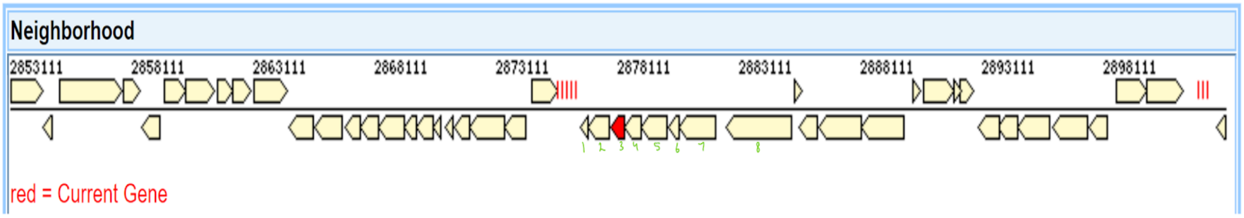


1. Cas2
2. Cas1
3. Cse3
4. Cas5
5. Cse4

A. Hyopthetical protein

6. Cse2
7. Cse1
8. Cas3

E. coli



1. Cas2
2. Cas1
3. Cse3
4. Cas5e
5. Cse4
6. Cse2
7. Cse1
8. Cas3

Figure 9: The chromosome map from IMG/M for Mrub_3014 and E. coli b2756 suggest that both genes are housed within a CRISPR-Cas operon. The figure displays that Mrub_3014 (top image red highlighted gene numbered 3) is part of an operon system and b2756

(bottom image highlighted gene numbered 3) is part of an operon system. The two images also display the similarities between the two genes. On the top image, the red highlighted gene Cse3 (Mrub_3014) is the gene of interest. Genes upstream of Cse3 are Cas1 and Cas2. Genes downstream of Cse3 are Cas5, Cse4, Cse2, Cse1, and Cas3. These genes that are flanking Cse3 are also found flanking Cse3 in *E. coli*. On the bottom image, the red highlighted gene Cse3 (b2756) is the gene of interest. Genes upstream of Cse3 are Cas1 and Cas2. Genes downstream of Cse3 are Cas5e, Cse4, Cse4, Cse2, Cse1 and Cas3. These genes that are flanking Cse3 are also found flanking Cse3 in *M. ruber.* *https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=GeneDetail&page=geneDetail&gene_oid=646674479*, *https://img.jgi.doe.gov/cgi-bin/m/main.cgi?section=GeneDetail&page=geneDetail&gene_oid=646314718*

Finally, the last figure provides more evidences why Mrub_3014 is part of an operon. Figure 10 shows the IMG/M Gene Neighborhood output, where genes are colored by their COG designation. The gene in red is mrub_3014 or its putative *casE* orthologs in different species. The CRISPR-Cas gene order is maintained in each of these species. Common gene order is an important indicator of gene function between species. Normally, gene order is highly variable between bacterial species; bacterial genomes show a high level of gene rearrangement, even between evolutionarily related species; genome instability can result from point mutations or from genome rearrangements such as deletions, duplications, amplifications, insertions, inversions, or translocations (Darmon *et al.,* 2014). Operon structure is an important indicator of gene function. “Operons are highly enriched in genes encoding related functions. Accordingly, operons often include genes encoding enzymes of consecutive steps in metabolic pathways (Zaslaver *et al.,* 2006) or genes encoding interacting proteins (Mushegian *et al.,* 1996; Huynen *et al.,* 2000). Conservation of a gene in an operon is thus a strong indication of functional neighborhood and can be used to make functional inferences (Overbeek *et al.,* 999; Moreno-Hagelsieb *et al.,* 2008). Operons are much more conserved than expected given rearrangement rates (de Daruvar *et al.,* 2002; Moreno-Hagelsieb *et al.,* 2008), presumably because of the advantages of such genetic organization (Lathe *et al.,* 2000; Omelchenko *et al.,* 2003; Price *et al.,* 2006)

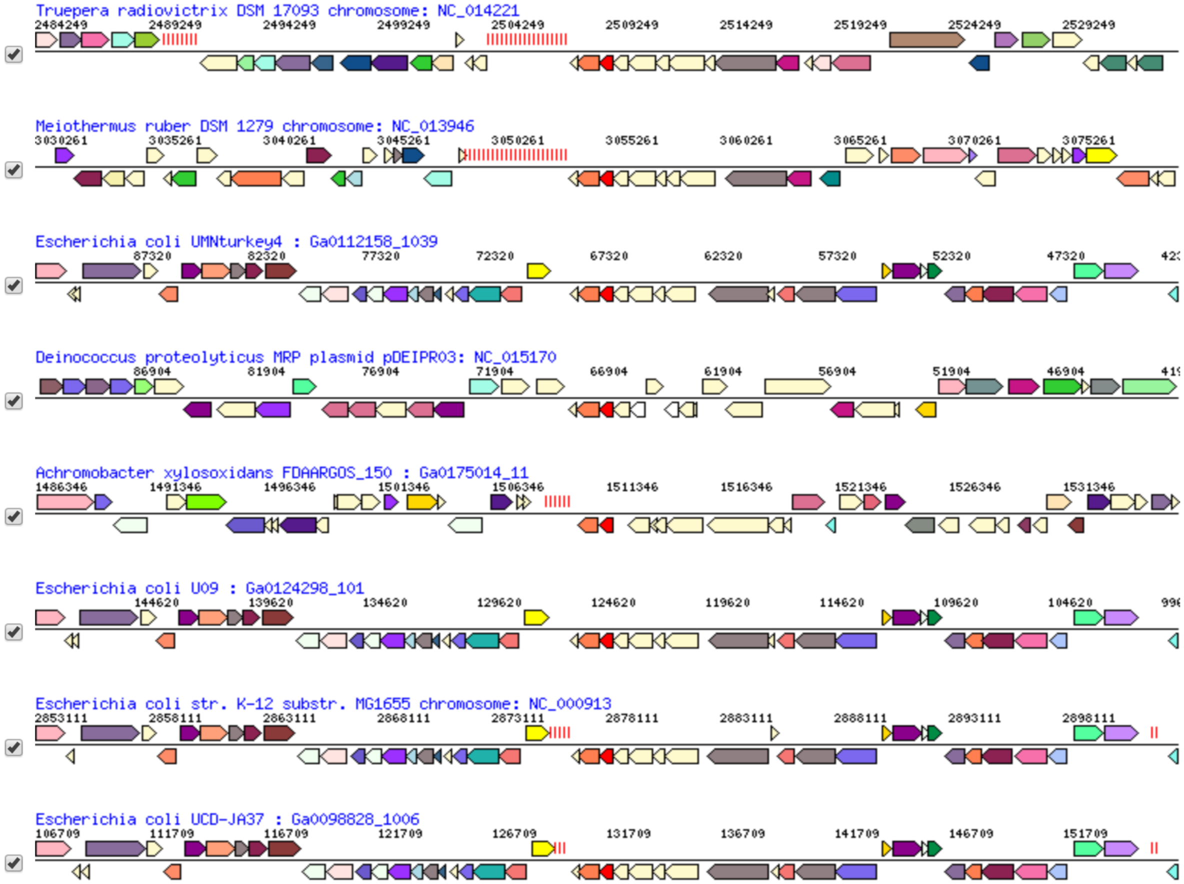


Figure 10. Gene Neighborhood output from IMG/M showing that Mrub_3014 must be part of an operon system. The picture displays that all gene with the same hits in other organisms that are closely related *to M. ruber* are in an operon. Therefore, it is reasonable for us to claim that Mrub_3014 is a part of an operon as well. https://img.jgi.doe.gov/cgi-bin/m/main.cgi

Conclusion:

Based on the data provided form variety of the databases, we conclude that *M. ruber* Mrub_3014 is orthologous to *E. coli* b2756. Throughout this research, Mrub_3014 was the gene of interest from *M. ruber* and b2756 was the gene of interest from the model organism, *E. coli.* Mrub_3014 is the locus tag for the *M. ruber* gene and b2756 is the locus tag for *E. coli* gene. Both of these genes are named *casE*, *cse3* and *cas6*e and their encoded protein is named casE. As stated in the beginning of this research, *casE* is a part of CRISPR-CAS system in both E. coli and M. ruber. We learned that casE in both organisms has a endonuclease activity that is very crucial and needed for sufficient cleavge of pre-CRISPR RNA. Furthermore, we learned casE is a part of an operon in both organisms. We provide evidences from databases such as BLAST, EcoCyc, KEGG and IMG/M to prove that Mrub_3014 and b3756 are very similar due to evolutionarily relatedness. Then we use databases such as TMHMM, PRED, and PSORT_B we confirmed that the protein of the two genes are located in the cytoplasm of the cell. Using the CDD database, TIGRFAM, and PFAM we further provided evidence that Mrub_3014 (*casE*) and b2756 (*casE*) are orthologous.

References:

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

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

Barrangou R, Fremaux C, Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. Science 315(5819):1709-12.

Darmon E and Leach DRF. 2014. Bacterial genome instability. Microbiol Mol Biol Rev 78(1):1-39.

de Daruvar A, Collado-Vides J, Valencia A. 2002. Analysis of the cellular functions of Escherichia coli operons and their conservation in bacillus subtilis. J Mol Evol 55(2):211-21.

Euzéby JP. 1997. List of bacterial names with standing in nomenclature: A folder available on the internet. Int J Syst Bacteriol 47(2):590-2.

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.

Huynen M, Snel B, Lathe W, Bork P. 2000. Predicting protein function by genomic context: Quantitative evaluation and qualitative inferences. Genome Res 10(8):1204-10.

Jiang F and Doudna JA. 2015. The structural biology of CRISPR-cas systems. Curr Opin Struct Biol 30:100-11.

Jiang Y, Chen B, Duan C, Sun B, Yang J, Yang S. 2015. Multigene editing in the escherichia coli genome via the CRISPR-Cas9 system. Appl Environ Microbiol 81(7):2506-14.

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

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/

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

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/

Liu S, Yuan Z, Yuan YA. 2015. Structural insights into specific crRNA G-rich sequence binding by meiothermus ruber Cse2. J Struct Biol 190(2):122-34.

L. Medina-Aparicio, J. E. Rebollar-Flores, A. L. Gallego-Hernández, A. Vázquez, L. Olvera, R. M. Gutiérrez-Ríos, E. Calva, I. Hernández-Lucas. 2011a. The CRISPR/cas immune system is an operon regulated by LeuO, H-NS, and leucine-responsive regulatory protein in salmonella enterica serovar typhi. Journal of Bacteriology 193(10):2396-407.

Loginova LG and Egorova LA. 1975. [Thermus ruber obligate thermophilic bacteria in the thermal springs of kamchatka]. Mikrobiologiia 44(4):661-5.

Lathe WC, Snel B, Bork P. 2000. Gene context conservation of a higher order than operons. Trends Biochem Sci 25(10):474-9.

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/

Makarova KS, Haft DH, Barrangou R, Brouns SJJ, Charpentier E, Horvath P, Moineau S, Mojica FJM, Wolf YI, Yakunin AF, et al. 2011. Evolution and classification of the CRISPR-cas systems. Nat Rev Microbiol 9(6):467-77.

Makarova KS, Wolf YI, Alkhnbashi OS, Costa F, Shah SA, Saunders SJ, Barrangou R, Brouns SJJ, Charpentier E, Haft DH, et al. 2015. An updated evolutionary classification of CRISPR-cas systems. Nat Rev Microbiol 13(11):722-36.

Makarova KS, Wolf YI, Koonin EV. 2013. The basic building blocks and evolution of CRISPR-CAS systems. Biochem Soc Trans 41(6):1392-400.

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

Moreno-Hagelsieb G and Janga SC. 2008. Operons and the effect of genome redundancy in deciphering functional relationships using phylogenetic profiles. Proteins 70(2):344-52.

Mushegian AR and Koonin EV. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci U S A 93(19):10268-73.

Omelchenko T, Vasiliev JM, Gelfand IM, Feder HH, Bonder EM. 2003. Rho-dependent formation of epithelial "leader" cells during wound healing. Proc Natl Acad Sci U S A 100(19):10788-93.

Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. 1999. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A 96(6):2896-901.

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

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904-9.

Shmakov S, Abudayyeh OO, Makarova KS, Wolf YI, Gootenberg JS, Semenova E, Minakhin L, Joung J, Konermann S, Severinov K, et al. 2015. Discovery and functional characterization of diverse class 2 CRISPR-cas systems. Mol Cell 60(3):385-97.

Tindall BJ, Sikorski J, Lucas S, Goltsman E, Copeland A, Glavina Del Rio T, Nolan M, Tice H, Cheng J, Han C, et al. 2010. Complete genome sequence of meiothermus ruber type strain (21).StandGenomic Sci 3(1):26-36.

Wright AV, Nuñez JK, Doudna JA. 2016. Biology and applications of CRISPR systems: Harnessing nature's toolbox for genome engineering. Cell 164(1-2):29-44.

Wu Y, Suhasini AN, Brosh RM. 2009. Welcome the family of FANCJ-like helicases to the block of genome stability maintenance proteins. Cell Mol Life Sci 66(7):1209-22.

Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, et al. 2010. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 26(13):1608-15.

Zaslaver A, Bren A, Ronen M, Itzkovitz S, Kikoin I, Shavit S, Liebermeister W, Surette MG, Alon U. 2006. A comprehensive library of fluorescent transcriptional reporters for Escherichia coli. Nat Methods 3(8):623-8.

Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, Volz SE, Joung J, van der Oost J, Regev A, et al. 2015. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-cas system. Cell 163(3):759-71.