2018

Mrub_1675, Mrub_1676, Mrub_1677, and Mrub_1679 genes are orthologs of b_3458, b_3457, b_3456, and b_3454 genes in *E. coli*, respectively, coding for ABC transporters. Mrub_1678 and b_3455, though perform similar tasks, are not orthologous

Ravi Patel
Augustana College, Rock Island Illinois

Alaina Hofmann
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

TMrub_1675, Mrub_1676, Mrub_1677, and Mrub_1679 genes are orthologs of *b_3458, b_3457, b_3456, and b_3454* genes in *E. coli,* respectively, coding for ABC transporters. *Mrub_1678 and b_3455*, though perform similar tasks, are not orthologous.

Ravi Patel
Alaina Hofmann
Dr. Lori R. Scott Laboratory
Biology Department, Augustana College
639 38th Street, Rock Island, IL 61201

Introduction

Meiothermus Ruber: What is it? And why study it?

Meiothermus Ruber (M.ruber), originally named *Thermus ruber,* was named from Greek origins, ‘meion’, meaning ‘lesser’ and ‘thermos’, meaning ‘hot’. The species name, ‘ruber’, is named after the pigment color of M.ruber, a bright red (Tindall *et al., 2010*). However,, the species is heterogenous (Lapage et al., 1952), in respect to the pigment color , and the other species in the genus display different pigmentation colors (Tenreiro et al., 1995), ranging from a pale yellow to a deep orange (Tindall *et al., 2010*).

M.ruber, originally named *Thermus ruber (Skerman)(*Lapage et al., 1952*),* is a gram negative, aerobic, rod shaped bacteria.The species was first found in and around Russian hot springs by Laginova (Loginova et al., 1987), where they thrive in their optimal temperature of about 60°C (Loginova et al., 1987). Interestingly enough,

however, the heat loving bacteria is not very well researched (Albuquerque et al ., 2009). There are only 28 publications for *M.ruber* (Scott et al., 2015) where as there are over 30,000 for *E.coli* and *Salmonella.* However, the importance of this study comes from the Joint Genome Institute and the Genomic Encyclopedia of Bacteria and Archaea (GEBA) project . The project aims to study lesser understood organisms (Phylogenetic Diversity et al ., 2015) because they may reveal processes or variants of processes that are not present in other well known organisms. Thus, the GEBA project aims to look into organisms that are often overlooked.

E. Coli **as a control to study** ***M.Ruber***

This project used a well known model organism, *E.coli* as a control and a point of comparison/contraction to better understand *M.ruber*. Model organisms, such as *E.coli*, have shown to grow very well in lab and replicate very rapidly and are not too difficult to maintain, which makes them good model organisms. These factors are some of the reasons why *E.coli* has been studied so extensively. However, the reason *E.coli* was chosen as the control for this project was that a BLAST search revealed that *E.coli's Pro C* protein sequence was very similar to that of *M.ruber*. It was later proved that *M.ruber_1345* gene was an ortholog of the *b_0386* gene of *E.coli,* which coded for Pyrroline-5-carboxylate reductase. To further investigate the orthology between *E.coli* and *M.ruber* we will look into a variety of branched chain amino acid ABC transporters.

ABC Transporter Proteins

Gases, polar molecules and small nonpolar molecules can cross the plasma membrane relatively easily. However, most substrates do not fall into these three categories and require a protein to cross the plasma membrane. ABC transporters are just some of the proteins cells use to transport molecules that fall outside of the three listed categories. Considering that most substrates are not gases, polar molecules or small nonpolar molecules, it is easy to make sense of the fact that a large portion of most genomes are dedicated to creating proteins that are transporters. ABC transporters, in specific, are a special type of transmembrane protein comprised of two nucleotide-binding domains (NBD), which initially bind to a substrate outside the cell; and two transmembrane domains (TMD), which are segments that cross the plasma membranes (Wilkens, et al ., 2015). ATP is hydrolyzed on on the NBD which causes a conformational change in the TMDs and allows for the solute of interest to enter the cell (Kaiyani et al., 2016).

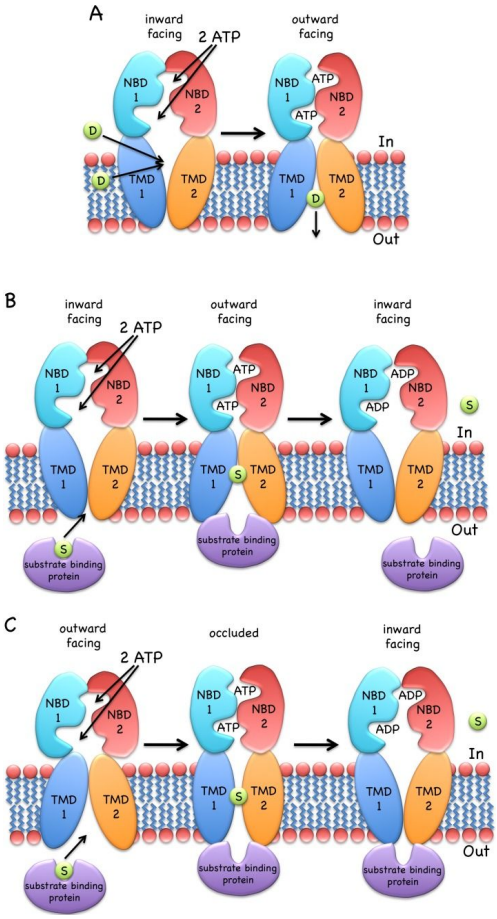


Figure 1 shows how to solute will bind to the receptor of NBD, which will allow for ATP to bind. ATP then hydrolyzed, causes the conformational change in the TMDs and the solute enters. Once ADP+Pi (the product of ATP hydrolysis) is released, the protein 'resets' (Kaiyani et al., 2016).

Bioinformatics:

The bioinformatics programs used in this project are very important to use in many applications among many biological practices. From the simple protein BLAST (Madden *et al.,* 2002), to the more indepth analyzation of genes, bioinformatics can give information in a wide variety of specific genes, pathways, and gene sequences. Access to these programs gives the science community immensely knowledgeable databases that is easily accessible to in research and projects such as this to save time. As science continues to make advances, bioinformatics tools will advance along with discoveries and be even more accessible to scientists and researchers.

Purpose/Hypothesis:

The purpose of this project is to determine if *Mrub_1675, Mrub_1676, Mrub_1677, Mrub_1678, and Mrub_1679* genes are orthologs of *b_3458, b_3457, b_3456, b_3455, and b_3454* genes in *E. coli,* respectively. We determined this by using multiple bioinformatics tools to show us the similarities and differences between the genes to tell if they are indeed orthologs of eachother. An important value to understand is the E-values. The E-values that come from the programs help to determine the significance of of the results. If the E-values generated are high values, the sequences are more likely to be aligned versus low E-values which indicate that the sequence is significantly different and therefore could be orthologs. To begin our original hypothesis, the *M. ruber* genes were BLASTed (Madden *et al.,* 2002) against *E. coli* and the *E. coli* genes were BLASTed (Madden *et al.,* 2002) against *M. ruber* and low E-values were obtained.

Since low E-values were generated, we can hypothesize that the genes are orthologs of eachother, but further bioinformatic analysis will further confirm the hypothesis.

Methods

In this gene annotation project, the GENI-ACT gene annotation instructions (Scott *et al.,* 2016) were followed as well as the addition of 15 BLAST (Madden *et al.,* 2002) hits in T-coffee (Notredame *et al.,* 2000), EcoCyc (Keseler *et al.,* 2013), and colored by KEGG (Kanehisa *et al.,* 2016) were supplementally used to generate bioinformatics data. To start the research, the *M. ruber* genes were BLASTed against *E. coli* and the *E. coli* genes were BLASTed against *M. ruber*. Once the genes were BLASTed, the GENI-ACT site instructions (Scott *et al.,* 2016) were followed and the bioinformatics tools were used to generate data to help confirm the orthologs. Using the T-coffee program (Notredame *et al.,* 2000), we used 15 BLAST (Madden *et al.,* 2002) hits for each gene. Instead of using MetaCyc for the ABC transporter pathway information, we used the EcoCyc (Keseler *et al.,* 2013) program. To be able to visualize the genes upstream and downstream of the genes in question, we used colored by KEGG (Kanehisa *et al.,* 2016) to colorfully see the genes involved in ABC transportation. The KEGG (Kanehisa *et al.,* 2016) genome pathway program was used to visualize the pathway of the genes and to help identify what mechanism of action the pathway was involved in. KEGG (Kanehisa *et al.,* 2016) was also used to identify locus tags, DNA coordinates, nucleotide sequences, and amino acid sequences of the genes in question. Next, EcoCyc (Keseler *et al.,* 2013), was used to visualize the pathway of the genes in a form

of their functionality in ABC transportation. EcoCyc (Keseler *et al.,* 2013) was also used to visualize the operon map for *E. coli* pathway. NCBI BLAST (Madden *et al.,* 2002) was used to show pairwise alignment between the genes and *M. ruber* and *E. coli*, the resulting bit score and E-values were used to show the orthologs of the genes. Integrated Microbial Genomes and Microbiomes (IMG/M) (Markowitz *et al.,* 2012) program was used for proposed DNA coordinates as well as identifying Shine-Delgarno sequences to help aid in identifying if the original start codon was called correctly. IMG/M (Markowitz *et al.,* 2012) was also used towards the end of the project to display a chromosome map of our genes in question and were able to analyze upstream and downstream genes. Another bioinformatics tool that was used was T-coffee (Notredame *et al.,* 2000) and it was used to find sequence alignments in different bacterias and give the possibility to align all of the different sequences. Being able to align multiple sequences, it allows the visualization of seeing if the start codon was called correctly. Another bioinformatics tool that was helpful in calling the original start codon was Weblogo (Crooks *et al*., 2016). Weblogo (Crooks *et al*., 2016) was a great visual tool to show the highly conserved amino acids that were in the sequences that helped to further confirm the start codon was called correctly. TMHMM (Krogh *et al.,* 2001), SignalP (Kall *et al.,* 2004) , LipoP (Juncker *et al.,* 2016) , PSORT-B (Yu *et al.,* 2010), and Phobius (Kall *et al.,* 2007) were all used for the visualization and confirmation of transmembrane proteins and to locate where the proteins reside. NCBI (Madden *et al.,* 2002) bioinformatics program was also used to get CDD results and COG information to further confirm the genes roles with accompanying bit scores and E-values. TIGRfam

(Haft *et al.,* 2001) and Pfam (Finn *et al.,* 2014, Finn *et al.,* 2016) programs were used to confirm the genes roles with accompanying bit scores and E-values as well with the use of the sequence family classifications, pairwise alignments, and HMM logos. Another important bioinformatics tool was PDB (Berman *et al.,* 2000) and it showed protein structures that were involved with the genes in our pathways. It also showed pairwise alignments, bit score, E-value, and other research that had been done.

Results

Table 1:

Bioinformatics tool used	*E. coli b_3454 (liv F)*	*Mrub_1679 (liv F)*
BLAST *E.coli* against *M. ruber*	Score: 241 bits E-value: 2e-81	
CDD Data (COG Category)	COG0410 ABC-type branched-chain amino acid transport system	COG0410 ABC-type branched-chain amino acid transport system
Cellular Localization	Cytoplasm	Cytoplasm
TIGRfam- protein family	TIGR03410 urea ABC transporter E-value: 4.7e-60	TIGR03410 urea ABC transporter, urea binding protein 2.24e-76
Pfam- protein family	PF00005: ATP-binding domain of ABC transporters E-Value: 3.3e-33	PF00005: ATP-binding domain of ABC transporters E-value: 2.42e-46

Protein Database	Crystal structure of an ABC-type branched-chain amino acid transporter (RPA4397) from Rhodopseudomonas palustris CGA009 at 1.50 A resolution E-value 0.0	ABC transporter 6.38e-36
KEGG pathway map	Branched Chain amino acid transport	Branched Chain amino acid transport

Table 1: The table above compares *E. coli b_3454 (liv F)* and *Mrub_1679 (liv F)* with the bioinformatics tools used. The bioinformatics tools used include NCBI BLAST (Madden *et al.,* 2002), IMG/M (Markowitz *et al.,* 2012), TIGRfam (Haft *et al.,* 2001), Pfam (Finn *et al.,* 2014, Finn *et al.,* 2016), PDB (Berman *et al.,* 2000), and KEGG (Kanehisa *et al.,* 2016).

Table 2:

Bioinformatics tool used	*E. coli b_3455 (liv G)*	*Mrub_1678 (liv G)*
BLAST *E.coli* against *M. ruber*	Score: 248 bits E-value: 7e-84	
CDD Data (COG Category)	COG0411 ABC-type branched-chain amino acid transport system	COG4177 ABC-type branched-chain amino acid transport system, permease component
Cellular Localization	Cytoplasm	cytoplasm
TIGRfam- protein family	TIGR03411 urea ABC transporter	TIGR03408

	3.7e-65	
Pfam- protein family	PF00005: ATP-binding domain of ABC transporters E-Value: 3.1e-32	PF00001: urea ABC transporter, permease protein UrtC 2.51e-18
Protein Database	Crystal structure of an ABC-type branched-chain amino acid transporter (RPA4397) from Rhodopseudomonas palustris CGA009 at 1.50 A resolution E-value 0.0	Branched-chain amino acid transport system (CL0142) 4.56e-13
KEGG pathway map	ABC Transporter	ABC transporter

Table 2: The table above compares *E. coli b_3455 (liv G)* and *Mrub_1678 (liv G)* with the bioinformatics tools used. The bioinformatics tools used include NCBI BLAST (Madden *et al.,* 2002), IMG/M (Markowitz *et al.,* 2012), TIGRfam (Haft *et al.,* 2001), Pfam (Finn *et al.,* 2014, Finn *et al.,* 2016), PDB (Berman *et al.,* 2000), and KEGG (Kanehisa *et al.,* 2016).

Table 3:

Bioinformatics tool used	*E. coli b_3456 (liv M)*	*Mrub_1677 (liv M)*
BLAST *E.coli* against *M. ruber*	Score: 184 bits E-value: 5e-54	
CDD Data (COG Category)	COG4177 ABC-type branched-chain amino acid transport system	COG4177 ABC-type branched-chain amino acid transport system, permease component
Cellular Localization	Transmembrane	Transmembrane
TIGRfam- protein family	TIGR03410 urea ABC transporter 1e-11	TIGR03408 urea ABC transporter, permease protein UrtC
Pfam- protein family	PF02653: Branched-chain amino acid transport system / permease component E-Value: 9e-60	Pf02653 Branched-chain amino acid transport system / permease component 4.56e-13
Protein Database	Crystal structure of an ABC-type branched-chain amino acid transporter (RPA4397) from Rhodopseudomonas palustris CGA009 at 1.50 A resolution E-value 0.0	No PDB results
KEGG pathway map	ABC Transporter	ABC transporter

Table 3: The table above compares *E. coli b_3456 (liv M)* and *Mrub_1677 (liv M)* with the bioinformatics tools used. The bioinformatics tools used include NCBI BLAST

(Madden *et al.,* 2002), IMG/M (Markowitz *et al.,* 2012), TIGRfam (Haft *et al.,* 2001), Pfam (Finn *et al.,* 2014, Finn *et al.,* 2016), PDB (Berman *et al.,* 2000), and KEGG (Kanehisa *et al.,* 2016).

Table 4:

Bioinformatics tool used	*E. coli b_3457 (liv H)*	*Mrub_1676 (liv H)*
BLAST *E.coli* against *M. ruber*	Score: 201 bits E-value: 7e-64	
CDD Data (COG Category)	COG0559 ABC-type branched-chain amino acid transport system	COG0559 Branched-chain amino acid ABC-type transport system,
Cellular Localization	Transmembrane	Transmembrane
TIGRfam- protein family	TIGR03410 urea ABC transporter 3.7e-8	TIGR03409 Urea ABC transporter
Pfam- protein family	PF02653:Branched-chain amino acid transport system / permease component E-Value: 6.7e-71	pfam02653 Branched-chain amino acid transport system / permease component 6.20e-18
Protein Database	Crystal structure of an ABC-type branched-chain amino acid transporter (RPA4397) from Rhodopseudomonas palustris CGA009 at 1.50 A resolution E-value 0.0	Solution Structure of PHAX-RBD in complex with ssRNA (2XC7)

KEGG pathway map	ABC Transporter	ABC transporter

Table 4: The table above compares *E. coli b_3457 (liv H)* and *Mrub_1676 (liv H)* with the bioinformatics tools used. The bioinformatics tools used include NCBI BLAST (Madden *et al.,* 2002), IMG/M (Markowitz *et al.,* 2012), TIGRfam (Haft *et al.,* 2001), Pfam (Finn *et al.,* 2014, Finn *et al.,* 2016), PDB (Berman *et al.,* 2000), and KEGG (Kanehisa *et al.,* 2016).

Table 5:

Bioinformatics tool used	*E. coli b_3458 (liv K)*	*Mrub_1675 (liv K)*
BLAST *E.coli* against *M. ruber*	Score: 173 bits E-value: 8e-52	
CDD Data (COG Category)	COG0683 ABC-type branched-chain amino acid transport system	COG0683 ABC-type branched-chain amino acid transport system
Cellular Localization	Cytoplasm	Cytoplasm
TIGRfam- protein family	TIGR03407 urea ABC transporter -166.5	Tigr03407 urea ABC transporter, urea binding protein
Pfam- protein family	PF13458: Periplasmic binding protein E-Value: 1.7e-61	PF13458: Periplasmic binding protein E-value: 2.42e-46

Protein Database	Crystal structure of an ABC-type branched-chain amino acid transporter (RPA4397) from Rhodopseudomonas palustris CGA009 at 1.50 A resolution E-value 0.0	Extracellular ligand binding receptor from Desulfohalobium retbaense DSM5692 E-value: 8.88e-16
KEGG pathway map	ABC Transporter	Branched Chain amino acid transport

Table 5: The table above compares *E. coli b_3458 (liv K)* and *Mrub_1675 (liv K)* with the bioinformatics tools used. The bioinformatics tools used include NCBI BLAST (Madden *et al.,* 2002), IMG/M (Markowitz *et al.,* 2012), TIGRfam (Haft *et al.,* 2001), Pfam (Finn *et al.,* 2014, Finn *et al.,* 2016), PDB (Berman *et al.,* 2000), and KEGG (Kanehisa *et al.,* 2016).

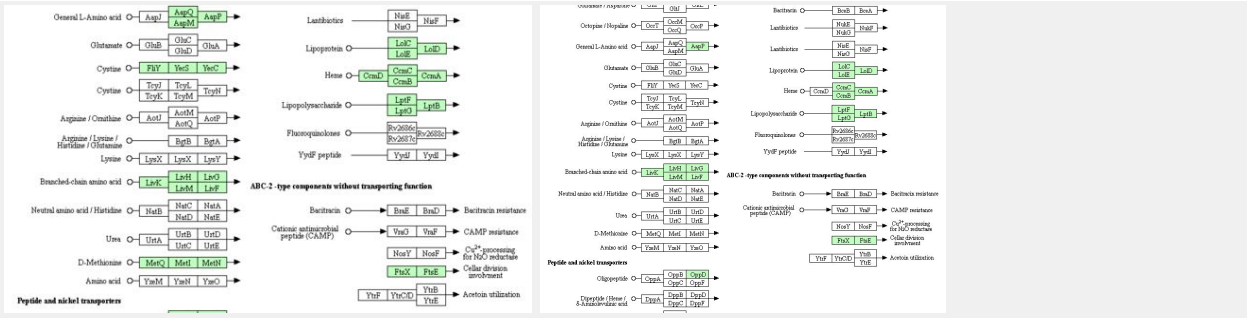
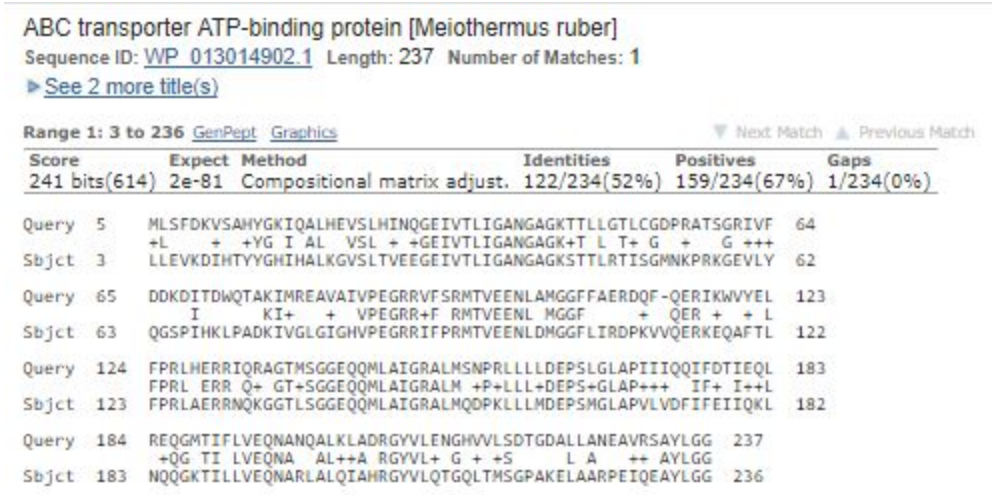


Figure 2: The pathway maps of *E.coli* (left) and *M.ruber (right)* are shown. Both have the branched chain amino acids.`

E.coli b_3454 / Mrub_1679

Table 1 summarizes the results generated from a variety of bioinformatics tools. The BLAST data shows that the e-value is 2e-81, which is significantly below the .001 cutoff and suggests that the sequence similarity is likely not by chance. The CDD data, also generated the same COG number, TIGRfam and Pfam all generated the same protein number as well, with very low e-values, which further shows that the two genes are not similar by chance. TMH, SignalP, LipoP, and PSORT-B all suggested that the proteins are found in the in the cytoplasm of the cell. They both also lack a cleavage site. The two proteins are, by the evidence provided, likely orthologous ABC transporters.

Figure 3 shows the BLAST results, when *E.coli* BLASTed against *M.ruber.*


ABC transporter ATP-binding protein [Meiothermus ruber]
Sequence ID: WP_013014902.1 Length: 237 Number of Matches: 1
See 2 more title(s)

Range 1: 3 to 236 GenPept Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps
241 bits(614)	2e-81	Compositional matrix adjust.	122/234(52%)	159/234(67%)	1/234(0%)

```
Query  5    MLSFDKVSAHYGKIQALHEVSLHINQGEIVTLIGANGAGKTTLLGTLCGDPRATSGRIVF  64
            +L    +  +YG I AL  VSL + +GEIVTLIGANGAGK+T L T+ G  +   G +++
Sbjct  3    LLEVKDIHTYYGHIHALKGVSLTVEEGEIVTLIGANGAGKSTTLRTISGMNKPRKGEVLY  62

Query  65   DDKDITDWQTAKIMREAVAIVPEGRRVFSRMTVEENLAMGGFFAERDQF-QERIKWVYEL  123
                I      KI+   +  VPEGRR+F RMTVEENL MGGF     +  QER +  + L
Sbjct  63   QGSPIHKLPADKIVGLGIGHVPEGRRIFPRMTVEENLDMGGFLIRDPKVVQERKEQAFTL  122

Query  124  FPRLHERRIQRAGTMSGGEQQMLAIGRALMSNPRLLLLDEPSLGLAPIIIQQIFDTIEQL  183
             FPRL ERR Q+ GT+SGGEQQMLAIGRALM +P+LLL+DEPS+GLAP+++  IF+ I++L
Sbjct  123  FPRLAERRNQKGGTLSGGEQQMLAIGRALMQDPKLLLMDEPSMGLAPVLVDFIFEIIQKL  182

Query  184  REQGMTIFLVEQNANQALKLADRGYVLENGHVVLSDTGDALLANEAVRSAYLGG  237
              +QG TI LVEQNA  AL++A RGYVL+ G + +S     L A   ++ AYLGG
Sbjct  183  NQQGKTILLVEQNARLALQIAHRGYVLQTGQLTMSGPAKELAARPEIQEAYLGG  236
```

Figure 3: BLAST results from *E.coli b_3454* BLASTed against *Mrub_1679* (Madden *et al.,* 2002)

E. coli b_3455* / *Mrub_1678

Table 2 summarizes the results generated from a variety of bioinformatics tools. The BLAST data generated shows an e-value of 7e-84, which suggests that the two proteins

are not similar by chance. However, the proteins generated 2 different COG numbers, 2 different TIGR numbers and two different Pfam numbers. Though the two proteins have shown to be found in the same location (TMH, SignalP, LipoP, and PSORT-B) and serve similar functions, they do not seem to be orthologous. This is not a shock, because the proteins are from different phyla, and more than likely found different ways to carry out similar functions. Figure 4 shows the BLAST results.

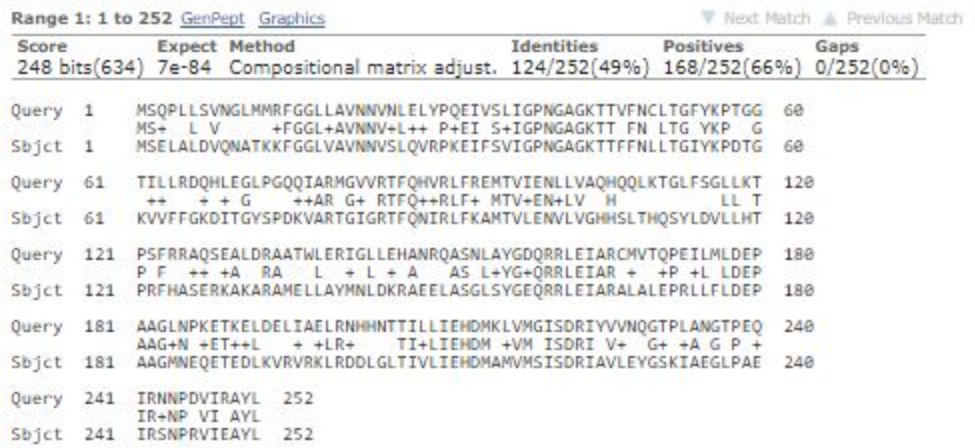


Figure 4: BLAST results from *E.coli b_3455* BLASTed against *Mrub_1678* (Madden *et al.,* 2002)

E. coli b_3456* / *Mrub_1677

Table 3 summarizes the results generated from a variety of bioinformatics tools. The BLAST data shows that the e-value is 5-e54, which is significantly below the .001 cutoff and suggests that the sequence similarity is likely not by chance. The CDD data, also generated the same COG number, TIGRfam and Pfam all generated the same protein number as well, with very low e-values, which further shows that the two genes are not similar by chance. TMH, SignalP, LipoP, and PSORT-B all suggested that the proteins are found embedded in the plasma membrane of the cell. They both also lack a

cleavage site. The two proteins are, by the evidence provided, likely orthologous ABC transporters. Figure 5 shows the BLAST results.

Range 1: 86 to 473 GenPept Graphics — Next Match — Previous Match

Score	Expect	Method	Identities	Positives	Gaps
184 bits(466)	5e-54	Compositional matrix adjust.	146/397(37%)	209/397(52%)	58/397(14%)

```
Query  77   FILPAIDGSTVKQKLFLVALLVLAVAWPFMVSRGTVDIATLTMIYIILGLGLNVVVGLSG  136
            F+LP +  ST+ +     A+L +AV    + +    ++     I+  L LGLNVVVG +G
Sbjct  86   FLLPNL--STLVRVALGAAILFIAVPIAGLTNSFLFELGIQIGIFAALALGLNVVVGQAG  143

Query  137  LLVLGYGGFYAIGAYTFALLNHYYGLGFWTCLPIAGL--------------MAAAAGFLL  182
            LL LG+  F+AIGAYT+ +          F    P  GL                AA  G L+
Sbjct  144  LLDLGFAAFFAIGAYTWGIFGSPQAAQFIPGYPSEGLPGNYLYLFMALAVITAAITGVLI  203

Query  183  GFPVLRLRGDYLAIVTLGFGEIVRILLLNNTE----ITGGPNGISQIPKPTLFGL-EFSR  237
            G P LRLRGDYLAIVTLG GE+VR+   NN +    IT GP GI+ + +P +  L EF R
Sbjct  204  GLPALRLRGDYLAIVTLGLGEVVRVFA-NNLDKPLNITNGPQGITPVNRPEVGPLTEFLR  262

Query  238  TAREGGWDTFSNFFGLKYD-PSDRVIFLYLVALLLVVLSLFVINRLLRMPLGRAWEALRE  296
                        +G   D P     F YL+ L+++ + + V  RL     GRAW A+RE
Sbjct  263  AIGA------ERLYGRPIDEPIAYAFFFYLLVLVVIGIVVLVNIRLANSRFGRAWVAIRE  316

Query  297  DEIACRSLGLSPRRIKLTAFTISAAFAGFAGTLFAARQGFVSPESFTFAESAFVLAIVVL  356
            DEIA +++G+     KL AF   AAF+G  G +FAA+Q FVSPESFT   S  +LA V+L
Sbjct  317  DEIAAKAMGIPLLPTKLLAFATGAAFSGAMGAIFAAKQTFVSPESFTLQASINILAFVIL  376

Query  357  GGMGS-QFAVILAAILLVVSRELMRDFNE----------------------------YSM  387
            GGMGS   AV+ AA + V++  +++DF++                             Y
Sbjct  377  GGMGSIGGAVVGAAAVTVLNIGILKDFSDLLNTWRQTGVTILGYNMANLPPQLNPAKYER  436

Query  388  LMLGGLMVLMMIWRPQGLLPMTRPQLKLKNGAAKGEQ  424
            L+ G +++LMMI+RP+GL+P  R + +L+    + ++
Sbjct  437  LVFGLILILMMIFRPEGLIPEQRHRAELQEARQEAKE  473
```

Figure 5: BLAST results from *E.coli b_3456* BLASTed against *Mrub_1677* (Madden *et al.,* 2002)

E. coli b_3457 / *Mrub_1676*

Table 4 summarizes the results generated from a variety of bioinformatics tools. The BLAST data shows that the e-value is 7e-64, which is significantly below the .001 cutoff and suggests that the sequence similarity is likely not by chance. The CDD data, also generated the same COG number, TIGRfam and Pfam all generated the same protein number as well, with very low e-values, which further shows that the two genes are not similar by chance. TMH, SignalP, LipoP, and PSORT-B all suggested that the proteins are found embedded in the plasma membrane of the cell. They both also lack a cleavage site. The two proteins are, by the evidence provided, likely orthologous ABC transporters. Figure 6 shows the BLAST results.

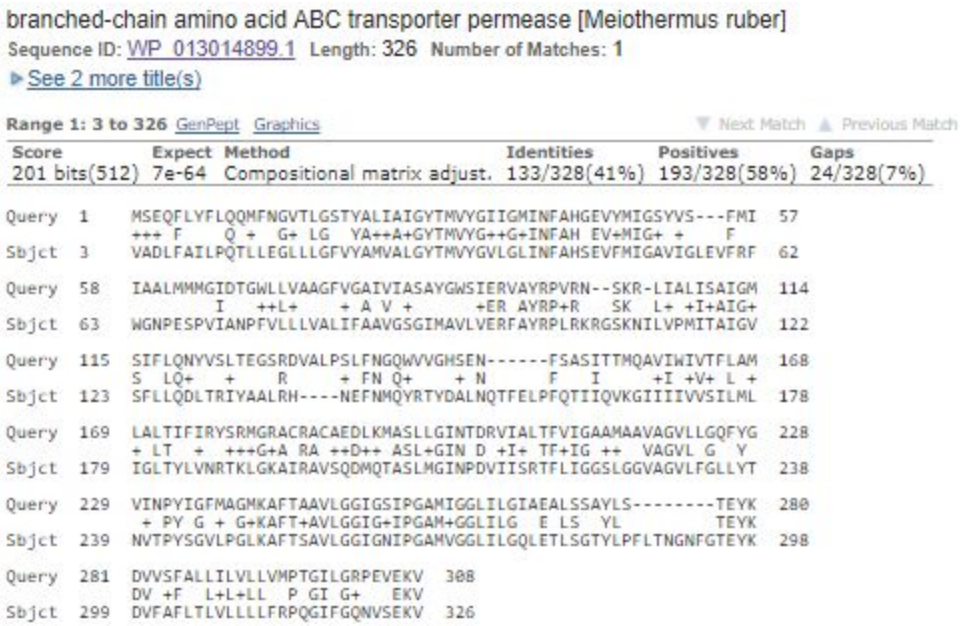
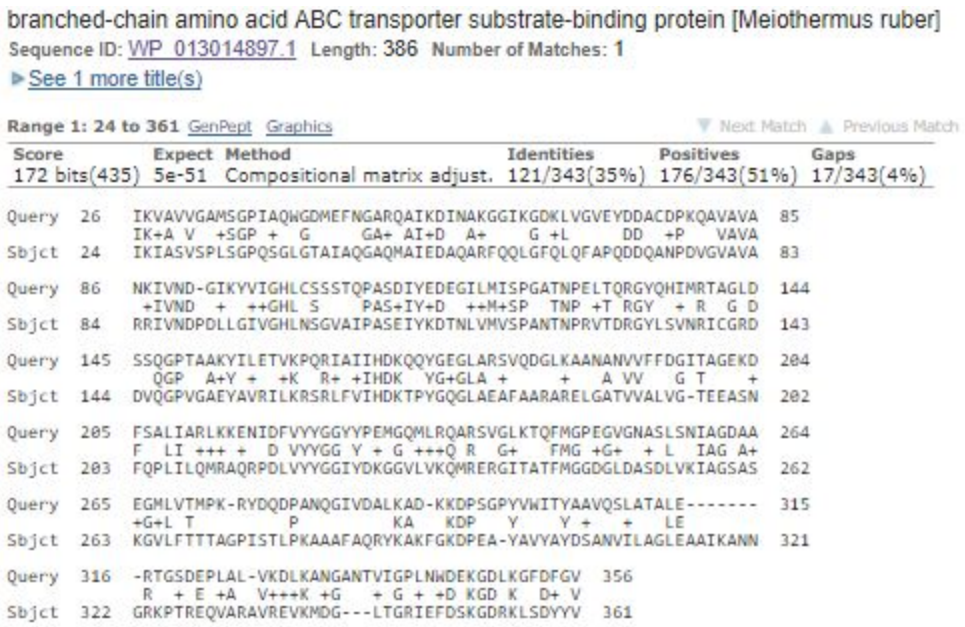


Figure 6: BLAST results from *E.coli b_3457* BLASTed against *Mrub_1676* (Madden *et al.,* 2002)

E. coli b_3458 / *Mrub_1675*

Table 5 summarizes the results generated from a variety of bioinformatics tools. The BLAST data shows that the e-value is 8e-52, which is significantly below the .001 cutoff and suggests that the sequence similarity is likely not by chance. The CDD data, also generated the same COG number, TIGRfam and Pfam all generated the same protein number as well, with very low e-values, which further shows that the two genes are not similar by chance. TMH, SignalP, LipoP, and PSORT-B all suggested that the proteins are found in the cytoplasm, on the interior of the cell. They both also lack a cleavage site. The two proteins are, by the evidence provided, likely orthologous ABC transporters. Figure 7 shows the BLAST results.



branched-chain amino acid ABC transporter substrate-binding protein [Meiothermus ruber]
Sequence ID: WP_013014897.1 Length: 386 Number of Matches: 1
▶ See 1 more title(s)

Range 1: 24 to 361 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
172 bits(435)	5e-51	Compositional matrix adjust.	121/343(35%)	176/343(51%)	17/343(4%)

```
Query  26   IKVAVVGAMSGPIAQWGDMEFNGARQAIKDINAKGGIKGDKLVGVEYDDACDPKQAVAVA  85
            IK+A V  +SGP +  G     GA+ AI+D  A+    G +L     DD  +P   VAVA
Sbjct  24   IKIASVSPLSGPQSGLGTAIAQGAQMAIEDAQARFQQLGFQLQFAPQDDQANPDVGVAVA  83

Query  86   NKIVND-GIKYVIGHLCSSSTQPASDIYEDEGILMISPGATNPELTQRGYQHIMRTAGLD  144
             +IVND  +  ++GHL S    PAS+IY+D  ++M+SP  TNP +T RGY  + R  G D
Sbjct  84   RRIVNDPDLLGIVGHLNSGVAIPASEIYKDTNLVMVSPANTNPRVTDRGYLSVNRICGRD  143

Query  145  SSQGPTAAKYILETVKPQRIAIIHDKQQYGEGLARSVQDGLKAANANVVFFDGITAGEKD  204
              QGP  A+Y +  +K  R+ +IHDK  YG+GLA +     +   A VV   G T    +
Sbjct  144  DVQGPVGAEYAVRILKRSRLFVIHDKTPYGQGLAEAFAARARELGATVVALVG-TEEASN  202

Query  205  FSALIARLKKENIDFVYYGGYYPEMGQMLRQARSVGLKTQFMGPEGVGNASLSNIAGDAA  264
            F  LI +++ +  D VYYGG Y + G +++Q R  G+   FMG +G+  + L  IAG A+
Sbjct  203  FQPLILQMRAQRPDLVYYGGIYDKGGVLVKQMRERGITATFMGGDGLDASDLVKIAGSAS  262

Query  265  EGMLVTMPK-RYDQDPANQGIVDALKAD-KKDPSGPYVWITYAAVQSLATALE-------  315
            +G+L T         P          KA   KDP   Y    Y +   +   LE
Sbjct  263  KGVLFTTTAGPISTLPKAAAFAQRYKAKFGKDPEA-YAVYAYDSANVILAGLEAAIKANN  321

Query  316  -RTGSDEPLAL-VKDLKANGANTVIGPLNWDEKGDLKGFDFGV  356
             R  + E +A  V+++K +G   + G + +D KGD K  D+ V
Sbjct  322  GRKPTREQVARAVREVKMDG---LTGRIEFDSKGDRKLSDYYV  361
```

Figure 6: BLAST results from *E.coli b_3458* BLASTed against *Mrub_1675* (Madden *et al.,* 2002)

Why Urea is relevant when discussing ABC transporters.

Urea, is clearly important when discussing ABC transporters, as seen in the results tables. Urea is formed as a result of deamination of amino acids. The same processes that facilitate the transport of amino acids, which are then broken down to components of urea (11).

Conclusion

M.ruber_1675* and *b_3458

M.ruber_1675 and *b_3458* both had the same COG grouping of COG0683. This grouping is the ABC-type branched-chain amino acid transport system, a periplasmic component. The E-value generated for *M.ruber_1675* was 2.42e-46 and for *E. Coli b_3458* it was 2.40e-102. The very low e-value, which is well below the .001 cut off, signals that the data is not generated by chance, and indicates significance. The proteins also had the same Pfam number, which means that they are in the same family in the Pfam database. The both share the PF13458 family, which is the periplasmic binding protein. *M.ruber*'s Pfam E-value was 2.42e-46 and *E.coli* had an E-value of 1.7e-61, the low E-value indicates that these proteins were not placed in the family by random chance. They also had the same TIGRfam hit as well, Tigr03407. *M.ruber* generated a Tigrfam E-value of 1.6e-05, while the *E.coli* generated an E-value of .00042. The cellular Localization signals indicated that both proteins are found in the Cytoplasm, and neither have any transmembrane helices. The next step would be to create a primer that replaces a highly conserved amino acid (Glutamate) into an Alanine and to observe the function. Figure 7 shows the conversion.

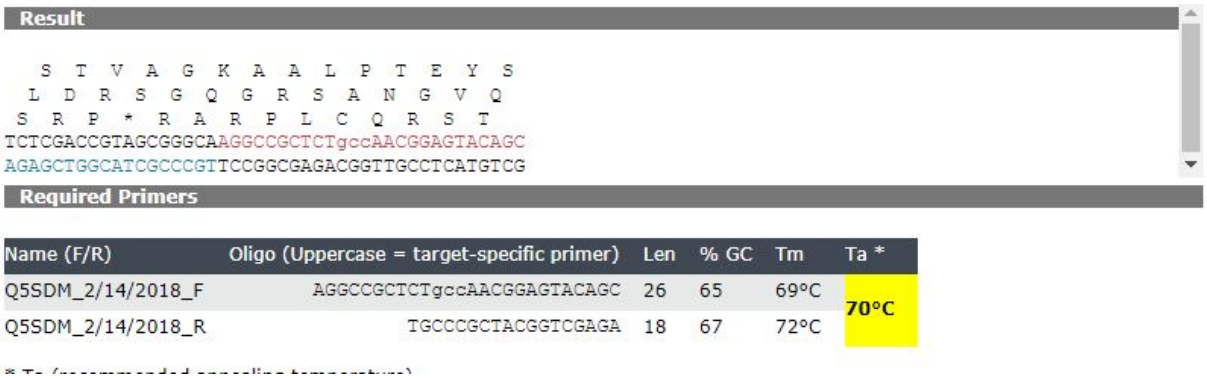


Figure 7: Substitution of a highly conserved Glutamate at position 15 into an Alanine.

M.ruber_1676* and *b_3457

M.ruber_1676 and *b_3457* both had the same COG groupings as well. Both proteins generated a COG number of 0559, which means they are in the same COG family. This family is the Branched-chain amino acid ABC –type transport system, a permease component. *M.ruber* generated an E-value of 2.39e-48 and *E.coli b_3457* generated an E-value of 4.54e-82. The very low E-values, well below the .001 cut off, indicate that these proteins are not in this family by chance, and quantify their significance. In addition, Pfam also generated results that showed both the proteins being in the same family, PF02653, which is the branched-chain amino acid transport system, a permease component. The *M.ruber* had an E-value of 6.20e-18, and the *E.coli* had an E-value of 6.7e-71, both these values are vastly below the minimum cut off, and is indicative of the fact that they were not placed into this family b random chance. TIGRfam also placed the proteins into the same family, they were both in TIGR03410, which is the urea ABC transporter. The *M.ruber* had an E-value of 6.20e-18 and the *E.coli* had an E-value of 2.7e-15, which are both well below the cutoff, and indicative of their significance. The cellular localization signal indicated that both proteins are found embedded in the membrane, and this is further proven by the fact that each one has multiple transmembrane helices.The next step would be to create a primer that replaces a highly conserved amino acid (Glutamate) into an Alanine and to observe the function. Figure 8

shows the conversion.

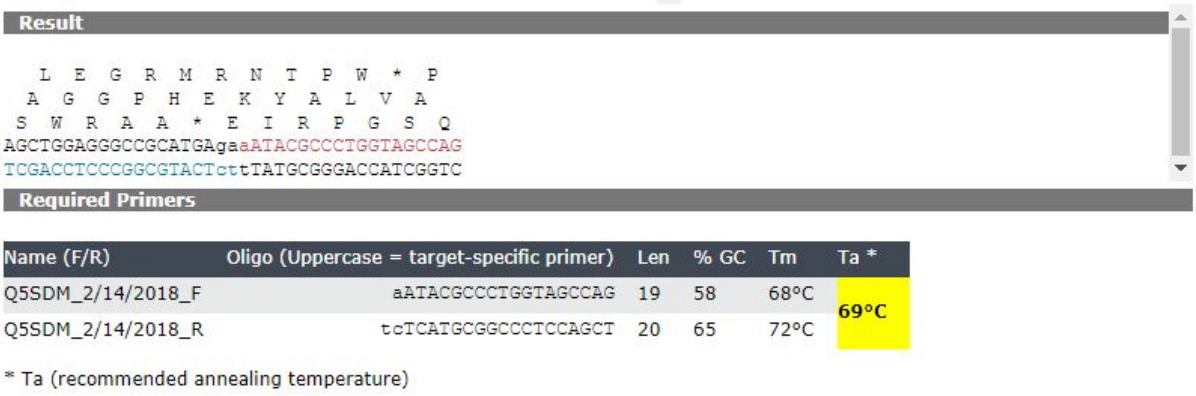


Figure 8: The substitution of Glutamate at position 2, to an Alanine.

M.ruber_1677 and *b_3456*

M.ruber_1677 and *b_3456* both had the same COG groupings. Both proteins generated a cog number of 4177. This value belongs to the ABC-type branched-chain amino acid transport system, a permease component. *M.ruber_1677* had an E-value 7.46e-34 while *E.coli b_3456* had an E-value of 6.05e-78. The low E-values indicate that these proteins were not placed into the family by chance, and that they are significant. Pfam also placed these two proteins in the same family; they were both placed in the PF02653, which is the Amino acid transport system, which is a permease component. *M.ruber* generated an E-value of 4.56e-13 and *E.coli* generated an E-value of 9.0e-60, both values are significantly below the cut off, and are indicative of the fact that they are not placed into the families by random chance. TIGRfam also placed both proteins in the same family. Both were found in the TIGR03410, which is the urea ABC transporter family. Both also had E-values well below the cut off. Both proteins were also found

embedded in the membrane, with multiple transmembrane helices for each one. The next step would be to create a primer that replaces a highly conserved amino acid (Glutamate) into an Alanine and to observe the function. Figure 9 shows the conversion.

Result

```
   G  L  A  G  G  V  Y  A  M  P  R  P  Y  S  S  P  L  T  M  P  L
  R  A  G  R  W  G  L  R  H  A  K  T  L  F  L  S  P  Y  D  A  F
 P  G  W  P  V  G  F  T  P  C  Q  D  P  I  P  L  P  L  R  C  L  W
CCGGGCTGGCCGGTGGGGTTTACGCCATgccAAGACCCTATTCCTCTCCCCTTACGATGCCTTTGG
GGCCCGACCGGCCACCCCAAATGCGGTACGGTTCTGGGATAAGGAGAGGGGAATGCTACGGAAACC
```

Required Primers

Name (F/R)	Oligo (Uppercase = target-specific primer)	Len	% GC	Tm	Ta *
Q5SDM_2/14/2018_F	TTTACGCCATgccAAGACCCTATTCCTCTCCCCTTACGATGCCTTTGG	48	52	77°C	72°C
Q5SDM_2/14/2018_R	CCCCACCGGCCAGCCCGG	18	89	81°C	

* Ta (recommended annealing temperature)

Figure 9: Shows the substituion of a highly conserved Glutamate at position 7 into an Alanine.

M.ruber_1679* and *b_3454

M.ruber_1679 and *b_3454* both had the same COG grouping; they both generated a cog number of 0410. This value belongs to the ABC-type branched-chain amino acid transport system, a ATPase component. *M.ruber_1679* had an E-value of 2.46e-114 and *E.coli b_3454* had an E-value of 5.11e-136. The low E-values indicate that these proteins were not placed into the family by random chance, and signify their significance. Both proteins also had the same Pfam family, they were both found in the PF00005, which is the ATP-binding domain of ABC transporters. *M.ruber* had an E-value of 2.42e-46 and *E.coli* had an E-value of 3.3e-33, which are both well below the cut off and are indicative that these proteins are not in the PF00005 family by random chance. TIGRfam also showed that both the proteins were in the TIGR03410 family,

which is the Urea ABC transporter, both had very low e-values which is indicative of the fact that the proteins were not placed into the families by random chance. The proteins are also found in the cytoplasm, with no transmembrane helices. The next step would be to create a primer that replaces a highly conserved amino acid (Glutamate) into an Alanine and to observe the function. Figure 10 shows the conversion.

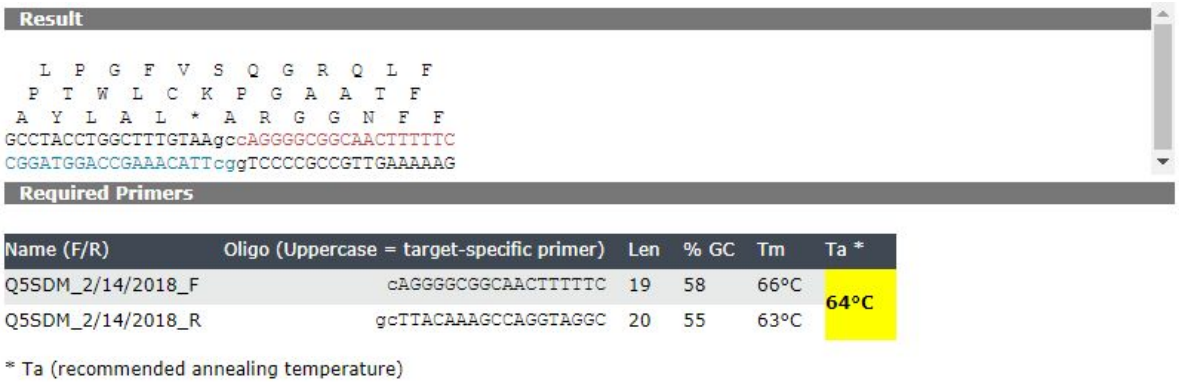
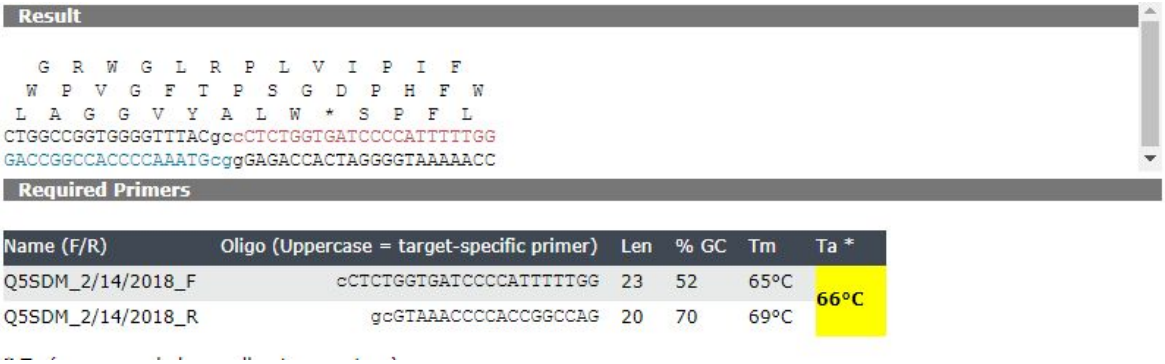


Figure 10: Shows the substitution of a highly conserved Glutamate at position 45, into an alanine.

M.ruber_1678* and *b_3455

M.ruber_1678 and *b_3455* had different COG groupings. *M.ruber_1678* had a COG number of 4177 with an E-value of 7.46e-34, while *b_3455* had a number of 4177 with an e-value of 1.61e-139. The different COG values combined with the low E-values, is indicative that these proteins are not orthologous. In addition, the Pfam results gave different results for the proteins, *M.ruber* was found in the PF00001 family, which is the urea ABC transporter a permease protein. *E.coli* was found in the PF00005, which was the ATP-binding domain of ABC transporters. Both proteins had very low E-values,

which is indicative of the fact that they were not placed in their respective families by random chance. Additionally, they both had different TIGRfam hits, *M.ruber* was placed into the TIGR03408 family while *E.coli* was placed into the TIGR03411 family, which both had very low E-values. Both the proteins were found in the cytoplasm with no transmembrane helices, however. Based on the difference in the bioinformatic tools, it is clear that *M.ruber_1678* and *b_3455* are not orthologous. This is not a shock, since they are from different phyla and are likely to have some major differences. The next step would be to create a primer that replaces a highly conserved amino acid (Glutamate) into an Alanine and to observe the function. Figure 11 shows the conversion.



Result

```
  G  R  W  G  L  R  P  L  V  I  P  I  F
 W  P  V  G  F  T  P  S  G  D  P  H  F  W
L  A  G  G  V  Y  A  L  W  *  S  P  F  L
CTGGCCGGTGGGGTTTACgccCTCTGGTGATCCCCATTTTTGG
GACCGGCCACCCCAAATGcggGAGACCACTAGGGGTAAAAACC
```

Required Primers

Name (F/R)	Oligo (Uppercase = target-specific primer)	Len	% GC	Tm	Ta *
Q5SDM_2/14/2018_F	cCTCTGGTGATCCCCATTTTTGG	23	52	65°C	66°C
Q5SDM_2/14/2018_R	gcGTAAACCCCACCGGCCAG	20	70	69°C	

Figure 11: Shows the substitution of a highly conserved Glutamate at position 15 into an alanine

Literature Cited

1. Tindall BJ, Sikorski J, Lucas S, Goltsman E, Copeland A, Glavina Del Rio T, Nolan M, Tice H, Cheng J, Han C, et al. Complete genome sequence of meiothermus ruber type strain (21). Standards in Genomic Sciences 2010;3(1):26-36.
2. Loginova LG, Egorova LA, Golovacheva RS, Seregina LM. Thermus ruber sp. nov., nom. rev. Int J Syst Bacteriol 1984; 34:498-499. doi:10.1099/00207713-34-4-498
3. Scott LR. Meiothermus ruber Genome Analysis Project. [Internet]. GENI-ACT; [2015 Dec 16]. Available from: http://geni-science.org/secure/projects/view/
4. Phylogenetic Diversity. [Internet] U.S. Department of Energy Joint Genome Institute; [2015 Dec 16]. Available from http://jgi.doe.gov/our-science/science-programs/microbialgenomics/phylogenetic-diversity/
5. 1. Wilkens S. Structure and mechanism of ABC transporters. F1000prime Reports 2015;7:14.
6. Skerman VBD, McGowan V, Sneath PHA. Approved Lists of Bacterial Names. Int J Syst Bacte
7. Lapage SP, Sneath PHA, Lessel EF, Skerman VBD, Seeliger HPR, Clark WA. 1992. International Code of Nomenclature of Bacteria (1990 Revision). Bacteriological Code. American Society for Microbiology, Washington, DC.

8. Albuquerque L, Ferreira C, Tomaz D, Tiago I, Veríssimo A, da Costa MS, Nobre MF. Meiothermus rufus sp. nov., a new slightly thermophilic red-pigmented species and emended description of the genus Meiothermus. Syst Appl Microbiol 2009; 32:306-313. PubMed doi:10.1016/j.syapm.2009.05.002
9. Tenreiro S, Nobre MF, da Costa MS. Thermus silvanus sp. nov. and Thermus chliarophilus sp. nov., two new species related to Thermus ruber but with lower growth temperatures. Int J Syst Bacteriol 1995; 45:633-639. PubMed doi:10.1099/00207713-45-4-633
10. 1. Kalyani DC, Munk L, Mikkelsen JD, Meyer AS. Molecular and biochemical characterization of a new thermostable bacterial laccase from meiothermus ruber DSM 1279. RSC Adv 2016;6(5):3910-8.
11. Meessen J. Urea synthesis. Chemie Ingenieur Technik 2014 Dec;86(12):2180-9.
12. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. PubMed
13. 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/.
14. Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank *Nucleic Acids Research*, 28: 235-242.
15. Betts, M. J., and Russell, R. B. 2003. Amino-Acid Properties and Consequences of Substitutions. Bioinformatics for Geneticists. 2003: 311–342.
16. Biolabs, N. E. Home - NEB | New England Biolabs. Home - NEB | New England Biolabs. https://www.neb.com/
17. Crooks, GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator, Genome Research, 14:1188-1190, (2004)

18. Finn RD, Bateman A, Clements J, et al. Pfam: the protein families database. Nucleic Acids Research. 2014;42 (Database issue):D222-D230. doi:10.1093/nar/gkt1223.

19. 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/

20. 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.

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

22. Kall L, Krough A, Sonnhammer E. 2004. A combined transmembrane topology and signal Lukas Käll, Anders Krogh and Erik L. L. Sonnhammer. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology*, 338(5):1027-1036, May 2004. (doi) (PubMed)

23. Käll,L., Anders Krogh and Erik L. L. Sonnhammer. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res.*, 35:W429-32, July 2007 (doi) (PubMed)

24. Kanehisa M, Sato Y, Kawashima M, Furumichi M. and Tanabe M. (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res., 44, D457–D462; [2016 Dec 6]. Available from: http://www.genome.jp/kegg/

25. 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.

26. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: Application to

complete genomes. Journal of Molecular Biology, 305(3):567-580, January 2001. (PDF, 959503 bytes)

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

28. 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/ BLAST tool: BLASTp tool from https://blast.ncbi.nlm.nih.gov/Blast.cgi

29. 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. 2015 Jan 28;43(Database issue):D222-2. doi: 10.1093/nar/gku1221. Epub 2014 Nov 20. [PubMed PMID: 25414356]

30. 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

31. Sonnhammer, ELL., G. von Heijne, and A. Krogh.A hidden Markov model for predicting transmembrane helices in protein sequences. In J. Glasgow, T. Littlejohn, F. Major, R. Lathrop, D. Sankoff, and C. Sensen, editors, Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology, pages 175-182, Menlo Park, CA, 1998. AAAI Press.

32. Tindall, B. J., Sikorski, J., Lucas, S., Goltsman, E., Copeland, A., Glavina Del Rio, T., … Lapidus, A. (2010). Complete genome sequence of Meiothermus ruber type strain (21T). Standards in Genomic Sciences, 3(1), 26–36. http://doi.org/10.4056/sigs.1032748

33. Yu, NY, 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):1608-161