2017

Mrub_1304, Mrub_2007 and Mrub_2006 are orthologs of *E. coli* b_3189, b_3972, and b_0091, respectively within the Peptidoglycan Biosynthesis Pathway

Mylaun E. Griffith
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Mrub_1304, Mrub_2007 and Mrub_2006 are orthologs of *E.coli* b_3189, b_3972, and b_0091, respectively within the Peptidoglycan Biosynthesis Pathway

Mylaun E. Griffith
Dr. Lori Scott Laboratory
Biology Department, Augustana College
639 38th St., Rock Island, IL, 61201
February 2017

Introduction

Reasons for using Meiothermus ruber

Meiothermus ruber (*M. ruber*) is a form of gram-negative, thermophilic, red-pigmented bacteria that thrives in aerobic and preferably hot environments ranging from 35-70 ºC (Tindall *et al.*, 2010). This organism was isolated initially in Russia from natural hot springs (Loginova *et al.*, 1975). There are 3,105 genes predicted to be in the initial automated annotation, 3,052 are putative protein-coding genes, and 53 are RNA genes (*e.g.*, tRNA, rRNA, *etc*.); thirty-eight pseudogenes were also identified in *M. ruber* (Lori Scott 2017). Yet, in general, there has not been many studies performed with the various genes mentioned above in *M. ruber*. Therefore, it is important to study them, in order to gain more knowledge of their functions and their contribution to science. However, *Escherichia coli (E.coli)* has been studied quite frequently and is usually used as a model organism because of the extensive research that has been done on its genes (Lori Scott 2017). Because there are so few studies done on *M. ruber* as opposed to other species within the Deinococcus-Thermus phylum, it is important to investigate the genes within this organism to further our knowledge within the field of Genomics. As DOE JGI states "Genome sequencing has revolutionized our understanding of microorganisms and the role they play in important processes, including pathogenesis, energy production, bioremediation, global nutrient cycles; and the origins, evolution, and diversity of life," so by researching *M. ruber* we contribute to the understanding of this organism in life (Lori Scott 2017)." In this project, we will further studies of *M. ruber* by studying three genes in this organism that are involved in peptidoglycan biosynthesis, using the genes in *E. coli* as a model.

The well-studied organism, *E. coli*, will be used as the control in my research. *E. coli* will serve as a model organism for the information that we lack about *M. ruber*. *E. coli* is a highly versatile, well-studied organism that is easily grown in the lab (Blount, 2015). Therefore, we have reason to use *E.coli* as our control within this research because not only is it well-studied, but also has genes that could be orthologous to the genes in *M.ruber*. The pathway being studied is peptidoglycan biosynthesis pathway which is used by many bacterium (Gautam *et al.*, 2011).

Peptidoglycan Biosynthesis

The cell wall is a key component of most eubacteria (Gautam *et al.*, 2011). Not only does the cell wall serve as a structural shape, but also provides strength and protection as well for these organisms (Gautam *et al.*, 2011). Disruptions to the cell wall could ultimately result in cell death, which is why the components that that cell wall contains are essential (Gautam *et al.*, 2011). Many bacteria can negatively impact humans causing illness. Therefore, understanding the structural and mechanistic properties of enzymes involved in cell wall biosynthesis, could

enable us to create inhibitors that could kill or prevent cell replication of the bacteria. Peptidoglycan is only found within eubacteria and has been prove to be a main component in bacterial cell walls (Gautam *et al.*, 2011). Peptidoglycan biosynthesis is a complex multi-step process that occurs initially in the cytoplasm and then moves to the plasma membrane (Gautam *et al*., 2011). In Figure 2, the enzymes expressed by the genes of interest within this research, *MurA, MurB*, and *MurC*, are involved in Stage I of peptidoglycan biosynthesis located in the cytoplasm. Because peptidoglycan plays an important role within the cell wall, it is important to study some the genes within its pathway of different organisms, especially *Meiothermus ruber.* We especially are examining *Meiothermus ruber* bacteria because it is understudied and has ornithine in its genus's peptidoglycan. This could mean that *Meiothermus ruber* may have a special cell wall based on the stressful environments it is exposed to. These stressful environments may cause *M. ruber* to have a different peptidoglycan biosynthesis for better chances of survival.

Figure 1 shows the reaction in the peptidoglycan biosynthesis pathway from substrate to product. In both E.coli and M.ruber, the enzyme UDP-N-acetylglucosamine 1-carboxyvinyltransferase (*MurA*) catalyzes the conversion of UDP-N-acetylglucosamine to UDP-N-acetylglucosamine enolpyruvate.The enzyme UDP-N-acetylenolpyruvoylglucosamine reductase (*MurB*) catalyzes the conversion of UDP-N-acetylglucosamine enolpyruvate to UDP-N-acetylmuramate. The UDP-N-acetylmuramate--L-alanine ligase catalyzes the conversion of UDP-N-acetylmuramate to UDP-N-MurNAc-L-Ala.

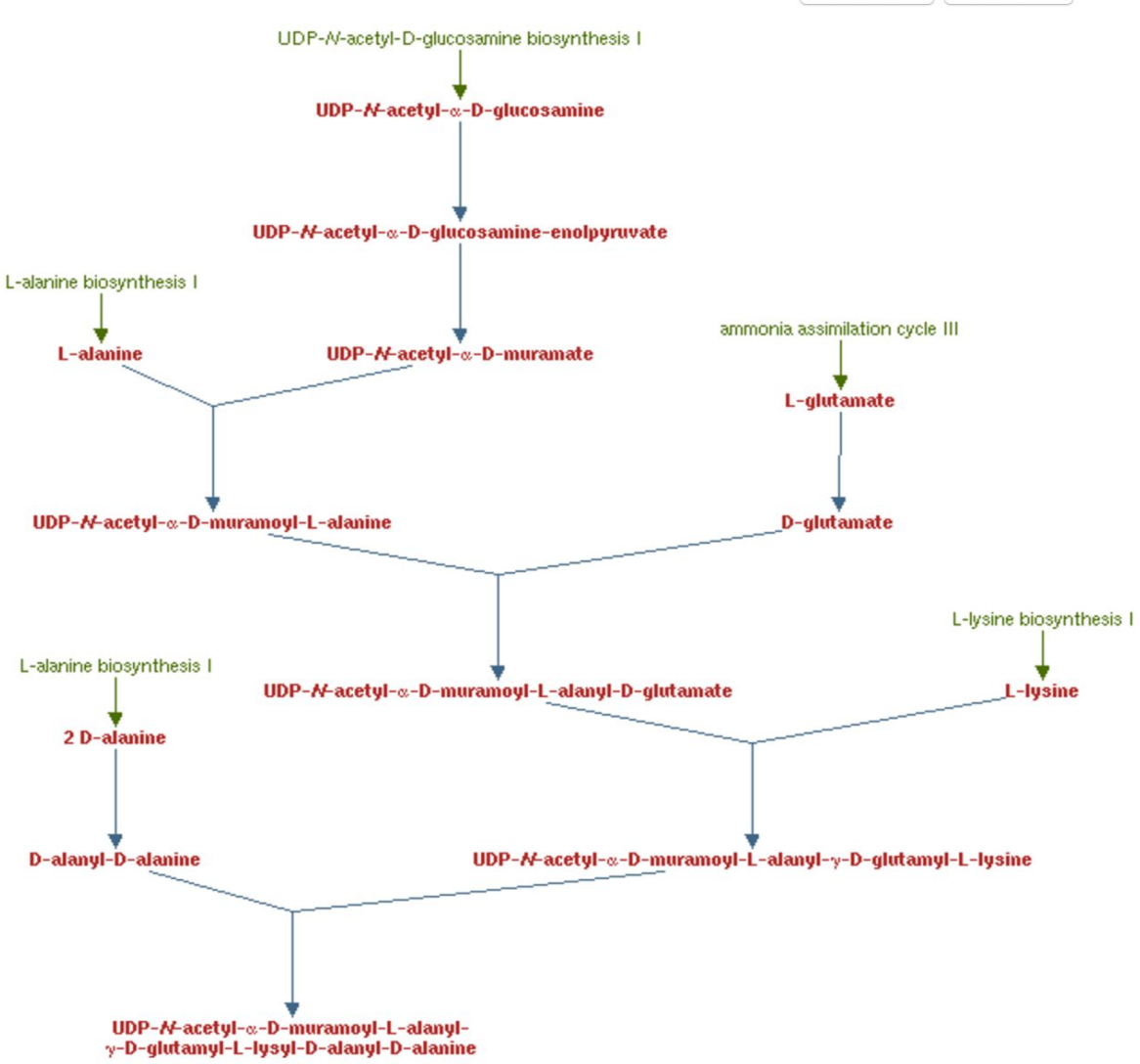


Figure 1. Peptidoglycan biosynthesis pathway displaying the enzymes involved with this reaction. Image taken from: https://metacyc.org

Figure 2 depicts the three stages that occur in peptidoglycan biosynthesis and their location within the cell. The proteins that will be discussed in this project are *MurA, MurB*, and *MurC*, which are all in stage I of peptidoglycan biosynthesis located in the cytoplasm. Stage II and III of peptidoglycan biosynthesis are located in the plasma membrane. However, the proteins associated with those stages will not be discussed in this research project.

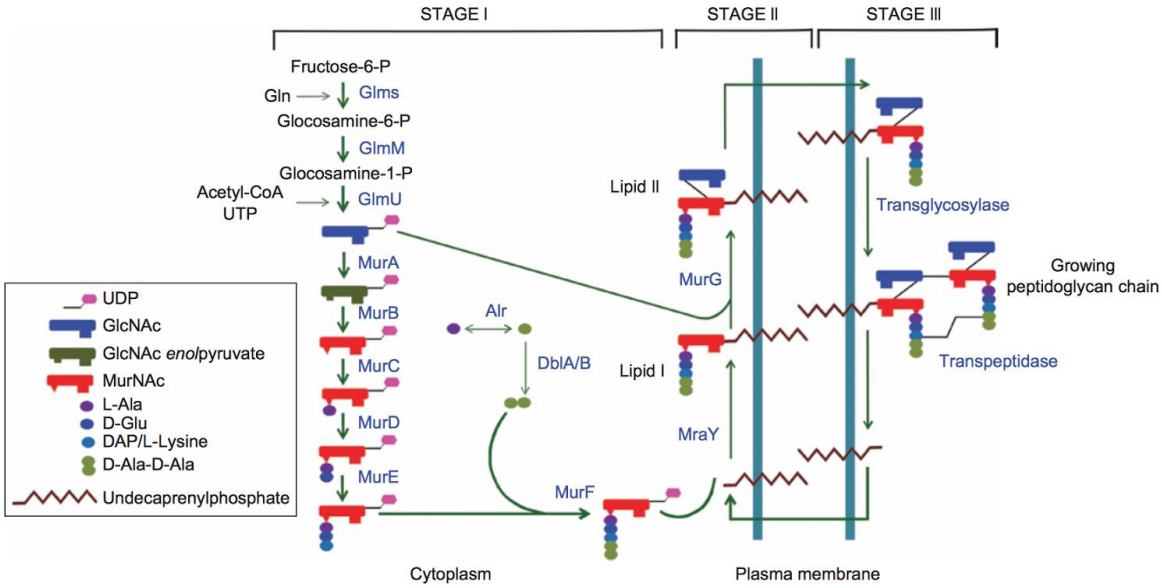


Figure 2. Representation of peptidoglycan biosynthesis and indication of each stages location. Image taken from: (Gautam *et al.* 2011) http://web.a.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=3&sid=f0a100be-6581-448e-9fc8-203d7e38a6d5%40sessionmgr4006&hid=4204

Bioinformatics

Bioinformatics are utilized in the field of biology to analyze data outputs. There are many bioinformatic programs available online that are free that could assist in one's research. Moving forward in technology, more advancements within the field of bioinformatics will continue to expand contributing to further success within biological research. The understanding of using these programs and being able to interpret the data is critical for comparing differences of organisms and their genes, as seen in the this project.

Purpose

Within this research project, we use a variety of bioinformatic programs to explore whether Mrub_1304 gene is orthologous with *E.coli* b_3189 gene, whether Mrub_2007 gene is orthologous with *E.coli* b_3972 gene, and whether Mrub_2006 gene is an ortholog of the *E.coli* b_0091 gene. Through the use of these bioinformatic tools, we will be able to identify the differences and similarities between these six genes. Due to the many similarities in nucleotide and amino acid sequence, and cellular localization between *E. coli* and *M. ruber* genes/proteins, I hypothesize that Mrub_1304,Mrub_2007,Mrub_2006 genes will be orthologs of the *E.coli*

b_3189*, E.coli* b_3972*, E.coli* b_0091 genes, respectively. In order to determine this hypothesis, knowledge of E-value significance was needed. The E-values help indicate whether or not the results are significant. A low E-value means that the two sequences being compared by a particular bioinformatics tool did not likely line up due to chance, which is the opposite for a high E-value. A low E-value is indicative of sequence similarity, which is assumed to indicate functional similarity. Prior to performing the research for the project, we conducted a BLAST search between Mrub_1304 gene and *E.coli* b_3189 gene, Mrub_2007 gene and *E.coli* b_3972 gene, and Mrub_2006 gene and *E.coli* b_0091 gene. The results yield very low E-values from these BLAST (2e-100, 5e-09, and 2e-81, respectively), which further lead us to our hypothesis (Madden 2002).

Methods

We used GENI-SCIENCE to research our genes within the peptidoglycan biosynthesis pathway to ensure our genes of interest had not been studied previously (Lori Scott 2017). We also utilized the GENI-ACT annotation website instructions, in order to collect data on the *E.coli* and *M.ruber* genes (Lori Scott 2017). To analysis the genes of interest in this project from *E. coli* and *M. ruber,* we used the following bioinformatics tools: BLAST (Madden 2002), CDD (Marchler *et al.*), T-Coffee (Notredame *et al.* 2000), WebLogo (Crooks *et al.* 2004), TMHMM (Krogh and Rapacki 2016), SignalP (Thomas *et al.* 2004), LipoP (Juncker *et al.* 2003), PSORT-B (Yu *et al.* 2010), Phobius (Kall *et al.* 2004), TIGRFAM (Haft *et al.* 2001), Pfam (Finn *et al.* 2016), PDB (Berman *et al.* 2000), IMG/EDU (Markowitz *et al.* 2012), KEGG (Kanehisa *et al.* 2016), and MetaCyc (Keseler *et al.* 2013). First, we performed a BLAST of Mrub_1304 gene vs. *E.coli* b_3189 gene, then Mrub_2007 gene vs. *E.coli* b_3972 gene, and finally *Mrub_2006* gene vs. *E.coli b_0091* gene to determine the similarities between the gene sets (Madden 2002). Once we were aware of the similarities between each of the sequences, we continued the research process by gathering the information needed for the different modules on the GENI-ACT site (Lori Scott 2017). We did this by using the appropriate bioinformatic programs suggested. After performing a BLAST search comparing the protein sequences, we then did a BLAST on each individual gene and selected the top 15 hits for *M.ruber* and top 10 hits for *E.coli* (Madden 2002). Then, we retrieved the CDD information for COG at the top of the BLAST results page (Marchler *et al.*). We used the top 15 hits for *M.ruber* and top 10 hits for *E.coli* from BLAST to input into the T-coffee tool, resulting in the multiple sequence alignment (Notredame *et al.* 2000). The multiple sequence alignment from T-coffee was used in the Weblogo tool to create the Sequence Logo (Crooks *et al.* 2004). Next, in order to find the location of the genes, we used TMHMM (Krogh and Rapacki 2016) , SignalP (Thomas *et al.* 2004), LipoP (Juncker *et al.* 2003), PSORT-B (Yu *et al.* 2010), and Phobius (Kall *et al.* 2004) by inserting the amino acid sequence into each of these programs. The TIGR (Haft *et al.* 2001), Pfam (Finn *et al.* 2016), and PDB (Berman *et al.* 2000) outputs were also retrieved through using

the amino acid sequence or gene locus tag. KEGG was used to see the genes present within the peptidoglycan biosynthesis pathway for each organism (Kanehisa *et al.* 2016). Then, MetaCyc was used to examine the reaction pathway for peptidoglycan biosynthesis, which is where the enzyme commission number was found for the genes (Keseler *et al.* 2013). The phyolgenic tree was formed in Phylogeny.fr with the T-coffee sequence (Notredame *et al.* 2000). For Gene Context, IMG/EDU was used to obtain the ortholog neighborhood of the organism as well as the gene content percentage (Markowitz *et al.* 2012). There were only a few deviations from the instructions within the course of our research. One deviation from the instructions was that we used the top 15 BLAST hits, instead of the recommended top 10 BLAST hits used for the T-coffee analysis. We choose the first 15 BLAST hits in *M.ruber* excluding the multispecies, and for *E.coli* we excluded *Escherichia coli,* then choose the first 10 BLAST hits (Madden 2002). The final modification we made included the use of the colored by KEGG for the Gene Context section of the Horizontal Transfer module (Kanehisa *et al.* 2016).

Table 1 summarizes the results of a variety of bioinformatics tools for *E.coli b_3189* gene and *Mrub_1304* gene. The information in the first row is the results of initial BLAST search discussed in the introduction (Madden 2002). The bit score is less informative. The two organism's protein lengths are slightly different and the two species are from different phyla, which means significant sequence divergence has likely occurred. However, the E-value of the BLAST alignment (2e-100) is more important; it indicates amino acid sequence of these enzymes cannot be attributed to chance, but the similarities are likely due to functional relatedness (Madden 2002). This evidence tells us that the two organisms are evolutionarily related and could share this gene. The second row indicates that both genes have the same COG number (COG0766) and name (*MurA*) with both having very low E-values showing significance from the CDD data tool (Marchler *et al.*). This also is evidence that both genes code for the same enzyme (UDP-N-acetylglucosamine 1-carboxyvinyltransferase) in the peptidoglycan biosynthesis pathway. Many of the bioinformatics tools used (i.e, TMHMM (Krogh and Rapacki 2016) , SignalP (Thomas *et al.* 2004), LipoP (Juncker *et al.* 2003), PSORT-B (Yu *et al.* 2010), and Phobius (Kall *et al.* 2004)) proposed that the cellular location of both genes is in the cytoplasm and there are no cleavage sites. TIGRfam showed that the genes have the same TIGR name (murA: UDP-N-acetylglucosamine 1-carboxyvinyl) and number (TIGR01072), as well as having very low E-values (Haft *et al.* 2001). Pfam also has the same Pfam name (EPSP synthase (3-phosphoshikimate-1-carboxyvinyltransferase)) and number (PF00275) for both genes (Finn *et al.* 2016). Along with the protein database giving different numbers (1UAE and 2F00) and names (Structure of UDP-N-acetylglucosamine enolpyruvyl transferase and 1.05 Angstrom Resolution Crystal Structure of UDP-N-acetylglucosamine 1-carboxyvinyltransferase from Acinetobacter baumannii in Covalently Bound Complex with (2R)-2-(phosphonooxy)propanoic Acid) (Berman *et al.* 2000). However, the difference in PDB numbers and names is that both protein sequences were crystallized from different organisms, but the same enzyme (*MurA*) was

crystallized in these organisms (Berman *et al.* 2000). Both genes also have the an enzyme commision number of E.C.2.5.1.7, derived from KEGG (Kanehisa *et al.* 2016). This collection of evidence is a strong indicator that *E.coli b_3189* gene and *Mrub_1304* gene are orthologs.

Results

Table 1. *E.coli* b_3189 gene and Mrub_1304 gene

Bioinformatics programs used	***E.coli* b_3189 gene**	**Mrub_1304 gene**
BLAST *E.coli vs. M.ruber*	**Score: 294** **E-value: 2e-100**	
CDD Data (COG category)	**COG Number: COG0766** **UDP-N-acetylglucosamine 1-carboxyvinyltransferase**	
	E-value: 0e0	**E-value: 3.24e-177**
Cellular Localization	**Cytoplasm**	
TIGRfam (Protein family)	**TIGR01072** **MurA: UDP-N-acetylglucosamine 1-carboxyvinyltransferase**	
	E-value: 1.3e-279	**E-value: 1.6e-175**
Pfam (Protein family)	**PF00275 (EPSP synthase (3-phosphoshikimate 1-carboxyvinyltransferase)**	
	E-value: 5.4e-131	**E-value: 9.1e-73**
Protein Database (PDB)	***E.coli* b_3189 gene: Structure of UDP-N-acetylglucosamine enolpyruvyl transferase** **Mrub_1304 gene: 1.05 Angstrom Resolution Crystal Structure of UDP-N-acetylglucosamine 1-carboxyvinyltransferase from Acinetobacter baumannii in Covalently Bound Complex with (2R)-2-(phosphonooxy)propanoic Acid.**	
	E-value: 1.7e-134	**E-value: 1.5e-150**

Enzyme commission number (E.C)	E.C.2.5.1.19- UDP-N-acetylglucosamine 1-carboxyvinyltransferase
KEGG pathway map	Peptidoglycan Biosynthesis Pathway

Table 2 summarizes the results of a variety of bioinformatics tools for *E.coli* b_3972 gene and Mrub_2007 gene. The information in the first row is the results of initial BLAST search discussed in the introduction (Madden 2002). The bit score is less informative. The two organism's protein lengths are slightly different and the two species are from different phyla, which means significant sequence divergence has likely occurred. However, the E-value of the BLAST alignment (5e-09) is more important; it indicates that the amino acid sequence of these enzymes cannot be attributed to chance, but the similarities are likely due to functional relatedness (Madden 2002). This evidence tells us that the two organisms are evolutionarily related and could share this gene. The second row indicates that both genes have the same COG number (COG0812) and name (*MurB*) with both having very low E-values showing significance from the CDD data tool (Marchler *et al.*). This also is evidence that both genes code for the same enzyme (UDP-N-acetylenolpyruvoylglucosamine reductase) in the peptidoglycan biosynthesis pathway. Many of the bioinformatics tools used (i.e, TMHMM (Krogh and Rapacki 2016) , SignalP (Thomas *et al.* 2004), LipoP (Juncker *et al.* 2003), PSORT-B (Yu *et al.* 2010), and Phobius (Kall *et al.* 2004)) proposed that the cellular location of both genes is in the cytoplasm and there are no cleavage sites. TIGRfam showed that the genes have the same TIGR name (murB: UDP-N-acetylenolpyruvoylglucosamine reductase) and number (TIGR00179), as well as having very low E-values (Haft *et al.* 2001). Pfam identified the same two Pfam names FAD binding domain and UDP-N-acetylenolpyruvoylglucosamine reductase,C-terminal domain, which correspond to the Pfam numbers PF01565 and PF02873 (Finn *et al.* 2016). The Protein Domain Database (PDB) identified sequence similarity to the same crystallized protein, which is numbered 2GQT and 1MBB and named UDP-N-acetylenolpyruvylglucosamine reductase (MurB) from *Thermus caldophilus* and (E)-enolbutyryl-UDP-N-acetylglucosamine as a mechanistic probe of UDP-N-acetylenolpyruvylglucosamine reductase (*MurB*) (Berman *et al.* 2000). However, the difference in PDB numbers and names is explained by the fact that both protein sequences were crystallized from different organisms, but the same enzyme (*MurB*) was crystallized in these organisms (Berman *et al.* 2000). Both genes also have the an Enzyme Commision number of E.C.1.3.1.98, as determined by KEGG (Kanehisa *et al.* 2016). This collection of evidence is a strong indicator that *E.coli* b_3972 gene and Mrub_2007 gene are orthologs.

Table 2. *E.coli* b_3972 gene and Mrub_2007 gene

Bioinformatics programs used	*E.coli* b_3972 gene	Mrub_2007 gene
BLAST *E.coli vs. M.ruber*	Score: 42.4 E-value: 5e-09	
CDD Data (COG category)	COG Number: COG0812 MurB:UDP-N-acetylenolpyruvoylglucosamine reductase	
	E-value: 5.47e-136	E-value: 2.40e-81
Cellular Localization	Cytoplasm	
TIGRfam (Protein family)	TIGR00179 MurB: UDP-N-acetylenolpyruvoylglucosamine reductase	
	E-value: 6e-163	E-value: 2.5e-35
Pfam (Protein family)	PF01565 (FAD binding domain) PF02873 (UDP-N-acetylenolpyruvoylglucosamine reductase,C-terminal domain)	
	E-value: 7.7e-27 5.5e-35	E-value: 1.1e-08 3.7e-19
Protein Database (PDB)	*E. coli* b_3972 gene: (E)-enolbutyryl-UDP-N-acetylglucosamine as a mechanistic probe of UDP-N-acetylenolpyruvylglucosamine reductase (MurB) Mrub_2007 gene: Crystal Structure of UDP-N-acetylenolpyruvylglucosamine reductase (MurB) from Thermus caldophilus	
	E-value: 2.7e-123	E-value: 1.0e-60
Enzyme commission number (E.C)	E.C.1.3.1.98- UDP-N-acetylenolpyruvoylglucosamine reductase	

KEGG pathway map	Peptidoglycan Biosynthesis Pathway

Table 3 summarizes the results of a variety of bioinformatics tools for *E.coli* b_0091 gene and Mrub_2006 gene. The information in the first row is the results of initial BLAST search discussed in the introduction (Madden 2002). The bit score is less informative. The two organism's protein lengths are slightly different and the two species are from different phyla, which means significant sequence divergence has likely occurred. However, the E-value of the BLAST alignment (2e-81) is more important; it indicates that the protein sequences of the two organisms are not aligned by chance; they share many of the same amino acids (Madden 2002). This evidence tells us that the two organisms are evolutionarily related and could share this gene. The second row indicates that both genes have the same COG number (COG0773) and name (*MurC*) with both having very low E-values showing significance from the CDD data tool (Marchler *et al.*). This also is evidence that both genes code for the same enzyme (UDP-N-acetylmuramate--L-alanine ligase) in the peptidoglycan biosynthesis pathway. Many of the bioinformatics tools used (i.e, TMHMM (Krogh and Rapacki 2016) , SignalP (Thomas *et al.* 2004), LipoP (Juncker *et al.* 2003), PSORT-B (Yu *et al.* 2010), and Phobius (Kall *et al.* 2004)) proposed that the cellular location of both genes is in the cytoplasm and there are no cleavage sites. TIGRfam showed that the genes have the same TIGR name (murC: UDP-N-acetylmuramate--alanine ligase) and number (TIGR01082), as well as having very low E-values (Haft *et al.* 2001). Pfam identified the same two Pfam names Mur ligase family, catalytic domain and Mur ligase middle domain, which correspond to the Pfam numbersPF01225 and PF08245 (Finn *et al.* 2016). The Protein Domain Database (PDB) identified sequence similarity to the same crystallized protein, which are numbered 4HV4 and and named 2.25 Angstrom resolution crystal structure of UDP-N-acetylmuramate--L-alanine ligase (*MurC*) from Yersinia pestis CO92 in complex with AMP and *Escherichia coli MurC*. However, the difference in PDB numbers and names is explained by the fact that both protein sequences were crystallized from different organisms, but the same enzyme (*MurC*) was crystallized in these organisms (Berman *et al.* 2000). Both genes also have the an Enzyme Commision number of E.C.6.3.2.8, as determined by KEGG (Kanehisa *et al.* 2016). This collection of evidence is a strong indicator that *E.coli* b_0091 gene and Mrub_2006 gene are orthologs.

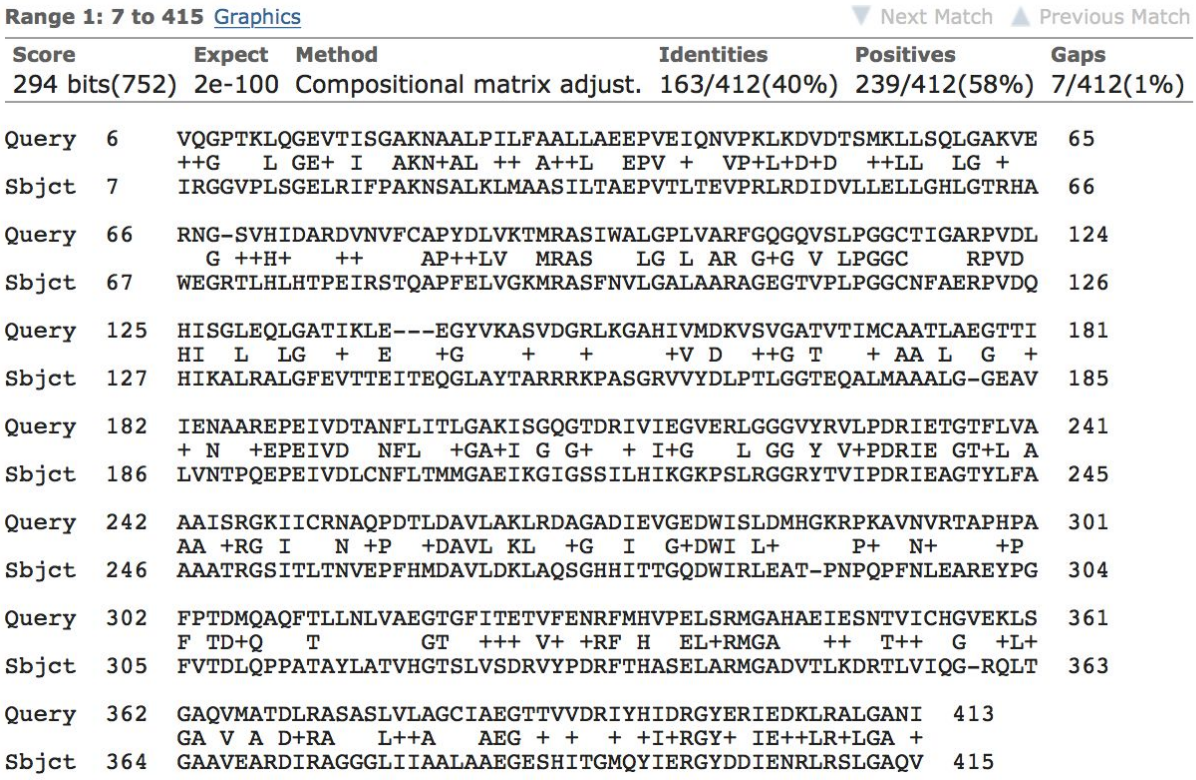
Table 3. *E.coli* b_0091 gene and Mrub_2006 gene

Bioinformatics programs used	*E.coli* b_0091 gene	Mrub_2006 gene

BLAST *E.coli* *vs. M.ruber*	**Score: 248** **E-value: 2e-81**	
CDD Data (COG category)	**COG Number: COG0773** **MurC: UDP-N-acetylmuramate--L-alanine ligase**	
	E-value: 0e+00	**E-value: 3.24e-177**
Cellular Localization	**Cytoplasm**	
TIGRfam (Protein family)	**TIGR01082** **MurC: UDP-N-acetylmuramate--L-alanine ligase**	
	E-value: 9.1e-285	**E-value: 3.9e-136**
Pfam (Protein family)	**PF01225 (FAD binding domain)** **PF08245 (Mur ligase middle domain)**	
	E-value: **9e-30** **5.7e-28**	**E-value:** **3.4e-20** **3.6e-19**
Protein Database (PDB)	***E.coli* b_0091 gene:** **Escherichia coli MurC** **Mrub_2006 gene:** **2.25 Angstrom resolution crystal structure of UDP-N-acetylmuramate--L-alanine ligase (murC) from Yersinia pestis CO92 in complex with AMP**	
	E-value: 1.1e-161	**E-value: 1.9e-41**
Enzyme commission number (E.C)	**E.C.6.3.2.8- UDP-N-acetylmuramate--L-alanine ligase**	
KEGG pathway map	**Peptidoglycan Biosynthesis Pathway**	

Figure 3 is a depiction of the BLAST between M.rub_1304 gene and *E. coli* b_3189 gene performed prior to starting the modules in GENI-ACT. The data shows 40% identities, meaning 40% of the amino acids were the same between the two protein sequences. The E-value

expressed by this BLAST is immensely low (2e-100). This very low E-value is an indicator that the two sequence did not align by random chance. The results of the BLAST served as the initial piece of evidence that the genes M.rub_1304 and *E. coli* b_3189 could be possible orthologs.



Range 1: 7 to 415 Graphics ▼ Next Match ▲ Previous Match

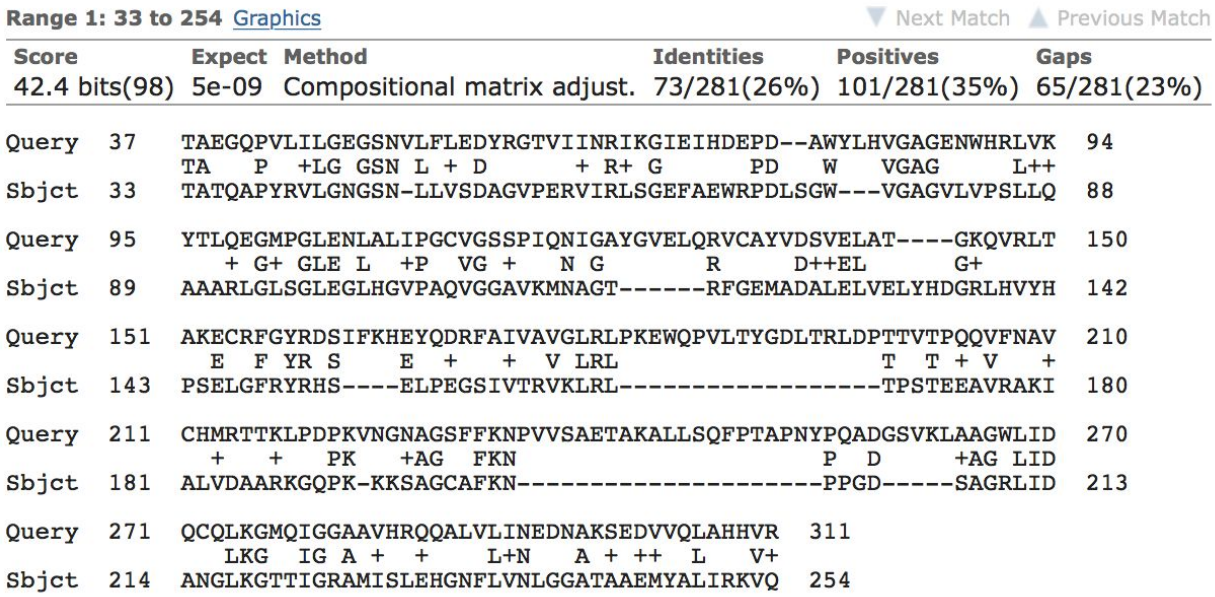
Score	Expect	Method	Identities	Positives	Gaps
294 bits(752)	2e-100	Compositional matrix adjust.	163/412(40%)	239/412(58%)	7/412(1%)

```
Query  6    VQGPTKLQGEVTISGAKNAALPILFAALLAEEPVEIQNVPKLKDVDTSMKLLSQLGAKVE  65
            ++G   L GE+ I  AKN+AL ++ A++L  EPV +  VP+L+D+D  ++LL  LG +
Sbjct  7    IRGGVPLSGELRIFPAKNSALKLMAASILTAEPVTLTEVPRLRDIDVLLELLGHLGTRHA  66

Query  66   RNG-SVHIDARDVNVFCAPYDLVKTMRASIWALGPLVARFGQGQVSLPGGCTIGARPVDL  124
             G ++H+   ++    AP++LV  MRAS   LG L AR G+G V LPGGC    RPVD
Sbjct  67   WEGRTLHLHTPEIRSTQAPFELVGKMRASFNVLGALAARAGEGTVPLPGGCNFAERPVDQ  126

Query  125  HISGLEQLGATIKLE---EGYVKASVDGRLKGAHIVMDKVSVGATVTIMCAATLAEGTTI  181
            HI  L  LG  +  E   +G    +   +     +V D  ++G T   + AA L  G  +
Sbjct  127  HIKALRALGFEVTTEITEQGLAYTARRRKPASGRVVYDLPTLGGTEQALMAAALG-GEAV  185

Query  182  IENAAREPEIVDTANFLITLGAKISGQGTDRIVIEGVERLGGGVYRVLPDRIETGTFLVA  241
            + N  +EPEIVD  NFL  +GA+I G G+  + I+G   L GG Y V+PDRIE GT+L A
Sbjct  186  LVNTPQEPEIVDLCNFLTMMGAEIKGIGSSILHIKGKPSLRGGRYTVIPDRIEAGTYLFA  245

Query  242  AAISRGKIICRNAQPDTLDAVLAKLRDAGADIEVGEDWISLDMHGKRPKAVNVRTAPHPA  301
            AA +RG I   N +P  +DAVL KL  +G  I  G+DWI L+     P+  N+    +P
Sbjct  246  AAATRGSITLTNVEPFHMDAVLDKLAQSGHHITTGQDWIRLEAT-PNPQPFNLEAREYPG  304

Query  302  FPTDMQAQFTLLNLVAEGTGFITETVFENRFMHVPELSRMGAHAEIESNTVICHGVEKLS  361
            F TD+Q   T       GT  +++ V+ +RF H  EL+RMGA   ++  T++  G  +L+
Sbjct  305  FVTDLQPPATAYLATVHGTSLVSDRVYPDRFTHASELARMGADVTLKDRTLVIQG-RQLT  363

Query  362  GAQVMATDLRASASLVLAGCIAEGTTVVDRIYHIDRGYERIEDKLRALGANI  413
            GA V A D+RA   L++A   AEG + +  + +I+RGY+ IE++LR+LGA +
Sbjct  364  GAAVEARDIRAGGGLIIAALAAEGESHITGMQYIERGYDDIENRLRSLGAQV  415
```

Figure 3. M.rub_1304 gene and *E. coli* b_3189 gene have similar protein sequences. This analysis was performed using NCBI BLAST bioinformatics program at http://www.ncbi.nlm.nih.gov. The query sequence is *E. coli* b_3189 gene protein sequence. The subject is M.rub_1304 gene protein sequence.

Figure 4 is a depiction of the BLAST between M.rub_2007 gene and *E. coli* b_3972 gene performed prior to starting the modules in GENI-ACT. The data shows 26% identities, meaning 26% of the amino acids were the same between the two protein sequences. The E-value expressed by this BLAST is low (5e-09). This low E-value is an indicator that the two sequence

did not align by random chance. The results of the BLAST served as the initial piece of evidence that the genes M.rub_2007 and *E. coli* b_3972 could be possible orthologs.



Range 1: 33 to 254 Graphics ▼ Next Match ▲ Previous Match

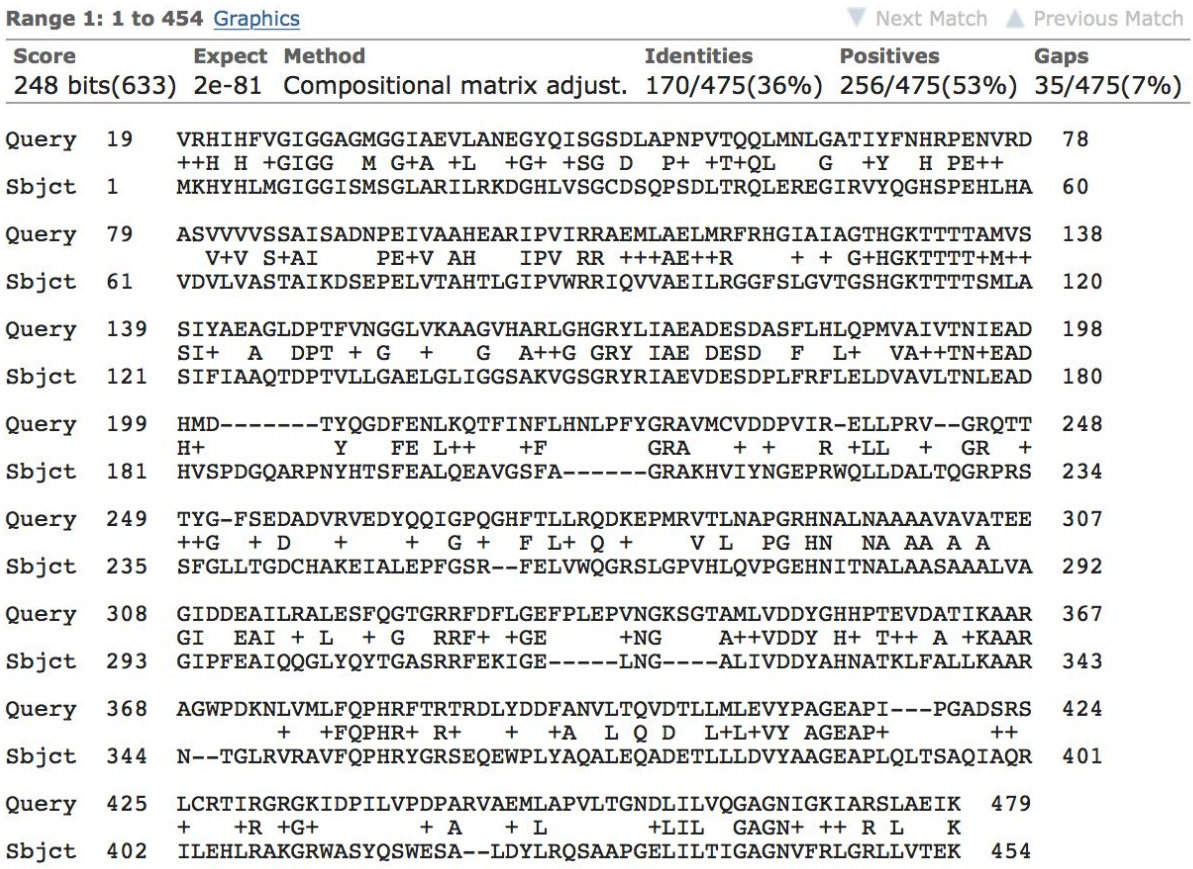
Score	Expect	Method	Identities	Positives	Gaps
42.4 bits(98)	5e-09	Compositional matrix adjust.	73/281(26%)	101/281(35%)	65/281(23%)

```
Query  37   TAEGQPVLILGEGSNVLFLEDYRGTVIINRIKGIEIHDEPD--AWYLHVGAGENWHRLVK  94
            TA    P  +LG GSN L + D      + R+ G      PD   W   VGAG     L++
Sbjct  33   TATQAPYRVLGNGSN-LLVSDAGVPERVIRLSGEFAEWRPDLSGW---VGAGVLVPSLLQ  88

Query  95   YTLQEGMPGLENLALIPGCVGSSPIQNIGAYGVELQRVCAYVDSVELAT----GKQVRLT  150
               + G+ GLE L  +P  VG +   N G       R     D++EL      G+
Sbjct  89   AAARLGLSGLEGLHGVPAQVGGAVKMNAGT------RFGEMADALELVELYHDGRLHVYH  142

Query  151  AKECRFGYRDSIFKHEYQDRFAIVAVGLRLPKEWQPVLTYGDLTRLDPTTVTPQQVFNAV  210
               E  F YR S    E  +   +  V LRL                   T  T + V   +
Sbjct  143  PSELGFRYRHS----ELPEGSIVTRVKLRL-----------------TPSTEEAVRAKI  180

Query  211  CHMRTTKLPDPKVNGNAGSFFKNPVVSAETAKALLSQFPTAPNYPQADGSVKLAAGWLID  270
               +   +   PK   +AG  FKN                      P  D     +AG LID
Sbjct  181  ALVDAARKGQPK-KKSAGCAFKN---------------------PPGD-----SAGRLID  213

Query  271  QCQLKGMQIGGAAVHRQQALVLINEDNAKSEDVVQLAHHVR  311
               LKG  IG A +  +    L+N   A + ++  L   V+
Sbjct  214  ANGLKGTTIGRAMISLEHGNFLVNLGGATAAEMYALIRKVQ  254
```

Figure 4. M.rub_2007 gene and *E. coli* b_3972 gene have similar protein sequence. This analysis was performed using NCBI BLAST bioinformatics program at http://www.ncbi.nlm.nih.gov. The query sequence is *E. coli b_3972* gene protein sequence. The subject is *M.rub_2007* gene protein sequence.

Figure 5 is a depiction of the BLAST between M.rub_2006 gene and *E. coli* b_0091 gene performed prior to starting the modules in GENI-ACT. The data shows 36% identities, meaning 36% of the amino acids were the same between the two protein sequences. The E-value expressed by this BLAST is low (2e-81). This low E-value is an indicator that the two sequence

did not align by random chance. The results of the BLAST served as the initial piece of evidence that the genes M.rub_2006 and E. coli b_0091 could be possible orthologs.



Range 1: 1 to 454 Graphics — Next Match — Previous Match

Score	Expect	Method	Identities	Positives	Gaps
248 bits(633)	2e-81	Compositional matrix adjust.	170/475(36%)	256/475(53%)	35/475(7%)

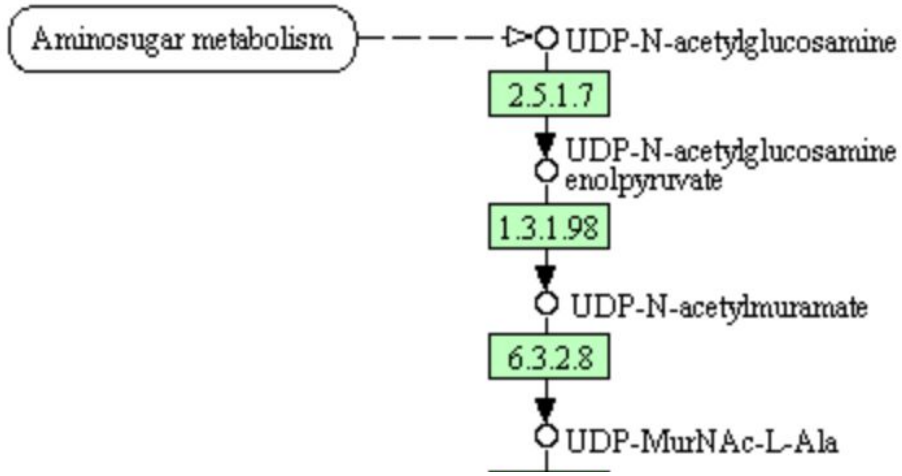
```
Query  19   VRHIHFVGIGGAGMGGIAEVLANEGYQISGSDLAPNPVTQQLMNLGATIYFNHRPENVRD  78
            ++H H +GIGG  M G+A +L  +G+ +SG D  P+ +T+QL   G  +Y  H PE++
Sbjct  1    MKHYHLMGIGGISMSGLARILRKDGHLVSGCDSQPSDLTRQLEREGIRVYQGHSPEHLHA  60

Query  79   ASVVVVSSAISADNPEIVAAHEARIPVIRRAEMLAELMRFRHGIAIAGTHGKTTTTAMVS  138
              V+V S+AI    PE+V AH   IPV RR +++AE++R    + + G+HGKTTTT+M++
Sbjct  61   VDVLVASTAIKDSEPELVTAHTLGIPVWRRIQVVAEILRGGFSLGVTGSHGKTTTTSMLA  120

Query  139  SIYAEAGLDPTFVNGGLVKAAGVHARLGHGRYLIAEADESDASFLHLQPMVAIVTNIEAD  198
            SI+  A  DPT + G  +   G  A++G GRY IAE DESD  F  L+  VA++TN+EAD
Sbjct  121  SIFIAAQTDPTVLLGAELGLIGGSAKVGSGRYRIAEVDESDPLFRFLELDVAVLTNLEAD  180

Query  199  HMD-------TYQGDFENLKQTFINFLHNLPFYGRAVMCVDDPVIR-ELLPRV--GRQTT  248
            H+          Y   FE L++   +F       GRA   + +   R +LL  +  GR  +
Sbjct  181  HVSPDGQARPNYHTSFEALQEAVGSFA------GRAKHVIYNGEPRWQLLDALTQGRPRS  234

Query  249  TYG-FSEDADVRVEDYQQIGPQGHFTLLRQDKEPMRVTLNAPGRHNALNAAAAVAVATEE  307
            ++G  + D   +    +  G +  F L+ Q +    V L  PG HN  NA AA A A
Sbjct  235  SFGLLTGDCHAKEIALEPFGSR--FELVWQGRSLGPVHLQVPGEHNITNALAASAAALVA  292

Query  308  GIDDEAILRALESFQGTGRRFDFLGEFPLEPVNGKSGTAMLVDDYGHHPTEVDATIKAAR  367
            GI  EAI + L  + G  RRF+ +GE     +NG    A++VDDY H+ T++ A +KAAR
Sbjct  293  GIPFEAIQQGLYQYTGASRRFEKIGE-----LNG----ALIVDDYAHNATKLFALLKAAR  343

Query  368  AGWPDKNLVMLFQPHRFTRTRDLYDDFANVLTQVDTLLMLEVYPAGEAPI---PGADSRS  424
                  +  +FQPHR+ R+   +  +A  L Q D  L+L+VY AGEAP+       ++
Sbjct  344  N--TGLRVRAVFQPHRYGRSEQEWPLYAQALEQADETLLLDVYAAGEAPLQLTSAQIAQR  401

Query  425  LCRTIRGRGKIDPILVPDPARVAEMLAPVLTGNDLILVQGAGNIGKIARSLAEIK  479
            +   +R +G+       + A   + L       +LIL  GAGN+ ++ R L   K
Sbjct  402  ILEHLRAKGRWASYQSWESA--LDYLRQSAAPGELILTIGAGNVFRLGRLLVTEK  454
```

Figure 5. M.rub_2006 gene and *E. coli* b_0091 gene have similar protein sequence. This analysis was performed using NCBI BLAST bioinformatics program at http://www.ncbi.nlm.nih.gov. The query sequence is *E. coli* b_3972 gene protein sequence. The subject is M.rub_2006 gene protein sequence.

Figure 6 shows the three enzymes that code for the genes of interest in this project, UDP-N-acetylglucosamine 1-carboxyvinyltransferase, UDP-N-acetylenolpyruvoylglucosamine reductase and UDP-N-acetylmuramate--L-alanine ligase, within both organisms. The enzymes colored green represent the presence of that enzyme in the organism. The enzyme commission

numbers within green highlighted box are the same in *Meiothermus ruber DSM 1279* and *Escherichia coli K-12 MG1655.* This was the second piece of evidence suggesting that the genes in these two different organisms are orthologous.

Panel A



Panel B

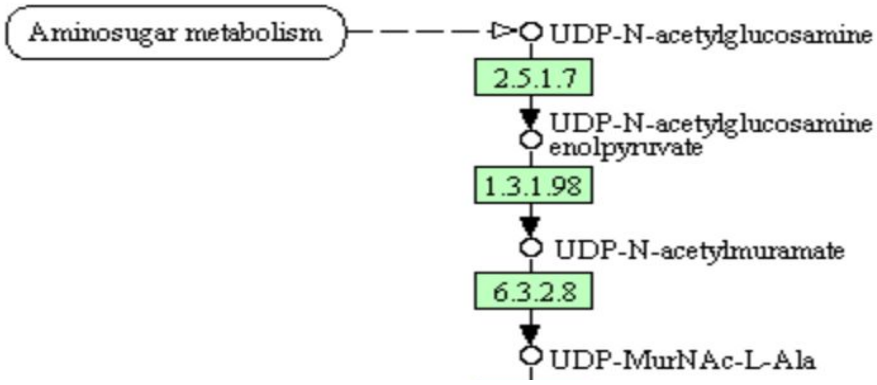
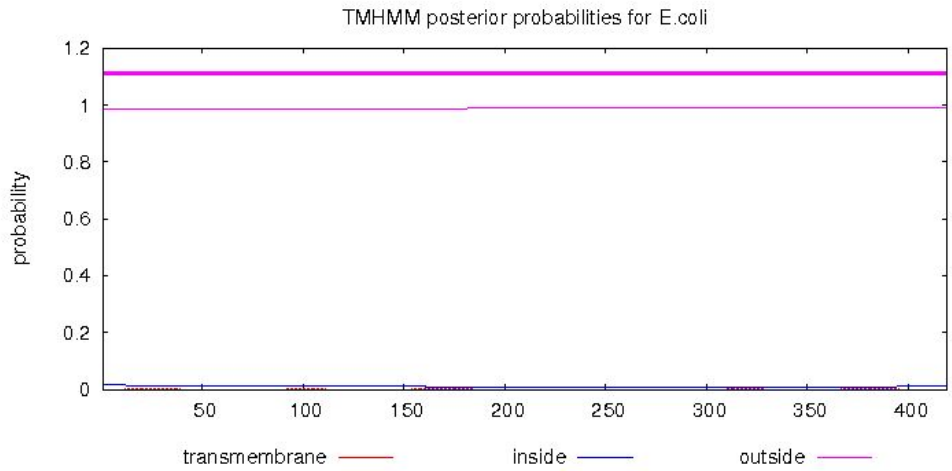


Figure 6. *Meiothermus ruber DSM 1279* and *Escherichia coli K-12 MG1655* genes of interest present in the peptidoglycan biosynthesis pathway. Panel A identifies the genes of interest in peptidoglycan biosynthesis pathway within *Escherichia coli K-12 MG1655.* Panel B identifies the genes of interest in peptidoglycan biosynthesis pathway within *Meiothermus ruber DSM 1279.* Available from: http://www.genome.jp/kegg/pathway.html

The charts in figures 7 shows the TMH hydropathy plots for both *E. coli* b_3189 and Mrub_1304. Red peaks indicate the transmembrane helices present, in which we have none. Therefore, this suggest that *E. coli* b_3189 and Mrub_1304 genes code for the same proteins in the cytoplasm instead of the membrane of these two organisms.

Panel A



Panel B

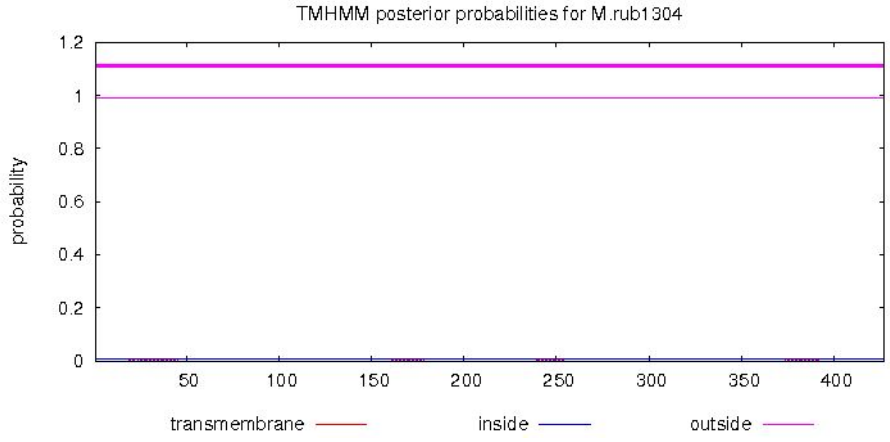
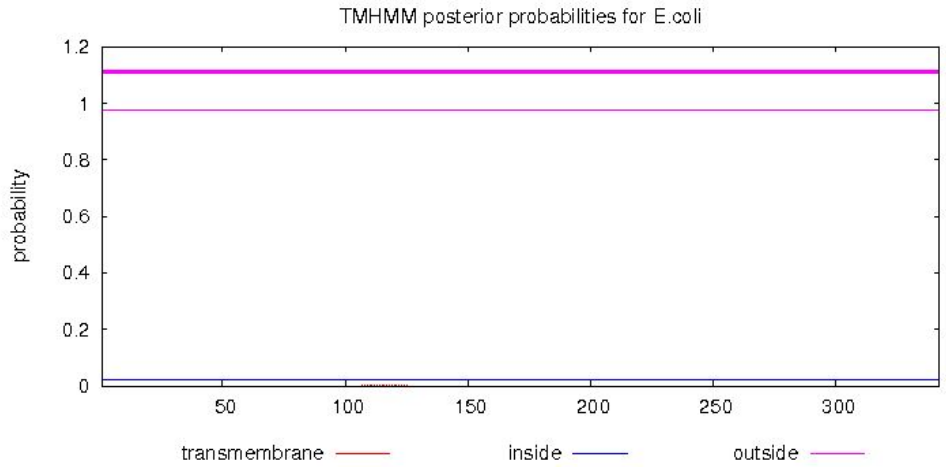


Figure 7. *E. coli* b_3189 and Mrub_1304 do not consist of TMH (transmembrane helices), which indicates a cytoplasmic location for these two proteins. Panel A displays the TMHMM for *E. coli* b_3189. Panel B shows the TMHMM for Mrub_1304. TMHMM Server v 2., bioinformatics program was used to create these two chart visuals. Avaliable from: http://www.cbs.dtu.dk/services/TMHMM/.

The charts in figures 8 shows the TMH hydropathy plots for both *E.coli* b_3972 and Mrub_2007. Red peaks indicate the transmembrane helices present, in which we have none. Therefore, this

suggest that *E.coli* b_3972 and Mrub_2007 genes code for the same proteins in the cytoplasm instead of the membrane of these two organisms.

Panel A



Panel B

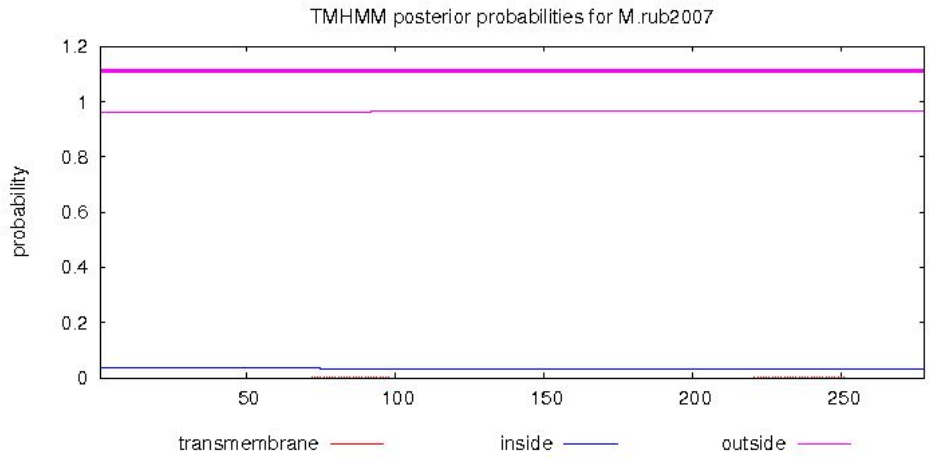
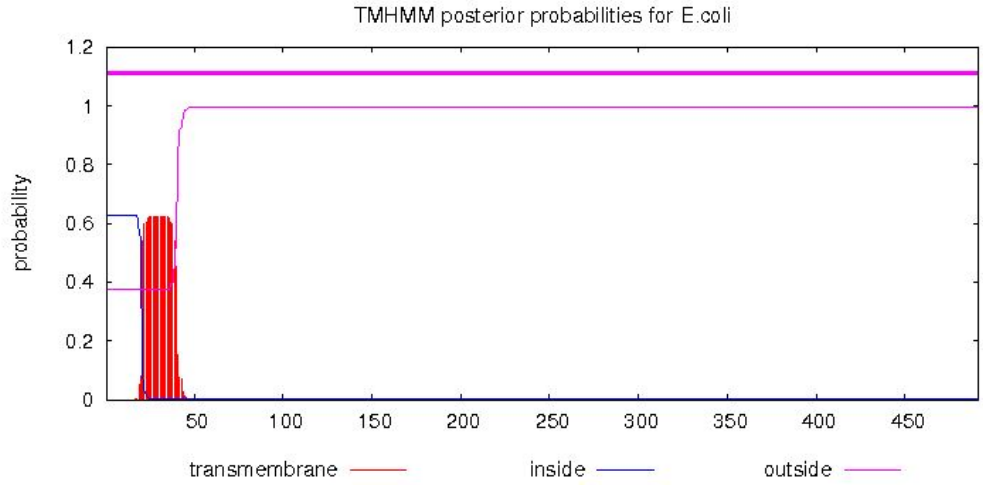


Figure 8. *E. coli* b_3972 and Mrub_2007 do not consist of TMH, which indicates a cytoplasmic location for these two proteins. Panel A displays the TMHMM for *E. coli* b_3972. Panel B shows the TMHMM for Mrub_2007. TMHMM Server v 2., bioinformatics program was used to create these two chart visuals. Avaliable from: http://www.cbs.dtu.dk/services/TMHMM/.

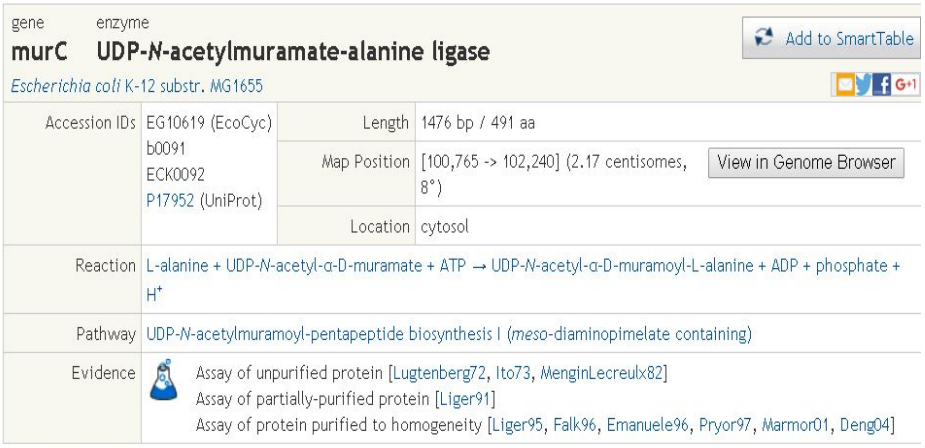
The charts in figures 9 shows the TMH hydropathy plots for both *E. coli* b_3189 and Mrub_1304. Red peaks indicates the presentence of a transmembrane helices present, but in *M.*

ruber. The red peak on the TMHMM of *E. coli* b0091 can be explained as follows. On the Ecocyc page for this enzyme, b0091 is identified as being cytosolic, which means there is experimental evidence to support this location (Keseler *et al.* 2013). Secondly, the TMHMM output for b0091 indicates 0 TMH (2nd line of TMH image). Thirdly, the HELP instructions for the TMHMM site explains that 18 or more amino acids must be found in N-terminal hydrophobic region to be a potential TMH; b0091 has only 12 aa (See line 3rd line of TMH output). Consequently, while this is clearly a short region of hydrophobic amino acids at the N-terminus of b0091, it is unlikely to be a signal peptide or transmembrane helices region. This refutes the idea of *E. coli* b_0091 being in a different location in the cell, other than the cytoplasm. Therefore, this suggest that *E. coli* b_0091 and Mrub_1304 genes are coded for in the cytoplasm.

Panel A




```
# E.coli Length: 491
# E.coli Number of predicted TMHs:  0
# E.coli Exp number of AAs in TMHs: 12.52848
# E.coli Exp number, first 60 AAs:  12.49834
# E.coli Total prob of N-in:        0.62573
# E.coli POSSIBLE N-term signal sequence
E.coli  TMHMM2.0       outside     1   491
```



Panel B

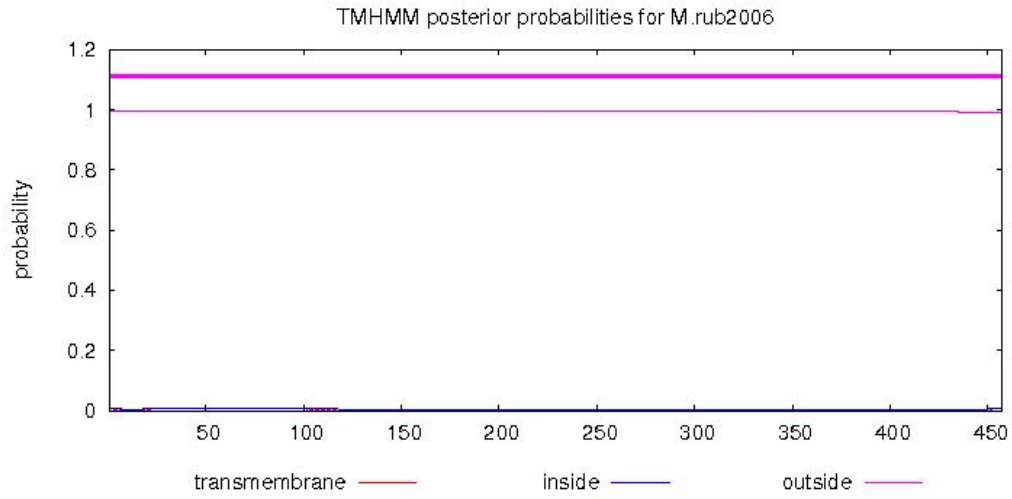
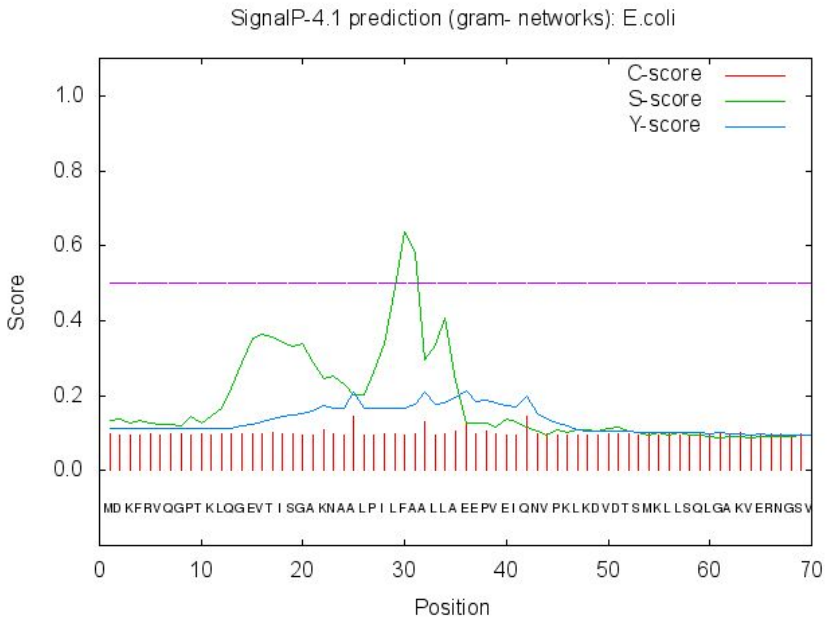


Figure 9. *E. coli* b_0091 consist of one TMH and Mrub_2006 does not consist of TMH, which indicates the location for these two proteins, which is the cytoplasm which is proven by EcoCyc (Keseler *et al.* 2013). Panel A displays the TMHMM for *E. coli* b_0091 as well as the EcoCyc page for this enzyme (Keseler *et al.* 2013). Panel B shows the TMHMM for Mrub_2006. TMHMM Server v 2., bioinformatics program was used to create these two chart visuals. Avaliable from: http://www.cbs.dtu.dk/services/TMHMM/.

The plots shown above in figure 10 are Signal P graph plots created for *E. coli* b_3189 and Mrub_1304. The purple line in these plots is the D value cutoff, which is calculated along with the S-score and Y-score to form a D value for a gene, which is used to predict protein cleavage sites. For both *E. coli* b_3189 and Mrub_1304 the D value (0.237,0.193) is below the cutoff

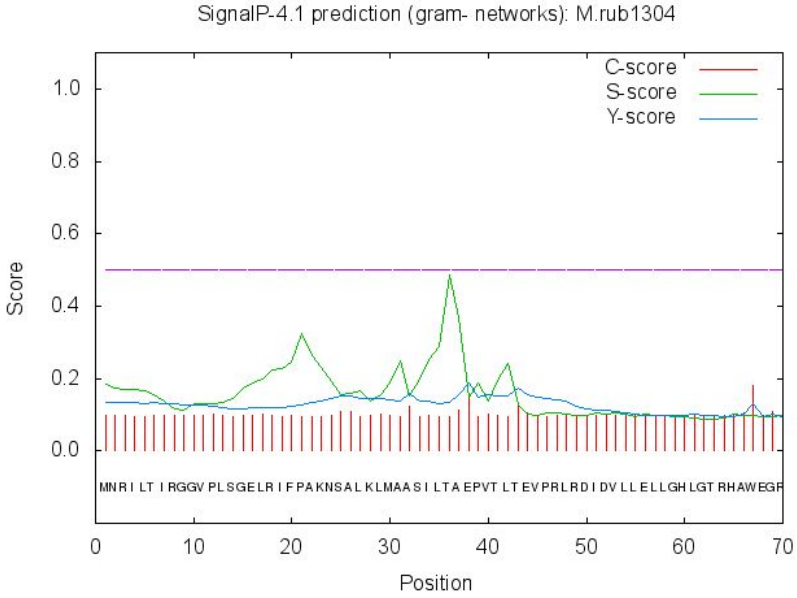
value (0.570, 0.570), respectively. This shows a constant feature in both genes, expressing that neither contain protein cleavage sites.

Panel A




```
D=0.237 D-cutoff=0.570 Networks=SignalP-noTM
```

Panel B

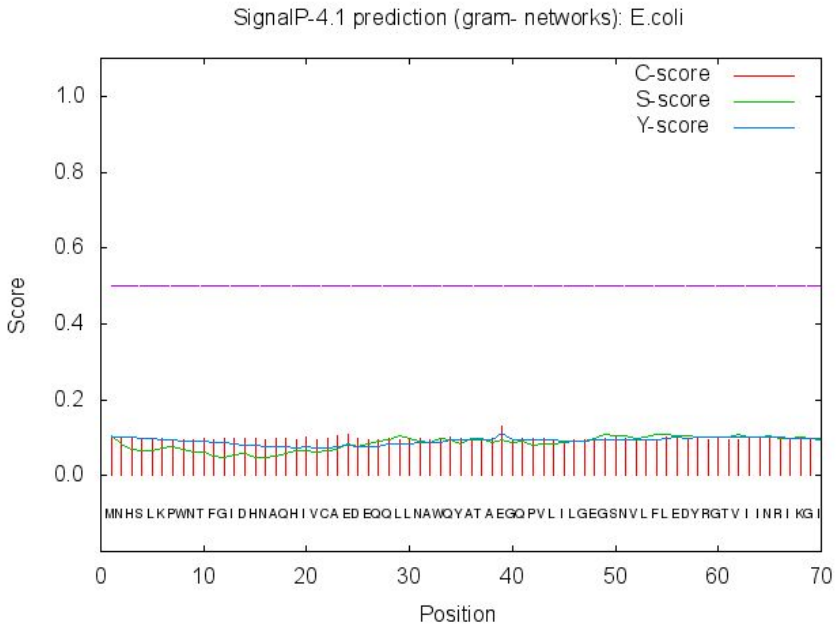


D=0.193 D-cutoff=0.570 Networks=SignalP-noTM

Figure 10. *E. coli* b_3189 and Mrub_1304 do not contain cleavage sites, which likely indicates a cytoplasmic location for these two proteins. The D values present in this figure for panel A and B are below the cutoff value. Panel A displays the plot for *E. coli* b_3189. Panel B shows the plot for Mrub_1304. Signal P server v 4.1 generated these plots. Available from: http://www.cbs.dtu.dk/services/SignalP/.

The plots shown above in figure 11 are Signal P graph plots created for *E. coli* b_3972 and Mrub_2007. The purple line in these plots is the D value cutoff, which is calculated along with the S-score and Y-score to form a D value for a gene, which is used to predict protein cleavage sites. For both *E. coli* b_3972 and Mrub_2007 the D value (0.094,0.260) is below the cutoff value (0.570, 0.570), respectively. This shows a constant feature in both genes, expressing that neither contain protein cleavage sites.

Panel A



D=0.094 D-cutoff=0.570 Networks=SignalP-noTM

Panel B

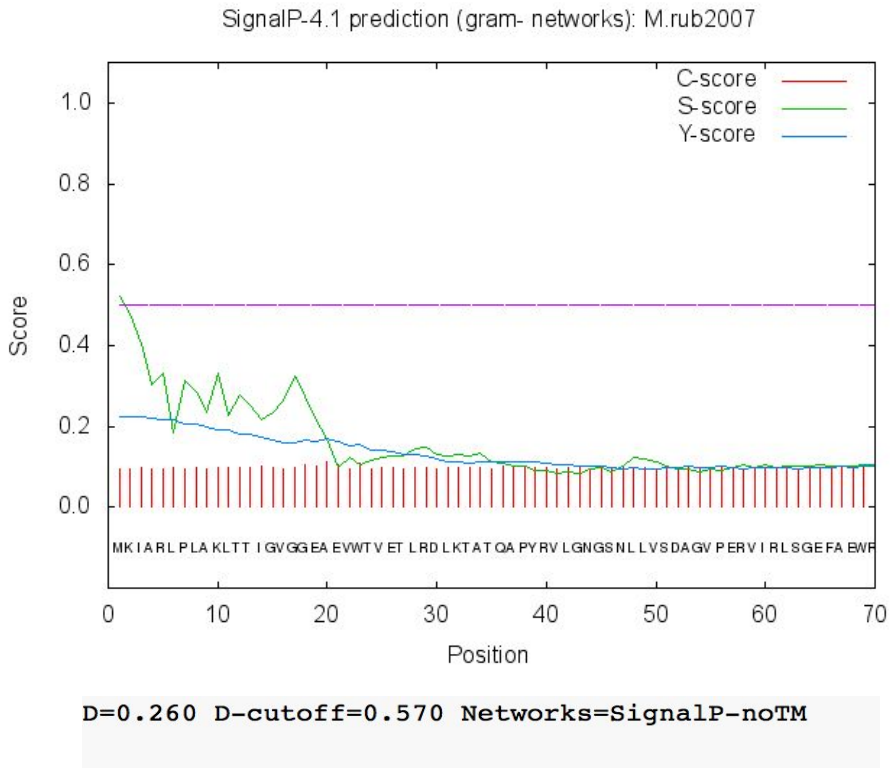
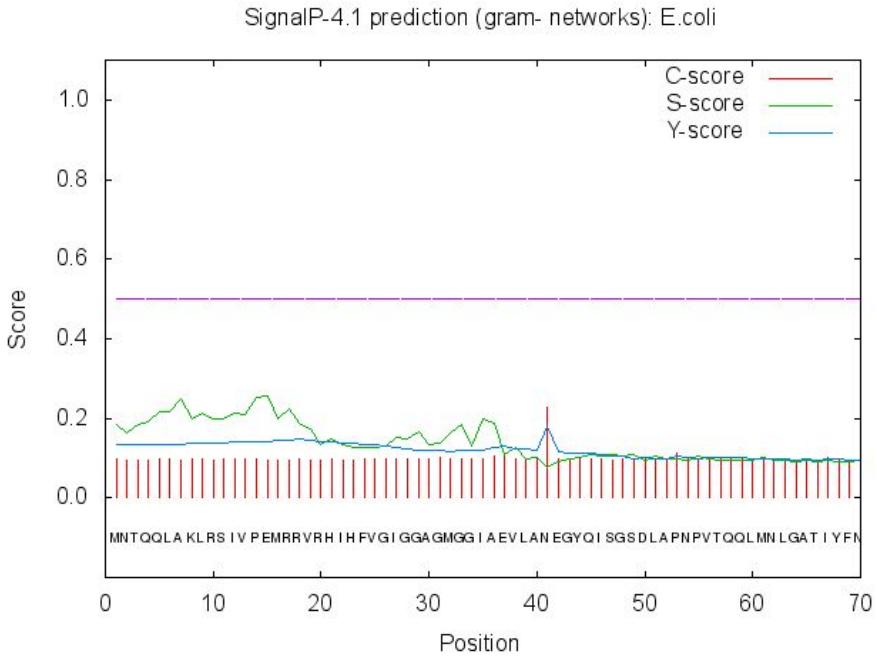


Figure 11. *E. coli* b_3972 and Mrub_2007 do not contain cleavage sites, which likely indicates a cytoplasmic location for these two proteins. The D values present in this figure for panel A and B are below the cutoff value. Panel A displays the plot for *E. coli* b_3972. Panel B shows the plot for Mrub_2007. Signal P server v 4.1 generated these plots. Available from: http://www.cbs.dtu.dk/services/SignalP/.

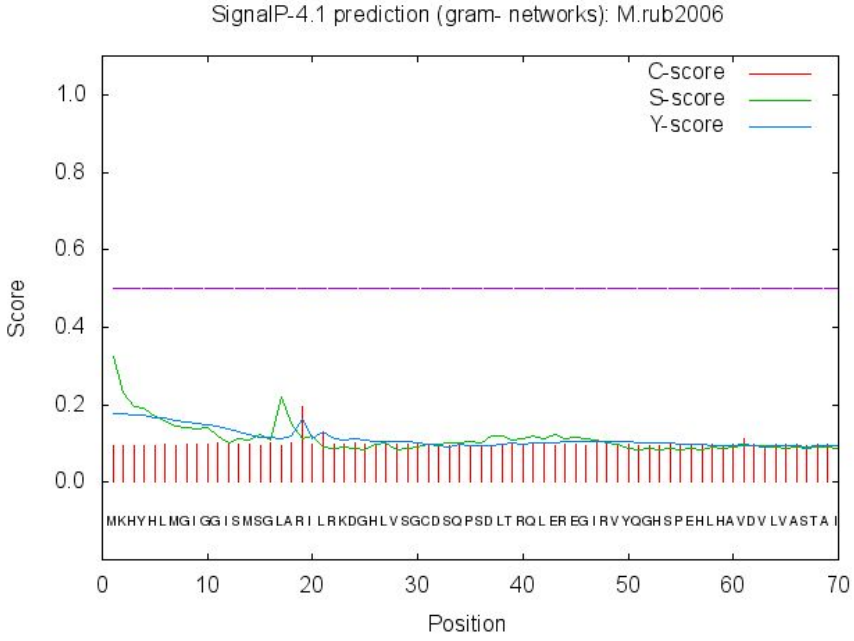
The plots shown above in figure 12 are Signal P graph plots created for *E. coli* b_0091 and Mrub_2006. The purple line in these plots is the D value cutoff, which is calculated along with the S-score and Y-score to form a D value for a gene, which is used to predict protein cleavage sites. For both *E. coli* b_0091 and Mrub_2006, the D value (0.177,0.162) is below the cutoff value (0.570, 0.570), respectively. This shows a constant feature in both genes, expressing that neither contain protein cleavage sites.

Panel A



D=0.177 D-cutoff=0.570 Networks=SignalP-noTM

Panel B

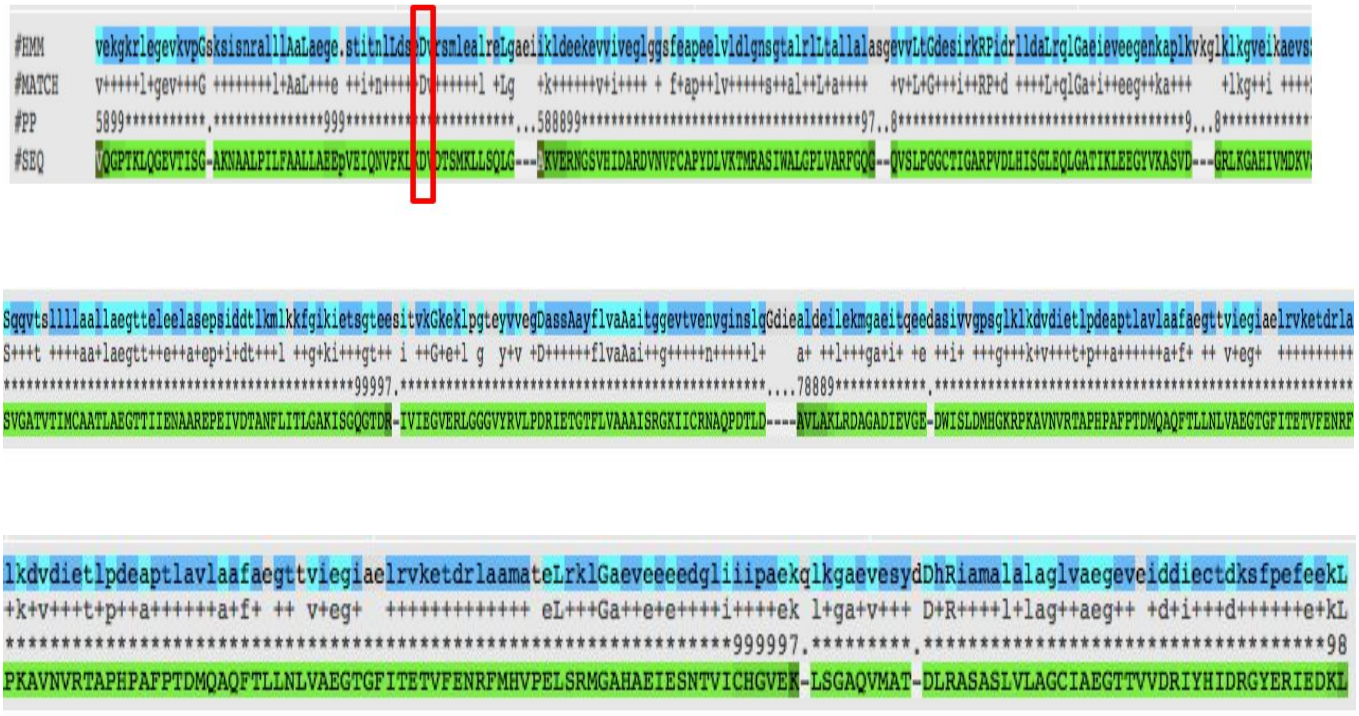


D=0.162 D-cutoff=0.570 Networks=SignalP-noTM

Figure 12. *E. coli* b_0091 and Mrub_2006 do not contain cleavage sites, which likely indicates a cytoplasmic location for these two proteins. The D values present in this figure for panel A and B are below the cutoff value. Panel A displays the plot for *E. coli* b_0091. Panel B shows the plot for Mrub_2006. Signal P server v 4.1 generated these plots. Available from: http://www.cbs.dtu.dk/services/SignalP/.

Looking at the pairwise alignments in figure 13, it is clear that *E. coli* b_3189 and Mrub_1304 have the same highly conserved aspartate, glycine, and proline residues toward the beginning and middle of the protein sequence at positions 44, 113, and 121, respectively. Aspartate will be the amino acid that is used from Mrub_1304 for site-directed mutagenesis to create a mutation. As opposed to comparing the two protein sequences against each other, like performed in BLAST, these pairwise alignments compare our protein sequence to a consensus sequence, comprised of multiple other proteins (Madden 2002). Because this data from *E. coli* b_3189 and Mrub_1304 shows that same consensus sequence, this provides further evidence suggesting that these two genes are orthologs.

Panel A



Panel B

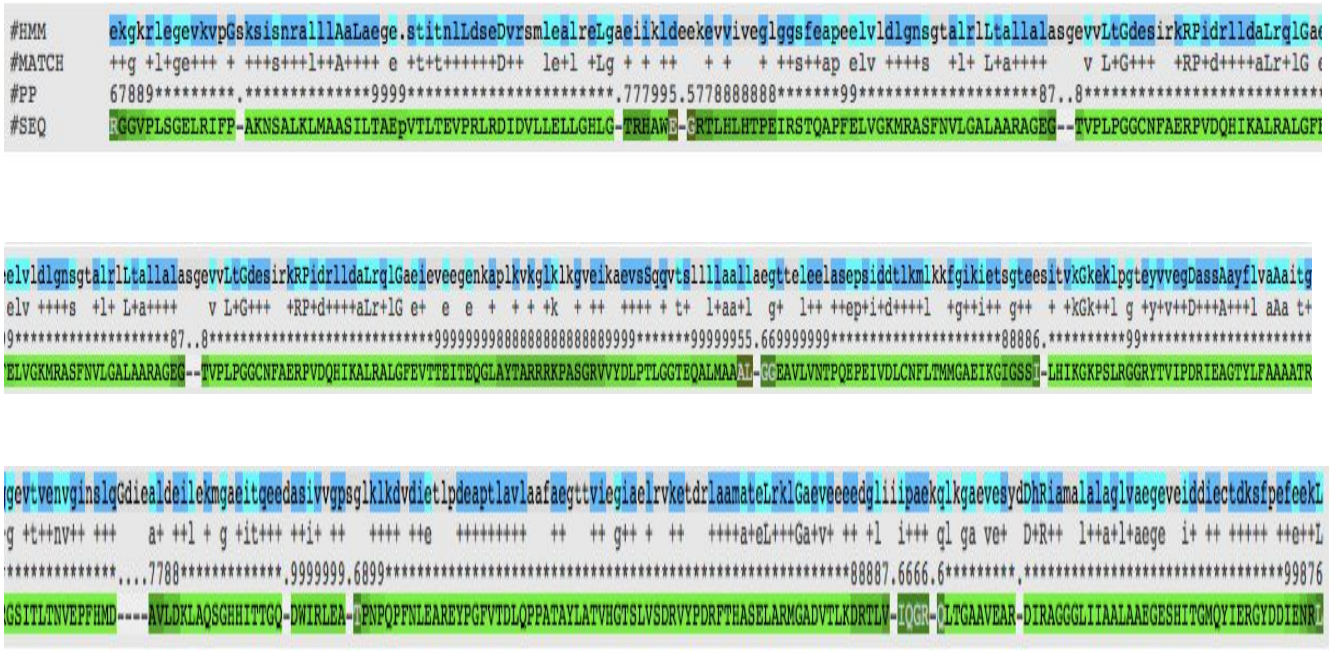


Figure 13. *E. coli* b_3189 and Mrub_1304 have many of the same highly conserved amino acids, which codes for the same domain (EPSP synthase (3-phosphoshikimate 1-carboxyvinyltransferase)). Panel A shows the pairwise alignment for *E. coli* b_3189. Panel B shows the pairwise alignment for *Mrub_1304.* The red box in panel A shows the amino acid, aspartate, which will be discussed in the process of site-directed mutagenesis within the conclusion. Both pairwise alignments were created using the Pfam website. Available from: https://pfam.sanger.ac.uk/search

Focusing on the pairwise alignments in figure 14, it is clear that *E. coli* b_3972 and Mrub_2007 have the same highly conserved glycine residues toward the beginning and end of the protein sequence in the FAD binding domain at positions 71 and 133. Also, these two genes have the same highly conserved glycine, phenylalanine, and valine residues toward the beginning and middle of the protein sequence in the UDP-N-acetylenolpyruvoylglucosamine reductase, C-terminal domain at positions 26, 29 and 91, respectively. As opposed to comparing the two protein sequences against each other, like performed in BLAST, these pairwise alignments compare our protein sequence to a consensus sequence, comprised of multiple other proteins (Madden 2002). Because this data from *E. coli* b_3972 and Mrub_2007 shows that same consensus sequences, this provides further evidence suggesting that these two genes are orthologs.

Panel A

FAD binding domain

UDP-N-acetylenolpyruvoylglucosamine reductase,C-terminal domain

Panel B

FAD binding domain

UDP-N-acetylenolpyruvoylglucosamine reductase,C-terminal domain

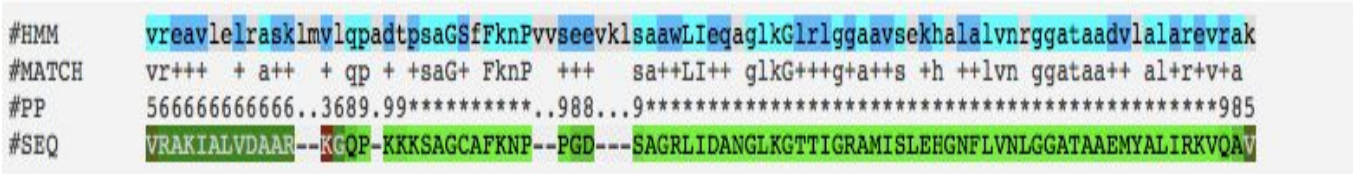


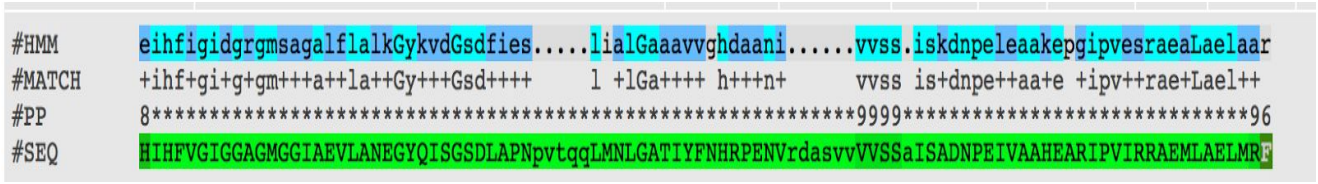
Figure 14. *E. coli* b_3972 and Mrub_2007 have some of the same highly conserved amino acids, which codes for the same two domains (FAD binding domain and UDP-N-acetylenolpyruvoylglucosamine reductase, C-terminal domain). Panel A shows the two

pairwise alignments for *E. coli* b_3972. Panel B shows the two pairwise alignment for Mrub_2007. Both pairwise alignments were created using the Pfam website. Available from: https://pfam.sanger.ac.uk/search

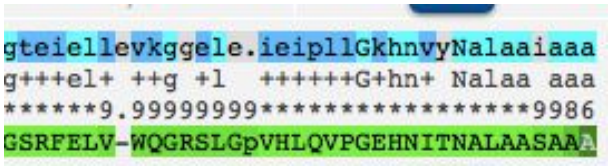
Looking at the pairwise alignments in figure 15, it is clear that *E. coli* b_0091 and Mrub_2006 have the same highly conserved glycine residues toward the beginning of the protein sequence in the Mur ligase family, catalytic domain at positions 23 and 28, respectively. Also, these two genes have the same highly conserved glycine, glutamate, and histidine residues toward the beginning and middle of the protein sequence in the Mur ligase middle domain at positions 6, 55 and 84, respectively. As opposed to comparing the two protein sequences against each other, like performed in BLAST, these pairwise alignments compare our protein sequence to a consensus sequence, comprised of multiple other proteins (Madden 2002). Because this data from *E. coli* b_0091 and Mrub_2006 shows that same consensus sequence, this provides further evidence suggesting that these two genes are orthologs.

Panel A

Mur ligase family, catalytic domain

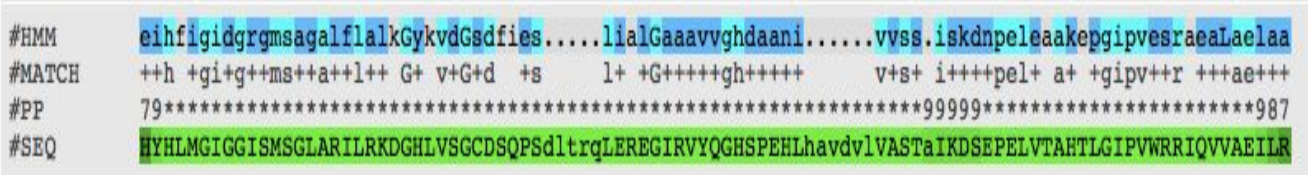


Mur ligase middle domain



Panel B

Mur ligase family, catalytic domain



Mur ligase middle domain

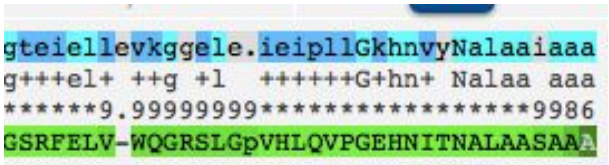
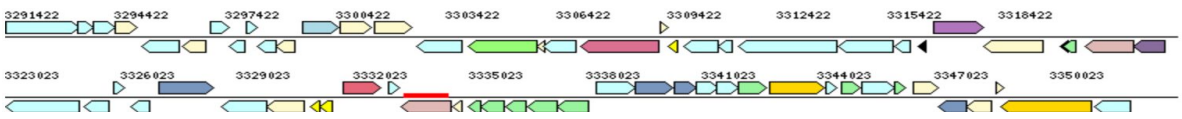


Figure 15. *E. coli* b_0091 and Mrub_2006 have some of the same highly conserved amino acids, which codes for the same two domains (Mur ligase family, catalytic domain, Mur ligase middle domain). Panel A shows the two pairwise alignment for *E. coli* b_0091. Panel B shows the two pairwise alignment for Mrub_2006. Both pairwise alignments were created using the Pfam website. Available from: https://pfam.sanger.ac.uk/search

Figure 16 shows the color of the gene of interest, which indicates a specific function. Both *E. coli* b_3189 and Mrub_1304, have the same color in the chromosome viewer, which shows that these genes have the same function. These genes are a pale pink color, which is representative of amino acid metabolism (Markowitz et al. 2012). Amino acid metabolism is a part of peptidoglycan biosynthesis in the fact that peptidoglycan is made up of amino acids that need to metabolized (Gautam *et al.* 2011). These genes of interest are not flanked by other genes with the same color as *E. coli* b_3189 and Mrub_1304 suggest these genes are not likely a part of an operon.

Gene Content A



Gene Content B

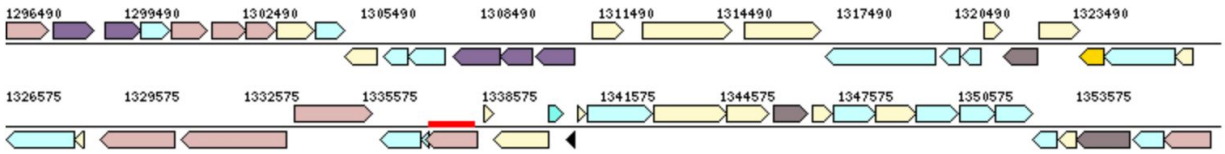
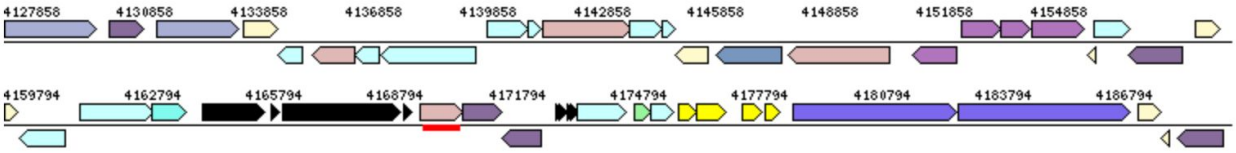


Figure 16. *E. coli* b_3189 and Mrub_1304 are not a part of an operon. Chromosome viewer was colored by KEGG. Panel A is *E. coli* b_3189 gene chromosome viewer. Panel B is Mrub_1304 gene chromosome viewer. IMG was used to obtain these images (Markowitz et al. 2012). Available from: https://img.jgi.doe.gov/cgi-bin/edu/main.cgi

Figure 17 shows the color of the gene of interest, which indicates a specific function. Both *E. coli* b_3972 and Mrub_2007, have the same color in the chromosome viewer, which shows that these genes have the same function. These genes are a pale pink color, which is representative of amino acid metabolism (Markowitz et al. 2012). Amino acid metabolism is a part of peptidoglycan biosynthesis in the fact that peptidoglycan is made up of amino acids that need to be metabolized (Gautam *et al.* 2011). These genes of interest are not flanked by other genes with the same color as *E. coli* b_3972 and Mrub_2007 suggest these genes are not likely a part of an operon.

Gene Context A



Gene Context B

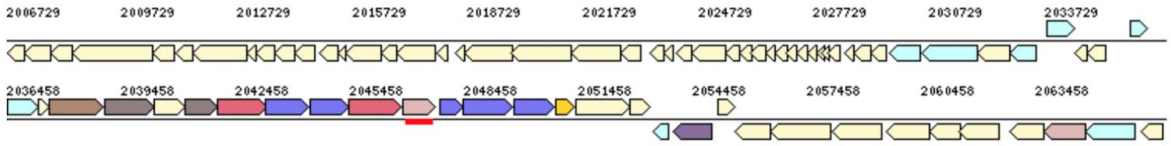
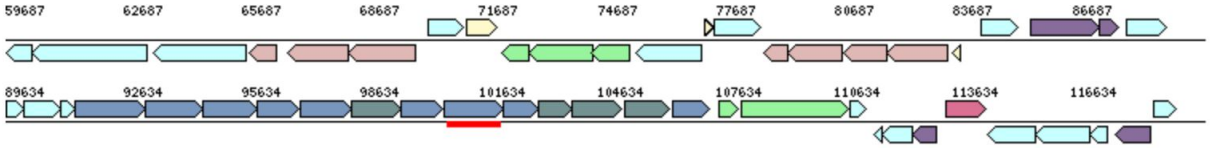


Figure 17. *E. coli* b_3972 and Mrub_2007 are not a part of an operon. Chromosome viewer was colored by KEGG. Panel A is *E. coli* b_3972 gene chromosome viewer. Panel B is Mrub_2007 gene chromosome viewer. IMG was used to obtain these images (Markowitz et al. 2012). Available from: https://img.jgi.doe.gov/cgi-bin/edu/main.cgi

Figure 18 shows the color of the gene of interest, which indicates a specific function. *E. coli* b_0091 and Mrub_2006, have the different colors in the chromosome viewer. Mrub_2006 gene has a dark pink color, which is representative of amino acid metabolism (Markowitz et al. 2012). *E. coli* b_0091 has a pale blue color, which is representative of metabolism of other amino acids (Markowitz *et al.* 2012). However, through further research in IMG, this difference in color is just indicator *E. coli* b_0091 is also a part of D-Glutamine and D-glutamate metabolism as well as peptidoglycan biosynthesis (Markowitz *et al.* 2012). Furthermore, these genes of interest are not flanked by other genes with the same color as *E. coli* b_0091 and Mrub_2006 which suggest these genes are not likely a part of an operon.

Gene Context A

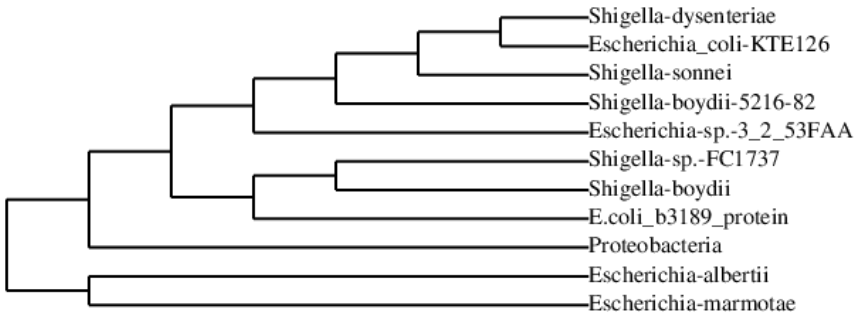


Gene Context B

Figure 18. *E. coli b_0091* and *Mrub_2006* are not a part of an operon. Chromosome viewer was colored by KEGG. Panel A is *E. coli* b_0091 gene chromosome viewer. Panel B is Mrub_2006 gene chromosome viewer. IMG was used to obtain these images (Markowitz et al. 2012). Available from: https://img.jgi.doe.gov/cgi-bin/edu/main.cgi

Figure 19 depicts the phylogenetic tree of *E. coli* and *M. ruber*. This data tells us that there was likely no horizontal gene transfer of the genes in *E. coli* b_3189 because the different species are within the same phylum. However, Mrub_1304 likely underwent horizontal gene transfer because the different species are not within the same phylum.

Panel A



Panel B

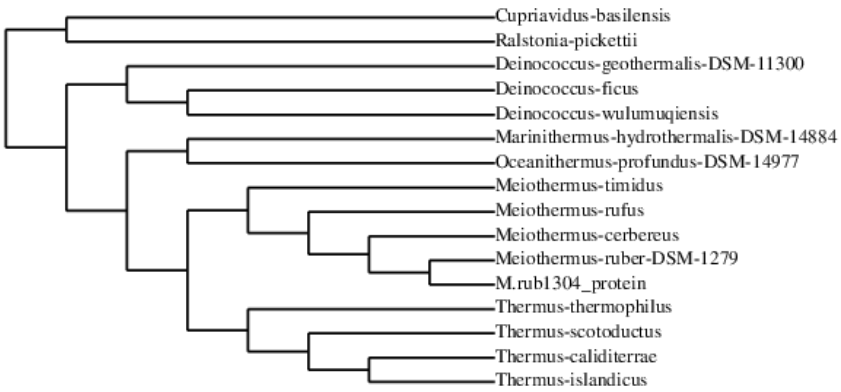
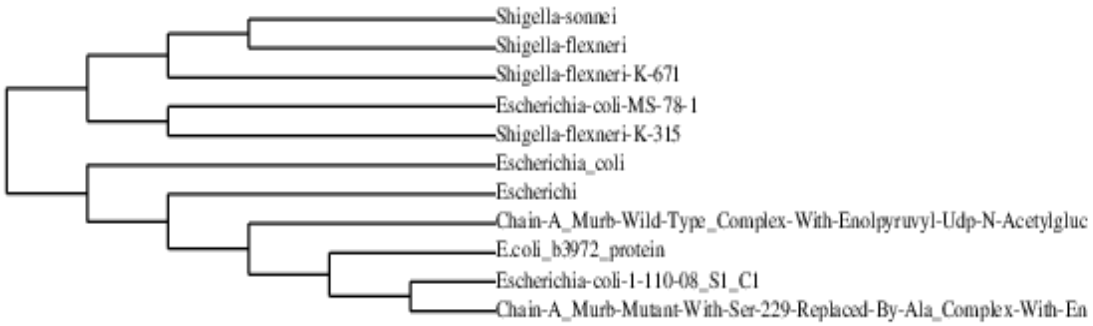


Figure 19. *E.coli* b_3189 and Mrub_1304 have not likely undergone horizontal gene transfer (HGT). Panel A is the phylogenetic tree for *E.coli* b_3189. Panel B is the phylogenetic tree for Mrub_1304. Available from: http://www.phylogeny.fr

Figure 20 looks at the the phylogenetic tree of *E. coli* and *M. ruber*. This data tells us that there was likely no horizontal gene transfer of the genes in *E. coli* b_3972 because the different species are within the same phylum. However, Mrub_2007 likely underwent horizontal gene transfer because the different species are not within the same phylum.

Panel A



Panel B

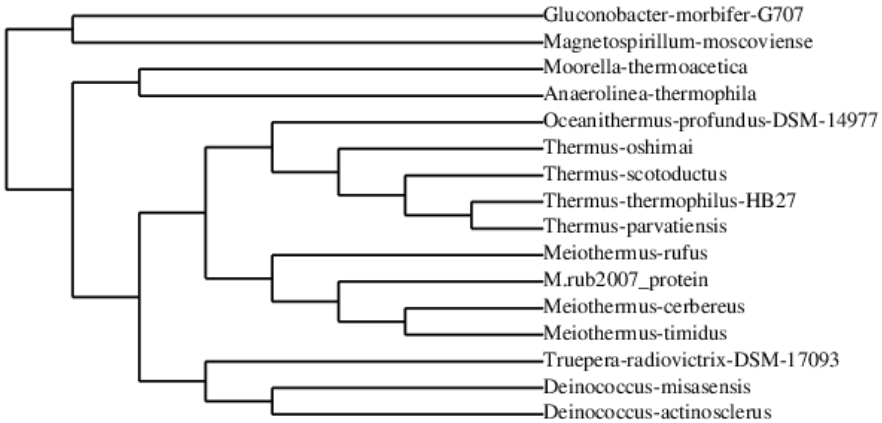
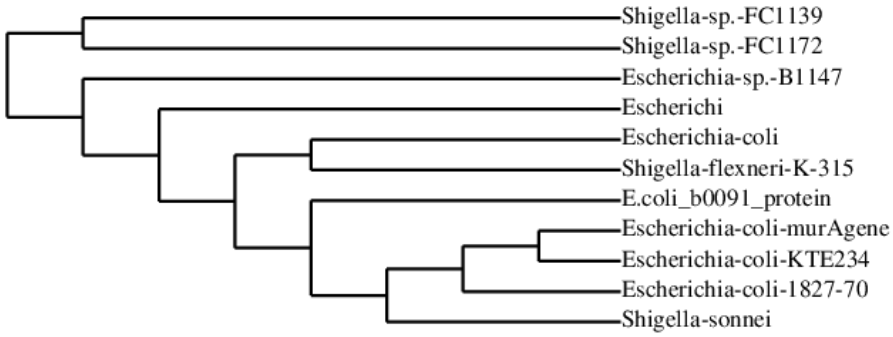


Figure 20. *E. coli* b_3972 and Mrub_2007 have not likely undergone horizontal gene transfer (HGT). Panel A is the phylogenetic tree for *E. coli* b_3972. Panel B is the phylogenetic tree for Mrub_2007. Available from: http://www.phylogeny.fr

Figure 21 shows the phylogenetic tree of *E. coli* and *M. ruber*. This data tells us that there was likely no horizontal gene transfer of the genes in *E. coli* b_0091 because the different species are within the same phylum. However, Mrub_2006 likely underwent horizontal gene transfer because the different species are not within the same phylum.

Panel A



Panel B

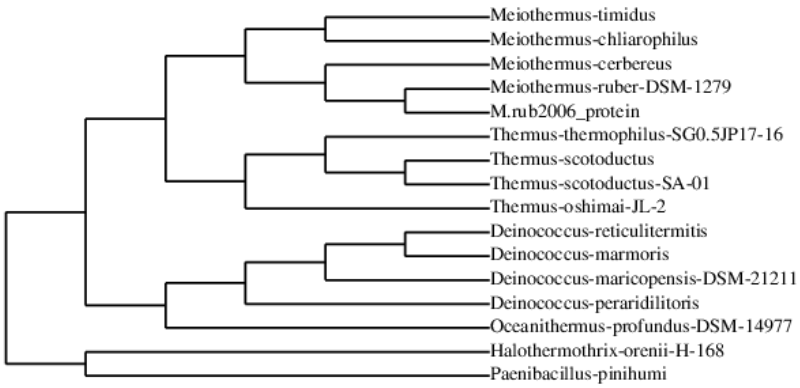


Figure 21. *E. coli* b_0091 and Mrub_2006 have not likely undergone horizontal gene transfer (HGT). Panel A is the phylogenetic tree for *E. coli* b_0091. Panel B is the phylogenetic tree for Mrub_2006. Available from: http://www.phylogeny.fr

Figure 22 shows the process of creating a missense mutation and receiving the forward and reverse primers needed for site-directed mutagenesis. The gene nucleotide sequence used for this process was from Mrub_1304. The red base pairs within the image represent the forward primer. The blue base pairs represent the reverse primer. This process of site-directed mutagenesis would help us understand that if a highly conserved amino acid undergoes a missense mutation, it would result in loss of function.

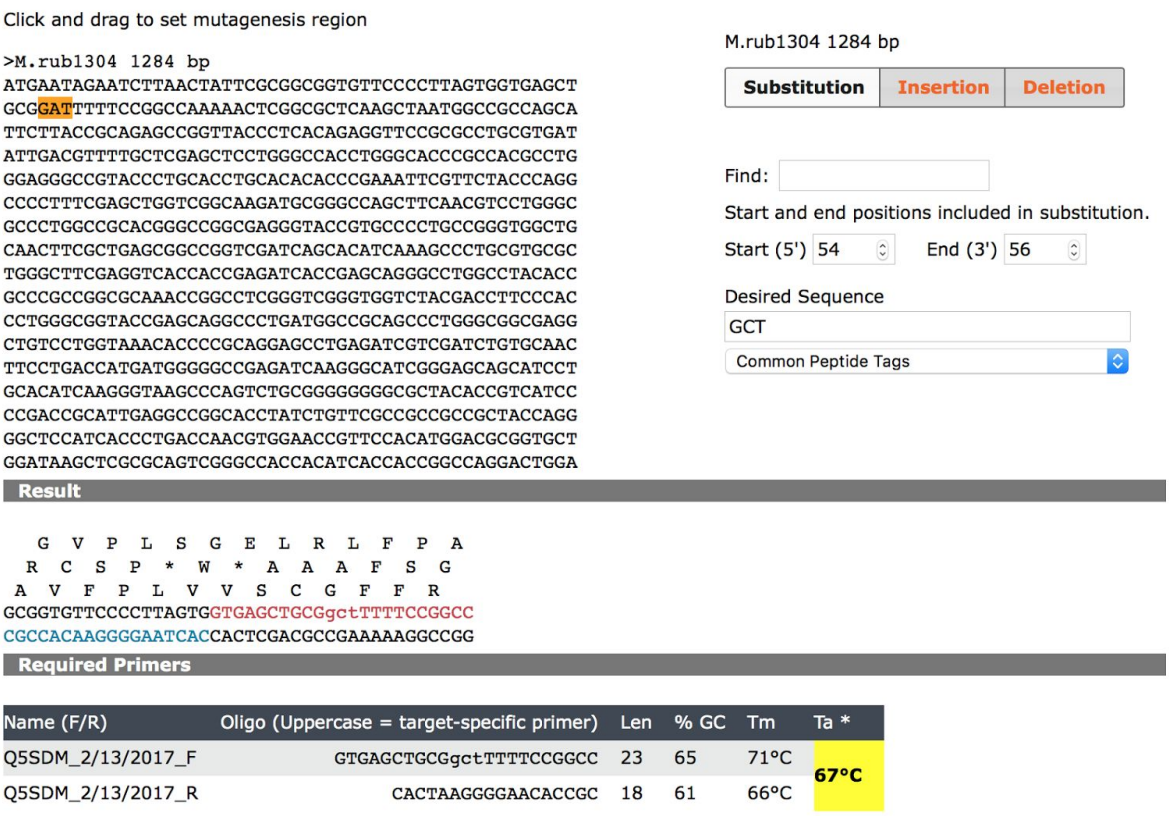


Figure 22. Creating a missense mutation in *M. ruber murA* via site-directed mutagenesis by substituting GCT for GAT at positions 54 and 56, which changes aspartate to alanine. Available from: http://nebasechanger.neb.com

Conclusion

The evidence obtained throughout this research brings us to the conclusion that Mrub_1304, Mrub_2007 and Mrub_2006 are orthologs of *E.coli* b_3189, b_3972, and b_0091, respectively, meaning that these two organisms likely have a common ancestor. These results support our hypothesis. The first piece of evidence was suggested by the BLAST output comparing the protein sequence of Mrub_1304, Mrub_2007 Mrub_2006 and *E.coli* b_3189, b_3972, b_0091, respectively (Madden 2002). Based on the very low E-values (2e-100, 5e-09, and 2e-81) and fairly high bit scores (294, 42.4, and 248), this showed us that strong sequence similarity is assumed to mean strong functional similarity. Additional bioinformatics tools that established the cellular location of the genes annotated in this study were TMHMM (Krogh and Rapacki 2016) , SignalP (Thomas *et al.* 2004), LipoP (Juncker *et al.* 2003), PSORT-B (Yu *et al.*

2010), and Phobius (Kall *et al.* 2004), indicating the location being in the cytoplasm. Also, Pfam (Finn *et al.* 2016) and TIGRfam (Haft *et al.* 2001) output data showed the protein sequence of Mrub_1304 and *E.coli* b_3189 to UDP-N-acetylglucosamine 1-carboxyvinyltransferase and its one domain. Mrub_2007 and *E. coli* b_3972 protein sequences matched to UDP-N-acetylenolpyruvoylglucosamine reductase and its two domains. Mrub_2006 and *E. coli* b_0091 protein sequences matched to UDP-N-acetylmuramate:L-alanine ligase and its two domains. The gene context images tell us that the genes are not a part of an operon. However this is not evidence of the predicted genes being orthologs, but does indicate gene function, which was similar in the between the genes predicted of being orthologs. Also, the phylogenetic tree generated and analyzed in this project displays the chances of horizontal gene transfer (HGT), which was unlikely because for each gene studied, the tree expressed that all the species with the genes are in the same phylum (Hornick *et al.* 2016). This applies to both *M.ruber* and *E. coli* genes of interest in this project. There were also many other bioinformatics programs used for this research yielding the same results each pair of predicted orthologs. There were two discrepancies within the data. Those discrepancies include *E. coli* b_0091 having one TMH present and the difference in color of Mrub_2006 and *E.coli* b_0091 on the chromosome viewer colored by KEGG. The first discrepancy is be explained by *E.coli* b_0091 being included in another pathway other than peptidoglycan biosynthesis pathway, which is D-Glutamine and D-glutamate metabolism. This also explains the difference color in the chromosome viewer compared to Mrub_2006. The second discrepancy is explained by the Ecocyc page for this enzyme, b0091, being identified as cytosolic, which means there is experimental evidence to support this location (Keseler *et al.* 2013). Secondly, the TMHMM output for b0091 indicates 0 TMH (as seen in Figure 9, Panel A). Thirdly, the HELP instructions for the TMHMM site explains that 18 or more amino acids must be found in N-terminal hydrophobic region to be a potential TMH; b0091 has only 12 aa (as seen in Figure 9, Panel A). Aside from these minor discrepancies, I believe the evidence overwhelming supports my hypothesis that Mrub_1304, Mrub_2007 and Mrub_2006 are orthologs of *E.coli* b_3189, b_3972, and b_0091, respectively.

Future research could include site-directed mutagenesis. If I were to study one of the genes from my project by site-directed mutagenesis, I would choose the gene Mrub_1304. Using the Pfam pairwise alignment (in figure 13), I would find the highly conserved amino acid within both Mrub_1304 and its ortholog *E.coli* b_3189 in order to select an amino acid. After finding the highly conserved amino acid for Mrub_1304, which is an aspartate residue, I would alter the codon for aspartate in nucleotide sequence. It is likely that the deletion or substitution of aspartate would cause loss of function because of how highly conserved the amino acid is and because it is a negatively charged amino acid. According to Betts and Russell (2003), aspartate plays an important role in the protein active and binding sites function. Therefore, through substitution of alanine with aspartate or the deletion of aspartate, there would more than likely be loss of function within this protein. However, I would chose to do a substitution mutation,

substituting the codon for aspartate with a codon for alanine. I chose alanine because it is known to be fairly non-reactive and not really involved in protein function (Betts and Russell, 2003). To undergo this mutation, using NEBaseChanger, the primers that would need to used are GTGAGCTGCGgctTTTTCCGGCC as the forward primer and CGCCACAAGGGGAATCAC as the reverse primer, refer to Figure 22 (NEBaseChanger, 2017). As explained above this could be a direction for future study on the genes discussed in my project.

Work Cited

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/.

Betts MJ and Russell RB. (2003). Amino Acid Properties and Consequences of Substitutions. *Bioinformatics for Geneticists.* [2017 Jan 10]. Available from: http://moodle.augustana.edu/pluginfile.php/294318/mod_resource/content/1/bettsRussell2003.pdf

Blount ZD. (2015). The unexhausted potential of *E. coli.* eLife;4: 1-12. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4373459/pdf/elife05826.pdf

Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, *Genome Research*, 14:1188-1190, 2004; [2016 Dec 6]. Available at: http://weblogo.berkeley.edu/

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/

Gautam A, Vyas R & Tewari R. (2011). Peptidoglycan biosynthesis machinery: A rich source of drug targets. Critical Reviews in Biotechnology; 31(4): 295–336. Available from: http://web.a.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=3&sid=f0a100be-6581-448e-9fc8-203d7e38a6d5%40sessionmgr4006&hid=4204

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.

Hornick L, Mosoiu A, Howe K, Hoekman R. [Internet]. (2016). Joint Genome Institute (JGI) United States Department of Energy. [2016 Jan 10]. Available from: http://jgi.doe.gov/about-us/

Kall L, Krough A, Sonnhammer E. 2004. A combined transmembrane topology and signal peptide prediction method. Journal of Molecular Biology 338(5):1027-36.

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/

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. MetaCyc: fusing model organism databases with systems biology *Nucleic Acids Research* 41:D605-612.

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/

Loginova LG, Egorova LA. (1975). Obligate thermophilic- bacterium *Thermus ruber* in hot springs of Kamchatka. Mikrobiologiya; 44:661-665.

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

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

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

NEBaseChanger [Internet]. (2017). New England Biolabs Inc. [2017 Jan 10]. Available from: http://nebasechanger.neb.com

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology 302 (1):205-17 Available from: http://www.ebi.ac.uk/Tools/msa/tcoffee/

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**

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

Scott LR. Meiothermus ruber Genome Analysis Project. [Internet]. GENI-ACT;[2017 Jan 31]. Available from: https://geni-science.org/secure/projects/view/

Thomas Nordahl Petersen, Søren Brunak, Gunnar von Heijne & Henrik Nielsen Discriminating signal peptides from transmembrane regions. 2011. *Nature Methods*, 8:785-786. Available from: http://www.cbs.dtu.dk/services/SignalP

Tindall et al. (2010). Complete genome sequence of Meiothermus ruber type strain. Stand Genomic Sci; 3(1): 26-36.