2017

Serine Biosynthesis and Glycine Biosynthesis/ Degradation: Mrub_0173 is Orthologous to *E. coli* b2913 (serA); Mrub_0125 is Orthologous to *E. coli* b4388 (serB); Mrub_2910 is Orthologous to *E. coli* b2551 (glyA).

Megan M. Janssen
Augustana College, Rock Island Illinois

Dr. Lori R. Scott
Augustana College, Rock Island Illinois

Serine Biosynthesis and Glycine Biosynthesis/Degradation: *Mrub_0173* is Orthologous to *E. coli* b2913 (*serA*); *Mrub_0125* is Orthologous to E. coli b4388 (*serB*); *Mrub_2910* is Orthologous to *E. coli* b2551 (*glyA*);

Megan M. Janssen
Dr. Lori R. Scott Laboratory
Biology Department, Augustana College
639 38th Street, Rock Island, IL 61201

INTRODUCTION

Why Study Meiothermus Ruber?

Meiothermus ruber (*M. ruber*) is a red-pigmented, thermophilic bacterium that is found in the Deinococcus-Thermus phylum (Tindall *et al.*, 2010). It is a gram negative, non-spore forming, rod-shaped, non-motile bacterium. The bacterium prefers to grow in high-temperature environments that range from 35-70°C. While organisms such as *Escherichia coli* and *Staphylococcus aureus* have over 40,000 publications available on PubMed, *M. ruber* has only about 38 publications (Scott, 2016). As you can see, there is a huge gap in information about the genes within *M. ruber*'s genome and their functions. Dr. Scott chose to study this lesser-known organism because it may lead to novel or variant processes that aren't found in well-studied bacteria. *M. ruber*'s genome has been sequenced by the Joint Genome Institute (JGI) as part of the Genome of Bacteria and Archaea Encyclopedia (GEBA) (Phylogenetic Diversity). Researching this microbe will advance the understanding of bacteria in general and may identify new biological pathways and processes.

E. coli as a control

Using a model organism is one way to fill in the gaps missing information for organisms that are understudied. *Escherichia coli* K12 MG1655, a common model organism, is relatively easy to grown in the laboratory and this allows it to be widely studied (Cooper 2000). *E.coli* has had its entire genome sequenced. All of its genes have been identified and many have been functionally confirmed (Keseler *et al.*, 2016). Aside from EcoCyc, there are many other online databases that also study *E. coli* K12 MG1655 and can be found on the EcoCyc website (Keseler *et al.*, 2016). Before starting the bioinformatics annotations, BLAST searches were conducted comparing *Mrub_0173* to *b2913*, *Mrub_2910* to *b2551*, and *Mrub_0125* to *b4388*. The results showed that the sequences between each gene pair were similar. As a result, we use we use E. coli as our control, not because is it easy to grow and extensively studied, but it also because it contains the genes that may be orthologous the *M. ruber* genes we are interested in.

Serine Biosynthesis

Serine is a non-essential amino acid that is functionally present in many proteins. Its chemical formula is $C_3H_7NO_3$. Serine is found in the cytoplasm, mitochondria, peroxisome, and can be extracellular. Furthermore, it is in urine, saliva, cerebrospinal fluid, blood, and all tissues (Metabocard for L-Serine). It is needed for cell membranes, the metabolism of fats and fatty acids, and muscle growth. Serine has a big role in purine, pyrimidine, and creatine pathways, as well as cellular multiplication in general (Tom *et al.*, 2003). The amino acid can be attained through dietary intake, protein and phospholipid degradation, biosynthesis from 3-phospho-glycerate, and from glycine. Different tissues during various stages of human development get serine through one of these ways.

Figure 1 shows the steps in the biosynthesis of serine. It begins by taking 3-phospho-D-glycerate and changing it into 3-phospho-hydroxopyruvate by D-3-phosphoglycerate dehydrogenase (*serA*). The next step transforms 3-phospho-hydroxopyruvate into 3-phsopho-L-serine through 3-phosphoserine aminotransferase (*serC*). The last step takes 3-phospho-L-serine and changes it into L-serine by phosphoserine phosphatase (*serB*). In this paper we are specifically looking at *serA* and *serB* in this pathway. *Mrub_0173* is a predicted ortholog to *E. coli b2913* (*serA*) and *Mrub_0125* is a suspected to be orthologous to *E.coli b4388* (*serB*).

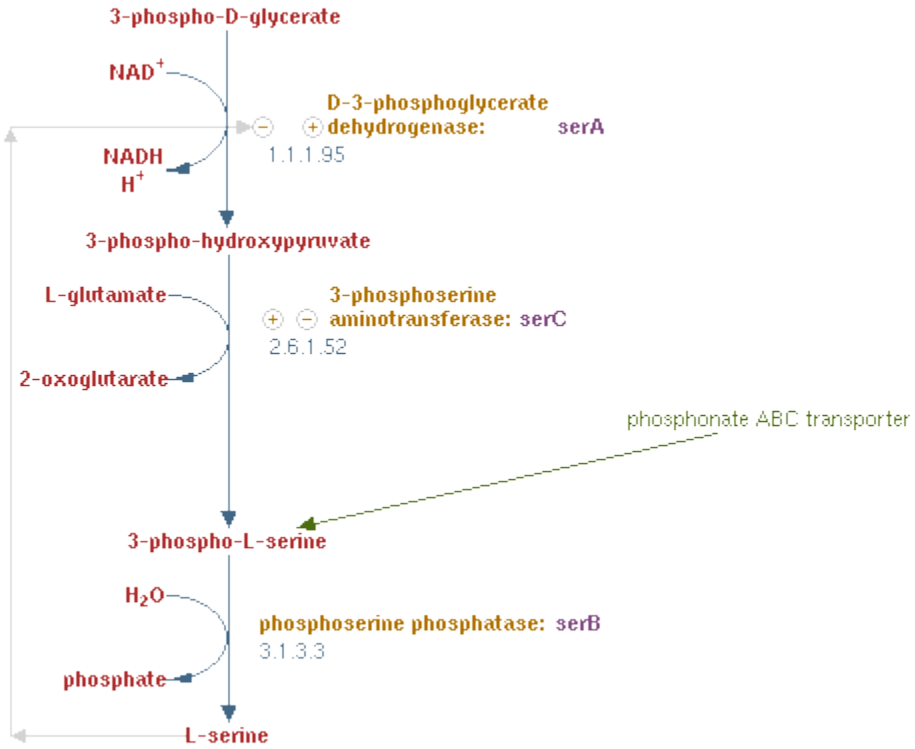


Figure 1. L-Serine biosynthesis pathway showing reactants, products, and the enzymes/genes that are involved with the catalysis of each reaction. Image from: https://ecocyc.org/ECOLI/NEW-IMAGE?type=PATHWAY&object=SERSYN-PWY

Glycine Biosynthesis and Degradation

Glycine is also a non-essential amino acid and it too has a big role in cellular growth. Its chemical formula is $C_2H_5NO_2$ and it is found in the mitochondria, lysosome, peroxisome, and can be extracellular. Furthermore, it is located in urine, saliva, cerebrospinal fluid, blood, and various tissues just like Serine (Metabocard for Glycine). Glycine takes part in production of DNA, phospholipids, and collagen. Studies have shown that growth in human cells were significantly better when the "non-essential" amino acids serine and glycine were added to cultures (Tom *et al.,* 2003). Glycine degradation is might be important to *M. ruber* because glycine percentages in proteins correspond to that protein's stability. A glycine substitution into a protein causes an increase in distance between the amino group and the alpha carbon and glycine removal reduces the hinge motion of the protein (Jacob *et al.*, 1999). So, a higher percentage of glycine in a protein increases the overall flexibility of that protein, while a lower percentage increases rigidity in that protein. Because *M. ruber* is a thermophilic bacterium that can grow at up to 70 ^{O}C, its proteins have to be more stable in order to function and preserve its structure at the high temperatures. So, glycine might be important when studying *M. ruber*.

Figure 2 illustrates the single step of glycine to serine through *glyA*. Starting with glycine, 5,10-methlyenetetrahydrofolate: glycine hydroxymethyltransferase (*glyA*), also called serine hydroxymethyltransferase, catalyzes the reaction to form L-serine. The reaction is reversible by using the same protein, GlyA. In this paper, *Mrub_2910* is suspected to be orthologous to *E.coli b2551* (*glyA*).

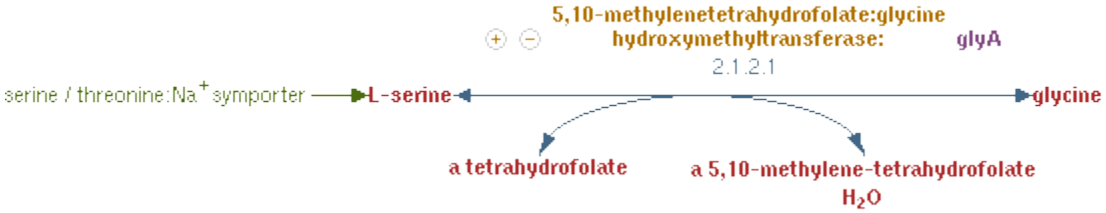


Figure 2. Glycine biosynthesis pathway showing reactants, products, and the enzyme/gene that are involved with the reaction. Image from: https://ecocyc.org/ECOLI/NEW-IMAGE?type=PATHWAY&object=GLYSYN-PWY

Bioinformatics

To people with careers in biological sciences, understanding and utilizing various bioinformatics tools are important pieces of knowledge. If they are free resources, these bioinformatics tools can save a lot of time if the user knows how to use and interpret the information. More and more scientific data will be stored in various databases, similar to the one used for this project, as bioinformatics technologies advance (Persidis, 1999). For that reason, a comprehensive understanding of available bioinformatics tools is, and will be, critical for succeeding in the field of biology.

Purpose/Hypothesis

This project consists of utilizing various bioinformatics tools to decide if the three *Meiothermus ruber* genes: *Mrub_0173, Mrub_2910, Mrub_0125* are orthologs to the respective *Escherichia coli* genes: *b2913, b2551, b4388*. We use the bioinformatics programs so we can find the similarities and differences between the genes and proteins encoded by them. For example, E-values are a common output for many bioinformatics tools. By using E-values, a value that correlates with significance, we can interpret the likelihood that two nucleotide or amino acid sequences have the same structure, and therefore similar function. High E-values mean that the sequences probably lined up due to chance alone. Low E-values mean that the sequence alignment was significant and random similarities are unlikely (NCBI). A BLASTp comparison between *Mrub_0173* and *E.coli b2913* yielded an E-value of 4e-41. The BLASTp of *Mrub_0125* and *E.coli b4388* produced an E-value of 8e-06. The BLASTp of *Mrub_2910* and *E.coli b2551* generated an E-value of 1e-154. Based on the notably low E-value given by the initial BLASTp comparison between each pair of genes, we hypothesize that the each pair of genes are orthologous to one another.

Methods

The GENI-ACT gene annotation website directions were followed with only a few changes in order to collect data on the six genes (http://www.geni-act.org/education/main/). Before starting the bioinformatics annotations, A BLASTp was performed to compare *Mrub_0173* to *b2913*, *Mrub_2910* to *b2551*, and *Mrub_0125* to *b4388* to determine similarities between the corresponding sequences. After concluding that each pair was similar to one another through the BLASTp searches, we started to annotate each gene by using the various bioinformatics programs to complete the associated 9 modules on GENI-ACT. To start, we filled in the first module by going to GENI-ACT's gene page for each gene and filling in the appropriate sequences and coordinates. Next, we confirm the start codon for each gene by using the IMG/M website (Markowitz *et al.,* 2012). Then, we fill in the second module by doing a BLAST with our amino acid sequence and filling in the CDD domain hits (NCBI; Marchler-Bauer *et al.*). We then choose 15 sequences from the BLAST and plugging them into T-Coffee to get a multiple sequence alignment (Notredame *et al.,* 2000). To finish the second module, we use WebLogo for the sequence logo (Haft *et al.,* 2001). Then we skip down to module 7, duplication and degradation, and fill it out using KEGG maps (Kanehisa *et al.,* 2016). Next, we go back up to the 3rd module, cellular localization data, and use NCBI, TMHMM, SignalP, LipoP, and PSORT-B to determine the protein's location in the cell (NCBI; Krogh *et al.,* 2016; Petersen *et al.,* 2011; Junker *et al.,* 2003; Yu *et al.,* 2010). The structure-based evidence module was filled out using TIGRfam, Pfam, and PDB (Haft *et al.,* 2001; Finn *et al.,* 2016; Berman *et al.,* 2000). The next module, enzymatic function, was filled out using KEGG and MetaCyc. (Kanehisa *et al.,* 2016; Keseler *et al.,* 2013). The last module, horizontal gene transfer, was completed with Phylogeny.fr and IMG/M (Dereeper *et al.,* 2008; Markowitz *et al.,* 2012). After completing all of the modules, we determined the correct annotation for each gene.

One change we had to the GENIACT instructions was that we used the top 15 BLAST hits for the T-coffee program instead of using the recommended top 10 BLAST. The BLAST search for these hits excluded the gene's bacteria species. Another change was that we omitted the Open Reading Frame module and the Paralog module for all of the *E. coli* genes. The bioinformatics tools utilized are cited in the works cited at the end.

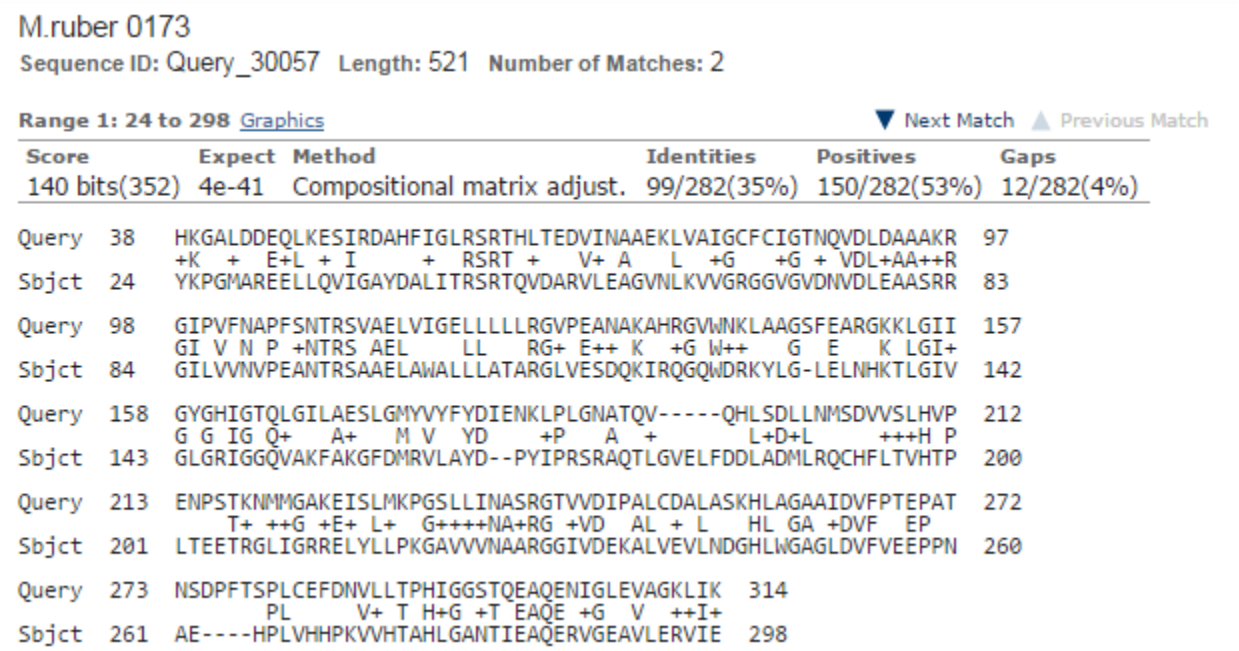
RESULTS

Table 1 is a summary of the results from the bioinformatics tools comparing *E. coli b2913*(*serA*) to *Mrub_0173*. The initial BLASTp result that we mentioned in the analysis is presented in the first row. The two amino acid sequences varied in length, so the low bit score is understandable. The E-value of the BLAST is 4e-41, which is close to zero. Because of this, we can assume that these two sequences share many of the same amino acids, suggesting functional similarities, and do not align simply by chance. A search of the CDD database pulled the same COG number (COG0111), SerA, for the two proteins. They both have significantly small E-values, which indicate that they could be the same enzyme in the serine biosynthesis pathway. The cellular location bioinformatics tools, a combination of TMH, SignalP, LipoP, and PSORT-B, suggest that both of the proteins are localized to the cytoplasm. The lack of a cleavage site indicates they are not membrane-bound nor traverse a membrane. This similarity in the location of the enzyme coded by the two genes is even more evidence that they are orthologs. Furthermore, the TIGRfam tool pulled the same hit from the database of TIGR01327 named PGDH: phosphoglycerate dehydrogenase and they both had extremely small E-values indicating sequence similaritiy. In addition, a search of the Pfam database indicated that both proteins have the same two domains, 2-Hacid_dh (PF00389) and 2-Hacid_dh_C (PF02826). The protein database (PDB) pulled two different names and numbers for each sequence, but they end up being the same protein in different organisms. *E. coli b2913*(*serA*) and *Mrub_0173* both had the same enzyme commission number, E.C.1.1.1.95 and were both predicted to be involved in the same step of serine biosynthesis, a sub-pathway of methane metabolism.

Table 1: *E. coli b2913*(*serA*) is orthologous to *Mrub_0173*

Bioinformatics tool used	***E. Coli b2913* gene (*serA*)**	***M. ruber Mrub_0173* gene**
BLAST *E.coli* against *M.ruber*	Score: 140 bits E-value: 4e-41	
CDD Data (COG category)	COG Number: COG0111 SerA	
	E-value: 2.12e-85	E-value: 1.73e-98
Cellular Localization	Cytoplasm of the cell	
TIGRfam – protein family	TIGR01327 PGDH: phosphoglycerate dehydrogenase	
	E-value: 5e-34	E-value: 2.7e-181
Pfam – protein family	1) PF00389 (2-Hacid_dh) 2) PF02826 (2-Hacid_dh_C)	
	E-values: 1) 1.1e-38 2) 3.1e-50	E-values: 1) 1.1e-32 2) 1e-58
Protein Database	1PSD – The allosteric Ligand site in the vmax-type cooperative enzyme phosphoglycerate dehydrogenase	3DDN – Crystal structure of hydroxypyruvic acid phosphate bound D-3-phosphoglycerate dehydrogenase in mycobacterium tuberculosis
	E-value: 0.0	E-value: 8.01702E-66
Enzyme commission number	E.C. 1.1.1.95 – Phosphoglycerate dehydrogenase	
KEGG pathway map	Pathway ID: 00680 Methane Metabolism	

The image in Figure 2 illustrates the results of the initial BLAST search of *E.coli serA* against *Mrub_0173*. The figure shows that 35% of the amino acids were exactly the same between the sequences and 150 amino acids were characteristically similar. With an E-value of 4e-41, which is close to zero, we can conclude that these two sequences did not line up by chance but represent structural and therefore functional similarity. For that reason, we can start to see that *E.coli serA* and *Mrub_0173* might share some major structural similarities. This BLAST is our initial indication that the two genes might be orthologous to one another.


M.ruber 0173

Sequence ID: Query_30057 Length: 521 Number of Matches: 2

Range 1: 24 to 298 Graphics ▼ Next Match ▲ Previous Match

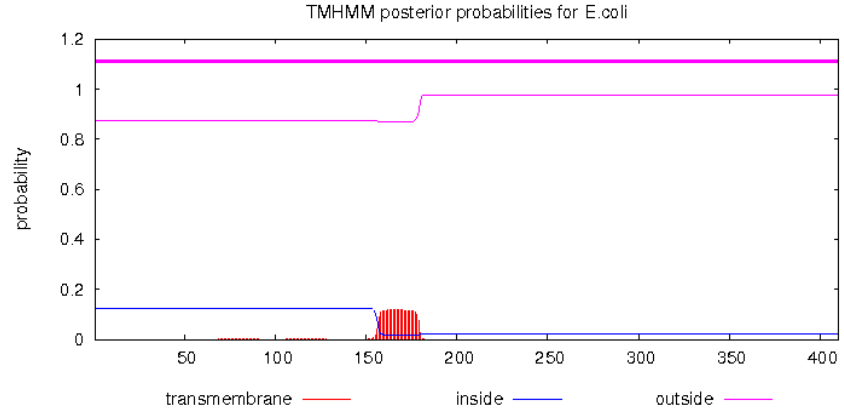
Score	Expect	Method	Identities	Positives	Gaps
140 bits(352)	4e-41	Compositional matrix adjust.	99/282(35%)	150/282(53%)	12/282(4%)

```
Query  38   HKGALDDEQLKESIRDAHFIGLRSRTHLTEDVINAAEKLVAIGCFCIGTNQVDLDAAAKR  97
            +K  +  E+L + I     +  RSRT +   V+ A   L  +G   +G + VDL+AA++R
Sbjct  24   YKPGMAREELLQVIGAYDALITRSRTQVDARVLEAGVNLKVVGRGGVGVDNVDLEAASRR  83

Query  98   GIPVFNAPFSNTRSVAELVIGELLLLLRGVPEANAKAHRGVWNKLAAGSFEARGKKLGII  157
            GI V N P +NTRS AEL    LL   RG+ E++ K  +G W++   G  E   K LGI+
Sbjct  84   GILVVNVPEANTRSAAELAWALLLATARGLVESDQKIRQGQWDRKYLG-LELNHKTLGIV  142

Query  158  GYGHIGTQLGILAESLGMYVYFYDIENKLPLGNATQV-----QHLSDLLNMSDVVSLHVP  212
            G G IG Q+   A+   M V  YD    +P   A  +       L+D+L     +++H P
Sbjct  143  GLGRIGGQVAKFAKGFDMRVLAYD--PYIPRSRAQTLGVELFDDLADMLRQCHFLTVHTP  200

Query  213  ENPSTKNMMGAKEISLMKPGSLLINASRGTVVDIPALCDALASKHLAGAAIDVFPTEPAT  272
                T+ ++G +E+ L+  G++++NA+RG +VD  AL + L   HL GA +DVF  EP
Sbjct  201  LTEETRGLIGRRELYLLPKGAVVVNAARGGIVDEKALVEVLNDGHLWGAGLDVFVEEPPN  260

Query  273  NSDPFTSPLCEFDNVLLTPHIGGSTQEAQENIGLEVAGKLIK  314
                   PL     V+ T H+G +T EAQE +G  V  ++I+
Sbjct  261  AE----HPLVHHPKVVHTAHLGANTIEAQERVGEAVLERVIE  298
```

Figure 2. *E.coli serA* and *Mrub_0173* have similar amino acid sequence. Query sequence: *E. coli serA;* Subject sequence: *Mrub_0173.* Analysis was performed using the NCBI BLAST bioinformatics tool at http://www.ncbi.nlm.nih.gov.

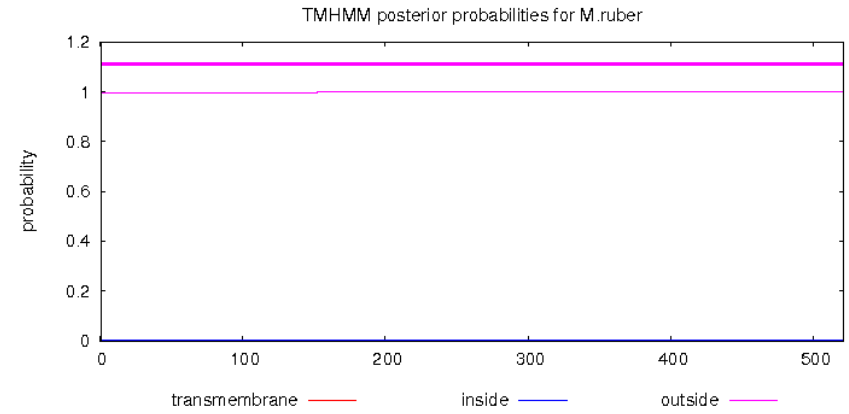
Figure 3 presents the TMH hydropathy plots for *E.coli serA* and *Mrub_0173*. Red peaks in a TMHMM plot that rise above a certain threshold represent the presence of transmembrane helices. Panel A shows a peak, but the height of the peak is not high enough to be significant. Subsequently, both of the TMHMM hydropathy charts show that the proteins coded by these genes are not in the membrane, but are in the cytoplasm.

```
# E.coli Length: 410
# E.coli Number of predicted TMHs:  0
# E.coli Exp number of AAs in TMHs: 2.67803
# E.coli Exp number, first 60 AAs:  0.0004
# E.coli Total prob of N-in:        0.12514
E.coli  TMHMM2.0        outside      1   410
```



Panel A

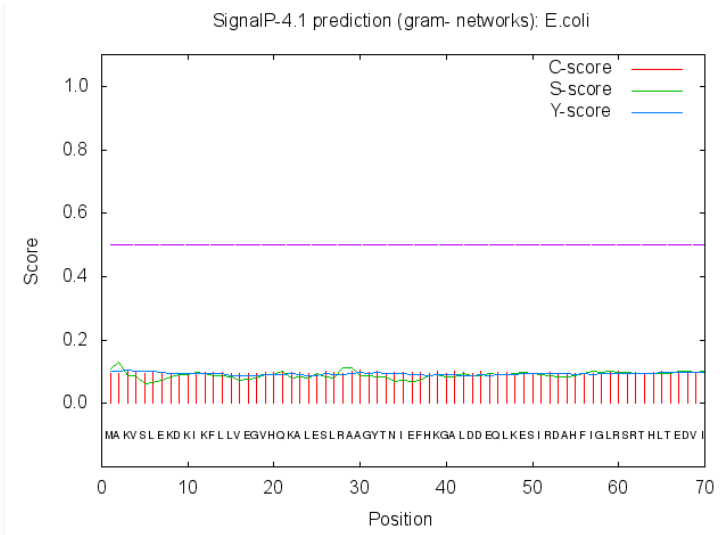
```
# M.ruber Length: 521
# M.ruber Number of predicted TMHs:  0
# M.ruber Exp number of AAs in TMHs: 0.01431
# M.ruber Exp number, first 60 AAs:  3e-05
# M.ruber Total prob of N-in:        0.00262
M.ruber TMHMM2.0        outside      1   521
```



Panel B

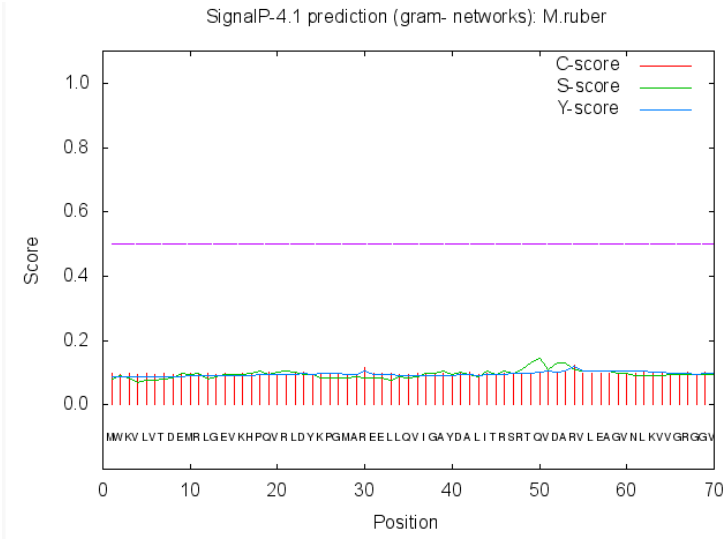
Figure 3. *E.coli serA* and *Mrub_0173* do not contain TMH regions; both predicted to be located in cytoplasm. Panel A shows the TMHMM for *E.coli b2913/serA;* Panel B shows the TMHMM for *Mrub_0173*. TMHMM Server v 2.0 http://www.cbs.dtu.dk/services/TMHMM was utilized to create the hydropathy charts.

The charts in Figure 4 are SignalP plots for *E.coli serA* and *Mrub_0173.* The SignalP bioinformatics tool is used in order to predict protein cleavage sites for proteins that might be bound to the cell membrane and/or pass through the membrane. It calculates a D-value by using the S-score and Y-score. If the D-value is lower than the cutoff value, represented by the purple line, then the protein does not have any cleavage sites. *E.coli serA* (Panel A) has a D-value of 0.094 that is not above the cutoff value of 0.570. *Mrub_0173* (Panel B) has a D-value of 0.107, which is also below the cutoff. That means both of the proteins do not contain cleavage sites, and therefore remain in the cytoplasm.



# Measure	Position	Value	Cutoff	signal peptide?
max. C	30	0.107		
max. Y	30	0.099		
max. S	2	0.131		
mean S	1-29	0.088		
D	1-29	0.094	0.570	NO

Panel A



# Measure	Position	Value	Cutoff	signal peptide?
max. C	54	0.124		
max. Y	54	0.118		
max. S	50	0.143		
mean S	1-53	0.095		
D	1-53	0.107	0.570	NO

Panel B

Figure 4. *E.coli serA* and *Mrub_0173* do not have cleavage sites; D values for both charts were below cutoff value. Panel A shows the plot for *E.coli b2913/serA*; Panel B shows the plot for *Mrub_0173*. Plots created by Signal P server v. 4.1 http://www.cbs.dtu.dk/services/SignalP.

Two more bioinformatics tools were used to determine cellular localization of the proteins. The LipoP tool predicted that both *E.coli serA* and *Mrub_0173* were located in the cytoplasm of the cell and did not have any cleavage sites. PSORT-B showed cytoplasmic score of 9.97 and

cytoplasmic membrane and periplasmic scores of 0.01 for *E.coli serA.* PSORT-B showed the exact same results for *Mrub_0173.* The final prediction from PSORT-B for both genes' cellular locations was in the cytoplasm. Because of this singular result for both, the 0.01 scores for the cytoplasmic membrane and periplasmic locations are insignificant. Since all of the cellular location bioinformatics tools indicate that the genes do not have any cleavage sites, we do not utilize the Phobius tool. All bioinformatics tools predict that both *E.coli serA* and *Mrub_0173* are in the cytoplasm (Table 1).

The pathway in Figure 5 shows the serine sub-pathway of methane metabolism. The green colored enzymes are allegedly to be present in the organism. We can see that both the *E.coli serA* and *Mrub_0173* are predicted to be both involved in the first step in the serine biosynthesis and code for the same enzyme, phosphoglycerate dehydrogenase. This is even more confirmation that the two genes are orthologous in nature.

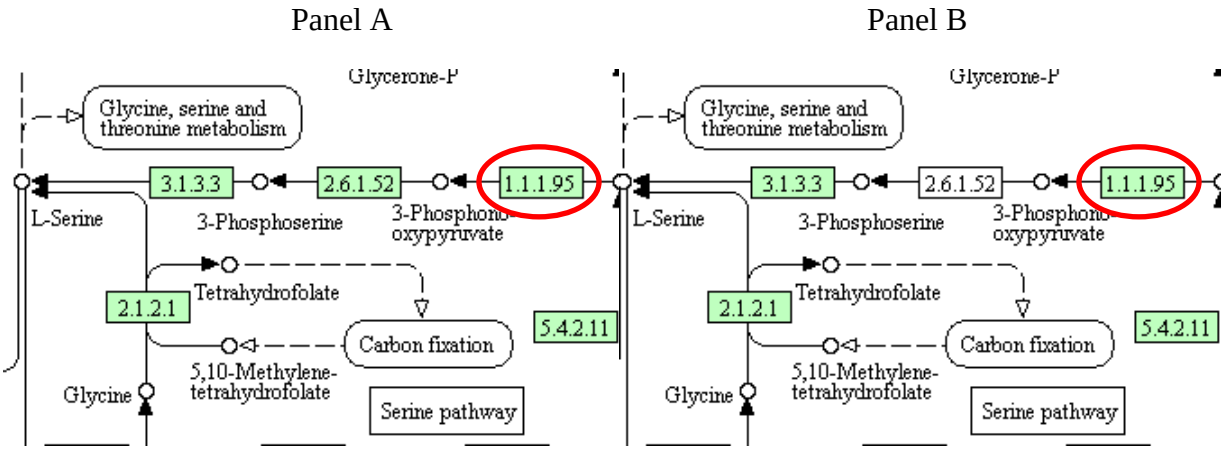
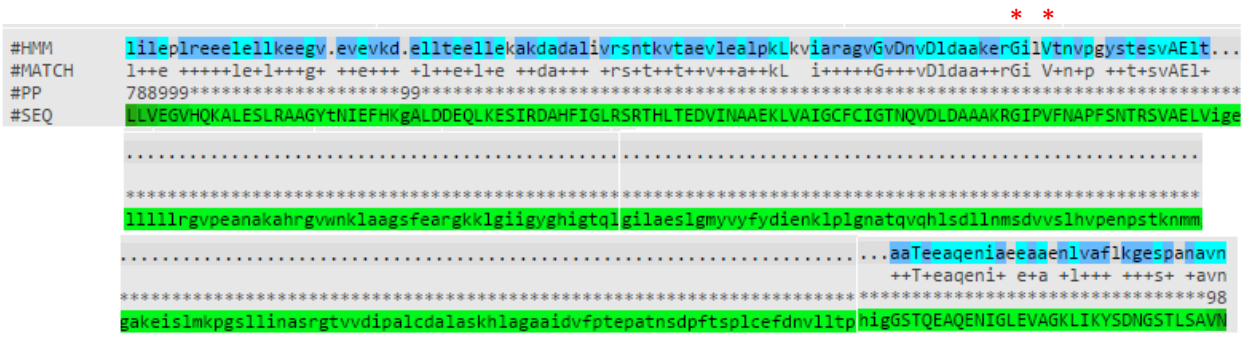


Figure 5. *E.coli serA* and *Mrub_0173* exist in the same biochemical pathway. Panel A shows the KEGG pathway when focusing on *Escherichia coli.* Panel B shows the KEGG pathway after selecting for *Meiothermus ruber.* These methane metabolism pathway maps are from The Kyoto Encyclopedia of Genes and Genomes (KEGG) database at http://www.genome.jp/kegg/pathway.html.

In order to determine the structural similarities between *E.coli serA* and *Mrub_0173*, we used TIGRFAM, Pfam, and PDB. TIGRFAM determines similar protein structures. For both genes, TIGRFAM resulted in TIGR01327 (Table 1). The name of this family is the PGDH: phosphoglycerate dehydrogenase. *E.coli serA* had a significant E-value of 5e-34 and a score of 124.3. *Mrub_0173* also had a significant E-value of 2.7e-181 and a score of 613.5.

A search of the Pfam database identified the protein family and number for each gene (Figure 6). They had a first hit result of PF00389 known as 2-Hacid_dh. *E.coli serA* and *Mrub_0173* had significant E-values of 1.1e-38 and 1.1e-32 respectively. *E.coli serA* had a score of 131.8 and *Mrub_0173* had a score of 112.3. The pairwise alignments in Figure 6 show us that both *E.coli serA* and *Mrub_0173* contain the same extremely conserved glycine and valine residues towards the middle of the protein sequence. The pairwise alignment compares each sequence to a consensus sequence that has been generated from various other proteins. Since the *E.coli serA* and *Mrub_0173* sequences pulled the same exact consensus sequence, this is even more evidence that the two genes are orthologs.

Panel A



Panel B

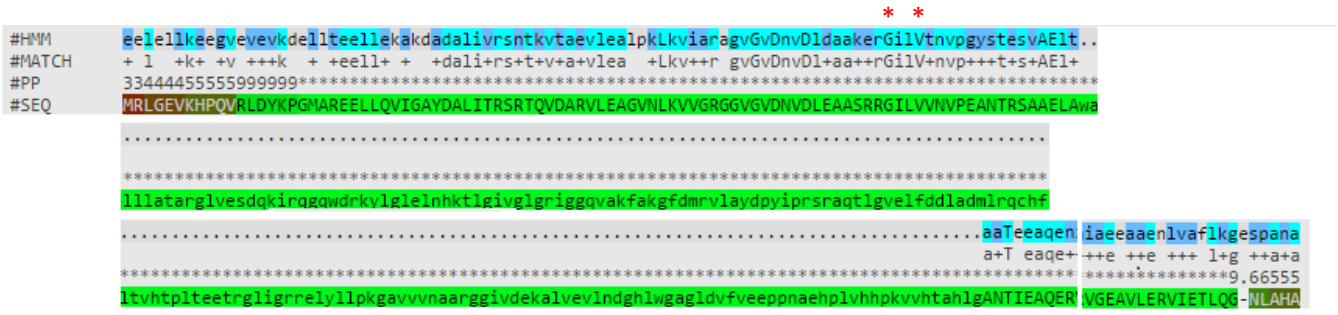


Figure 6. Pfam shows *E.coli serA* and *Mrub_0173* have the same highly conserved amino acids; Both genes code for same family, 2-Hacid_dh (PF00389). Panel A displays the pairwise alignment for *E.coli b2913/serA* (#SEQ) against the consensus sequence (#HMM). Panel B displays the pairwise alignment for *Mrub_0173* (#SEQ) against the consensus sequence (#HMM). The conserved glycine and valine are marked with a red asterisks. These alignments were made by Pfam at http://pfam.sanger.ac.uk/search.

A PDB hit reveals sequence similarity to a protein whose crystal structure has been determined. For *E.coli serA*, the PDB code was 1PSD and the name was "The allosteric Ligand site in the vmax-type cooperative enzyme phosphoglycerate dehydrogenase." The resulting E-value for this was 0.0. *Mrub_0173* has a PDB code of 3DDN with the name being "Crystal structure of hydroxypyruvic acid phosphate bound D-3-phosphoglycerate dehydrogenase in mycobacterium tuberculosis." It had an E-value of 8.01702e-66. Although it resulted in two different structures, we can see that both of the names include phosphoglycerate dehydrogenase. We may infer that the results aren't exactly the same because PDB pulled the same enzyme from two different organisms. This may be because *M.ruber* is more closely related the organism that the PDB pulled for its result.

The images from Figure 7 show the ortholog neighborhoods of *E.coli serA* and *Mrub_0173* from the IMG/EDU gene finder tool. The *E.coli serA* gene, underlined in red in Panel A, is next to a few genes going in the same direction. Because they are not similar colors, we can conclude that the genes are not in an operon. This is the same result for *Mrub_0173.* The *Mrub_0173* gene is underlined in red in Panel B and is also next to a few genes that are going in the same direction. Those genes are not similar colors to the *Mrub_0173* gene, so we can conclude they do not have similar functions and therefore are not a part of an operon.

Panel A

Panel B

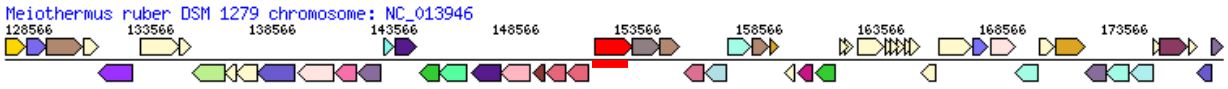
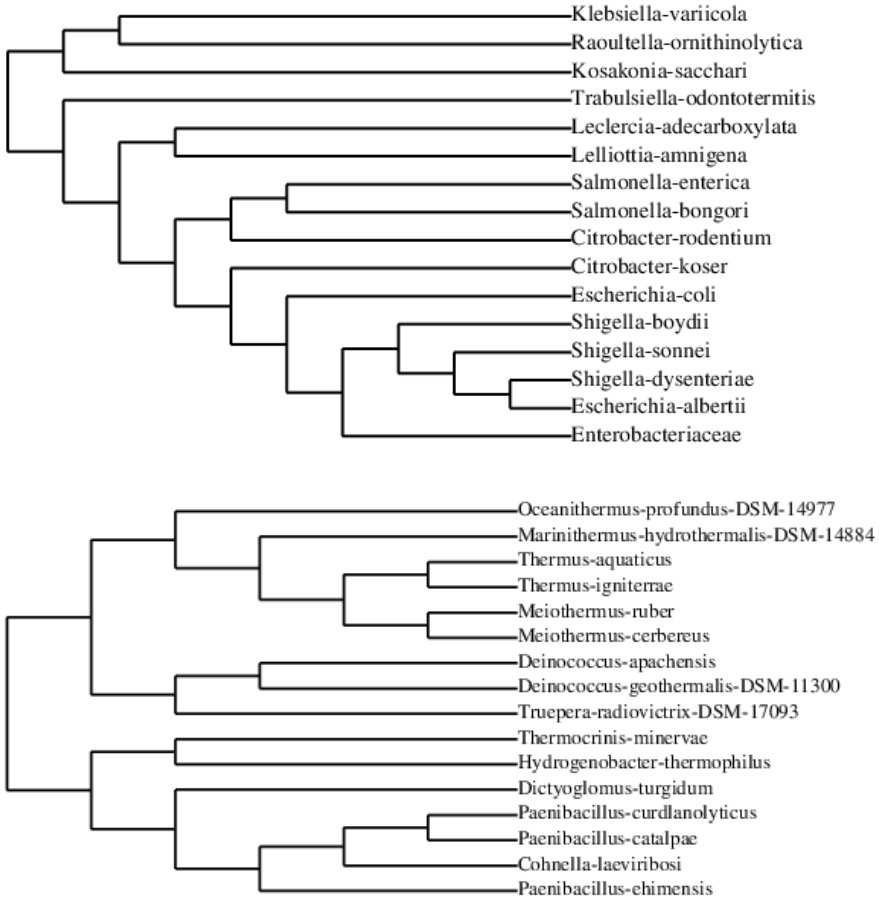


Figure 7. *E.coli serA* and *Mrub_0173* genes are not a part of an operon. Panel A illustrates the *E.coli b2913/serA* Chromosome Viewer; Panel B is displaying the *Mrub_0173* Chromosome Viewer. Red lines indicate the location of each gene on their respective viewer. Neighborhood region images from IMG/EDU at https://img.jgi.doe.gov/.

Phylogeny.fr is a bioinformatics tool that creates phylogenetic trees from the multiple sequence alignments from T-Coffee. It provides evidence for or against horizontal gene transfer (HGT) being an option in for these particular genes. Figure 8 shows the *E.coli serA* tree (Panel A) shows that all of the surrounding species are from the phylum Proteobacteria. This means that it is unlikely that HGT occurred for this gene. Panel B in Figure 8 shows the *Mrub_0173* tree. It shows that the proteins most closely related to the one encoded by *Mrub_0173* are from the same phylum, Deinococcus-Thermus. Although there are other species from the Aquificae, Dictyoglomus, and Firmicutes, phylum on the map, they aren't directly surrounding the *M. ruber* species. Therefore, it is also not very likely that HGT occurred for the *Mrub_0173* gene.



Panel A

Panel B

Figure 8. Horizontal gene transfer is not likely because of each query is surrounded by numerous species within its respective phylum. Panel A is the phylogenetic tree for *E.coli b2913/serA.* Panel B is the phylogenetic tree for *Mrub_0173.* Multiple sequence alignments from T-Coffee were used for each gene to build the respective phylogenetic trees. The phylogenetic trees are from Phylogeny.fr at http://www.phylogeny.fr/

Table 2 is a summary of the results from the bioinformatics tools comparing *E. coli b4388*(*serB*) to *Mrub_0125*. The BLASTp result that initially preformed is presented in the first row. The two amino acid sequences were different lengths, so the low bit score is understandable. The E-value of the BLAST is 8e-06, which is close to zero. We can assume that these two sequences did not align simply by chance and share many of the same amino acids, suggesting functional similarities, because of this result. The CDD database pulled the same COG number (COG0560), SerB, for the two proteins. Both results had significantly small E-values, which indicate that they could be the same enzyme in the serine biosynthesis pathway. A combination of TMH, SignalP, LipoP, and PSORT-B suggest that the two proteins are localized to the cytoplasm. The lack of a cleavage site indicates they are not membrane bound or traverse a membrane. The similarity in the location of the products of these two genes is even more evidence that they are orthologs. Moreover, while the TIGRfam tool resulted two different, but similar results, literature sites that phosphoserine phosphatase is actually in the HAD Hydrolase family (Arora et al. 2014). The TIGRfam number for *E.coli b4388* was TIGR00338 which is named serB: phosphoserine phosphatase SerB and the number for *Mrub_0125* was TIGR01490 which is named HAD-SF-IB-hyp1: HAD hydrolase, family IB. They both had extremely small E-values for their respective TIGRfam result, indicating the results were not likely to be due to chance. Furthermore, the Pfam database pulled the same domains for the two proteins, HAD – haloacid dehalogenase-like hydrolase (PF12710). The PDB pulled two different names and numbers for each sequence, but this may be because *E. coli* and *M.ruber* are not closely related and therefore PDB pulled different related specie's enzymes for each. *E. coli b4388*(*serB*) and *Mrub_0125* both had the same enzyme commission number, E.C.3.1.3.3 and were both predicted to be involved the a sub-pathway of methane metabolism, serine synthesis.

Table 2: *E. coli b4388*(*serB*) is orthologous to *Mrub_0125*

Bioinformatics tool used	***E. Coli b4388* gene (*serB*)**	***M. ruber Mrub_0125* gene**
BLAST *E.coli* against *M.ruber*	Score: 45.4 bits E-value: 8e-06	
CDD Data (COG category)	COG Number: COG0560 SerB	
	E-value: 1.73e-83	E-value: 1.13e-19
Cellular Localization	Cytoplasm of the cell	
TIGRfam – protein family	TIGR00338 serB: phosphoserine phosphatase SerB	TIGR01490 HAD-SF-IB-hyp1: HAD hydrolase, family IB
	E-value: 1.6e-144	E-value: 9.5e-22
Pfam – protein family	1) PF12710 (HAD - haloacid dehalogenase-like hydrolase)	
	E-values: 1) 1.3e-18	E-values: 1) 2.9e-20
Protein Database	3N28 – Crystal structure of probable phosphoserine phosphatase from vibrio cholerae, unliganded form	3VFF – The crystal structure of the protein with unknown function from Bordetella pertussis Tohama I
	E-value: 4.39845E-81	E-value: 5.17358E-6
Enzyme commission number	E.C. 3.1.3.3 – Phosphoserine phosphatase	
KEGG pathway map	Pathway ID: 00680 Methane Metabolism	

The image in Figure 9 demonstrates the results of the initial BLASTp search of *E.coli serB* against *Mrub_0125*. This figure displays the 23% of the amino acids were identical between the two sequences and 49 out of 108 amino acids were characteristically similar. The E-value was 8e-06, which is close to zero. From this we can conclude that these two sequences did not line up by chance, but represent structural and functional similarities. This is where we get our initial indication that *E.coli serB* and *Mrub_0125* share some key similarities. This first BLAST suggests that the two genes might be orthologs.

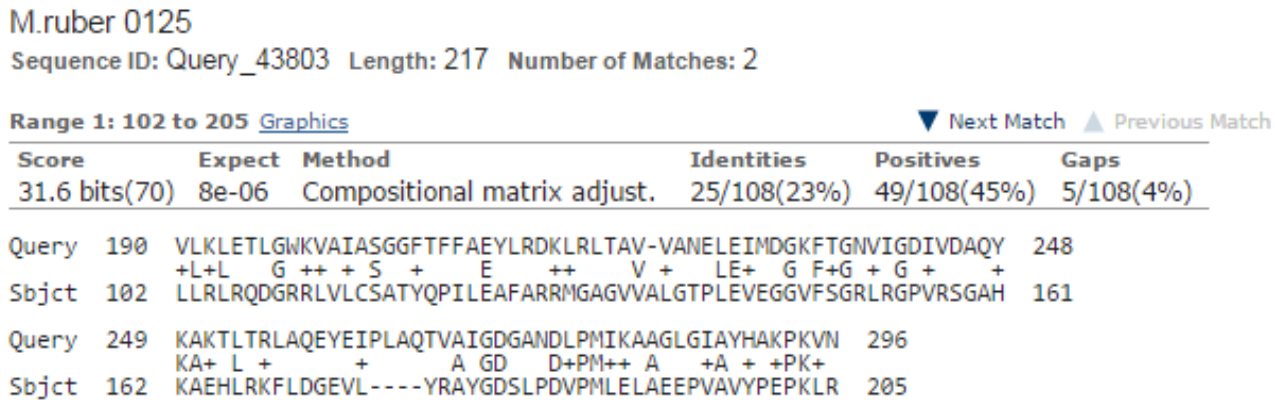
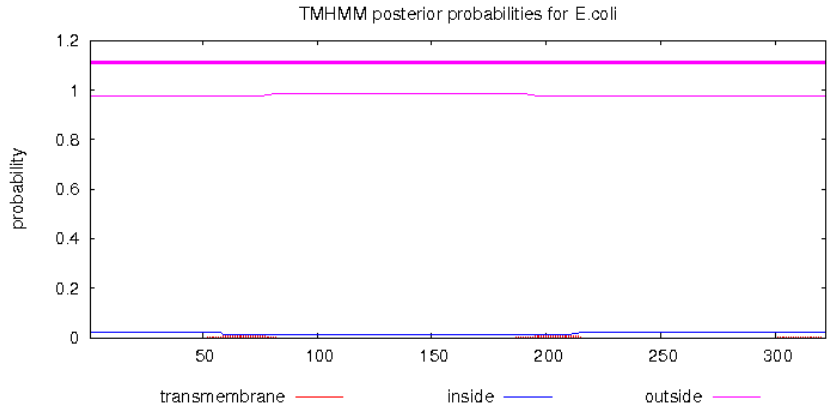


Figure 9. *E.coli serB* and *Mrub_0125* have similar amino acid sequence. Query sequence: *E. coli serB.* Subject sequence: *Mrub_0125.* Analysis was performed using the NCBI BLAST bioinformatics tool at http://www.ncbi.nlm.nih.gov.

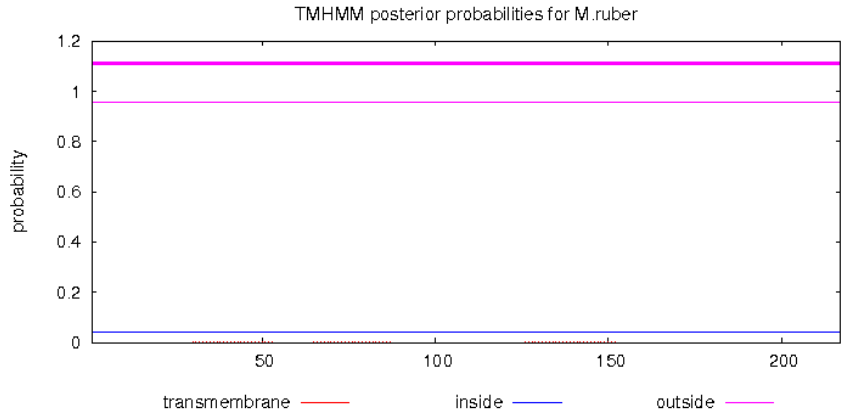
Figure 10 is a representation of the TMHMM hydropathy plots for *E.coli serB* and *Mrub_0125*. Again, red peaks in a TMHMM plot that rise above a certain threshold represent the presence of transmembrane helices. Neither panel show any red peaks are shown on the plot, so neither gene has any transmembrane helices. Subsequently, both of the TMHMM hydropathy charts show that the proteins coded by these genes are not in the membrane, but are in the cytoplasm.

```
# E.coli Length: 322
# E.coli Number of predicted TMHs:  0
# E.coli Exp number of AAs in TMHs: 0.45508
# E.coli Exp number, first 60 AAs:  0.03227
# E.coli Total prob of N-in:        0.02162
E.coli  TMHMM2.0        outside      1   322
```



Panel A

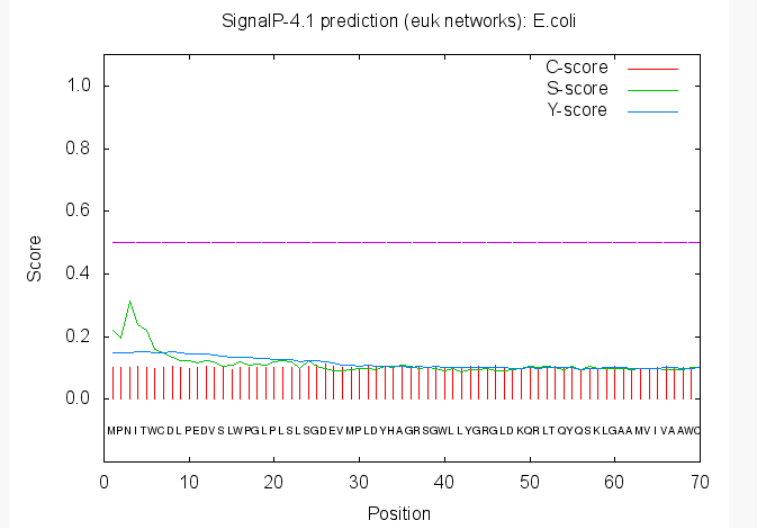
```
# M.ruber Length: 217
# M.ruber Number of predicted TMHs:  0
# M.ruber Exp number of AAs in TMHs: 0.11952
# M.ruber Exp number, first 60 AAs:  0.08974
# M.ruber Total prob of N-in:        0.04252
M.ruber TMHMM2.0        outside      1   217
```



Panel B

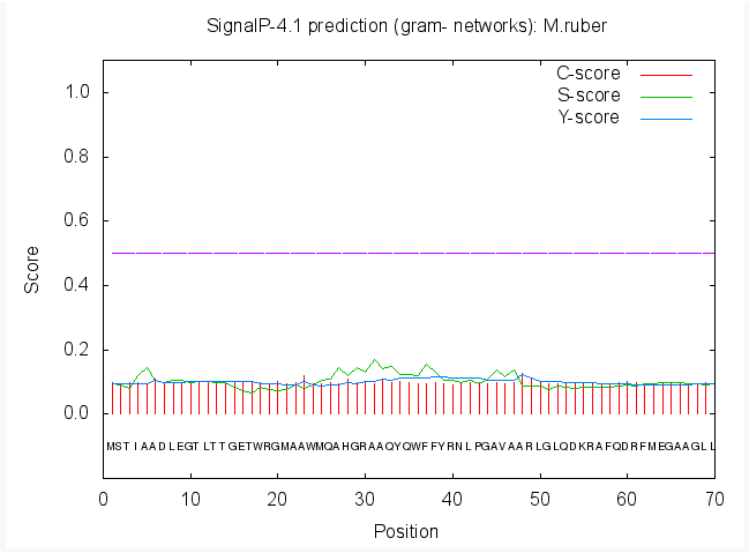
Figure 10. *E.coli serB* and *Mrub_0125* do not contain TMH regions; both proteins predicted to be located in cytoplasm. Panel A shows the TMHMM for *E.coli b4388/serB;* Panel B shows the TMHMM for *Mrub_0125*. TMHMM Server v 2.0 http://www.cbs.dtu.dk/services/TMHMM was utilized to create the hydropathy charts.

The maps in Figure 11 are SignalP graphs for *E.coli serB* and *Mrub_0125.* As previously explained, the SignalP bioinformatics tool is used to predict protein cleavage sites. It calculates a D-value by using the S-score and Y-score. If the D-value is lower than the cutoff value, represented by the purple line, then the protein does not have any cleavage sites for proteins that might be bound to the cell membrane. *E.coli serB* (Panel A) has a D-value of 0.168 that is not above the cutoff value of 0.450. *Mrub_0125* (Panel B) has a D-value of 0.116, which is also below the cutoff. These results suggest that both of the proteins do not contain cleavage sites, and therefore remain in the cytoplasm.




```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    26       0.114
  max. Y    11       0.145
  max. S     3       0.315
  mean S     1-10    0.187
       D     1-10    0.168   0.450   NO
```

Panel A




```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    48       0.126
  max. Y    48       0.122
  max. S    31       0.171
  mean S     1-47    0.109
       D     1-47    0.116   0.570   NO
```

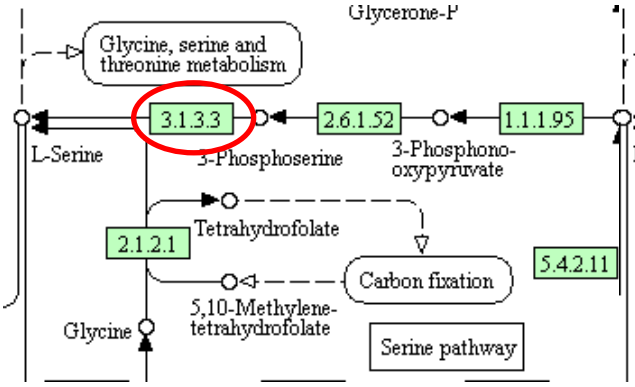
Panel B

Figure 11. *E.coli serA* and *Mrub_0173* do not have cleavage sites; D values for both charts were below cutoff value. Panel A shows the plot for *E.coli b2913/serA*; Panel B shows the plot for *Mrub_0173*. Plots created by Signal P server v. 4.1 http://www.cbs.dtu.dk/services/SignalP.

The other two bioinformatics tools used to determine cellular localization of the proteins were LipoP and PSORT-B. The LipoP tool projected that both *E.coli serB* and *Mrub_0125* were located in the cytoplasm of the cell and neither had any cleavage sites. PSORT-B showed cytoplasmic score of 9.97 and cytoplasmic membrane and periplasmic scores of 0.01 for *E.coli serB.* PSORT-B showed the same results for *Mrub_0125.* The final calculation from PSORT-B for the cellular locations of both genes was in the cytoplasm. The final results were the same for both genes and so the 0.01 scores for the cytoplasmic membrane and periplasmic locations are not significant. We did not utilize the Phobius bioinformatics tool because the other cellular location bioinformatics tools indicate that the genes do not have any cleavage sites. All bioinformatics tools reveal that both *E.coli serB* and *Mrub_0125* are in the cytoplasm (Table 2).

Figure 12 presents the serine sub-pathway of methane metabolism. The enzymes that are colored green are purportedly existent in the organism. We can see that both the *E.coli serB* and *Mrub_0125* are both predicted to be involved in the first step in the serine biosynthesis and code for the same enzyme, phosphoserine phosphatase. This confirmation is helpful in determining if the two genes are orthologous

Panel A



Panel B

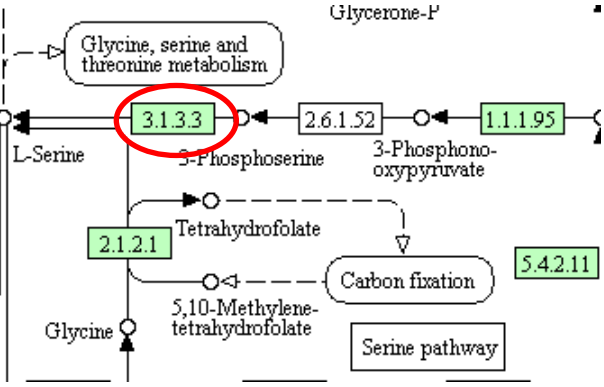


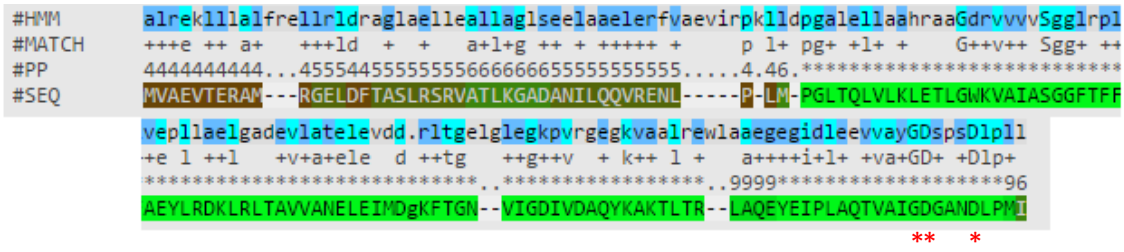
Figure 12. *E.coli serB* and *Mrub_0125* occur in the same biochemical pathway. Panel A shows the Serine KEGG pathway when focusing on *Escherichia coli.* Panel B shows the Serine KEGG pathway after selecting for *Meiothermus ruber.* These methane metabolism pathway maps are from The Kyoto Encyclopedia of Genes and Genomes (KEGG) database at http://www.genome.jp/kegg/pathway.html.

.

The structural similarities between *E.coli serB* and *Mrub_0125* were determine by using TIGRfam, Pfam, and PDB. TIGRfam determines similar protein structures. TIGRfam number for *E.coli b4388* was TIGR00338 which is named serB: phosphoserine phosphatase SerB. On the other hand, *Mrub_0125* had a TIGRfam number of TIGR01490 which is named HAD-SF-IB-hyp1: HAD hydrolase, family IB (Table 2). While the results for TIGRfam aren't exactly the same, it turns out that they are still representing relatively the same thing. According to Arora et al., phosphoserine phosphatase enzymes are part of the haloacid dehalogenase (HAD) superfamily. This means that if we were looking at just this bioinformatics tool we would be able to determine that they are, at the very least, in the same superfamily. *E.coli serB* had a significant E-value of 1.6e-144 and a score of 491.4. *Mrub_0125* also had a significant E-value of 9.5e-22 and a score of 83.5.

A search of the Pfam database identified the same results for both of the genes. The first hit result was PF12710, HAD – haloacid dehalogenase-like hydrolase. *E.coli serB* and *Mrub_0125* both had significant E-values of 1.3e-18 and 2.9e-20 respectively. *E.coli serB* had a score of 67.9 and *Mrub_0125* had a score of 73.3. Figure 13 displays the pairwise alignments and shows that both *E.coli serB* and *Mrub_0125* include the same conserved glycine and aspartic acid amino acids at the end of the protein sequence. The alignment associates each sequence to a consensus sequence that is pulled from many other proteins. The *E.coli serB* and *Mrub_0125* sequences drew the same consensus sequence and, because of this, it is even more evidence that the two genes are orthologs.

Panel A



Panel B

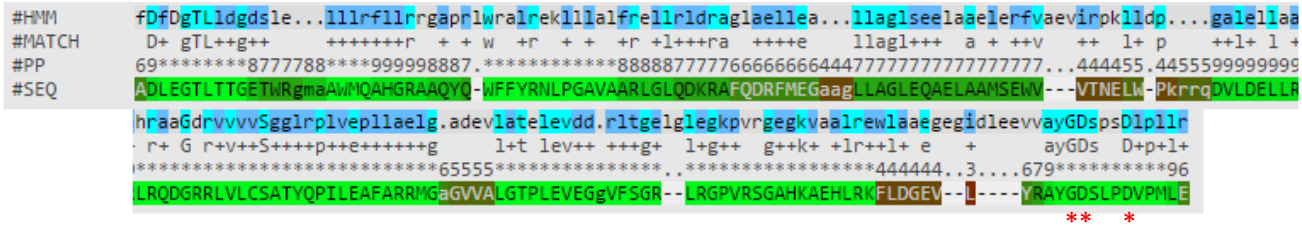
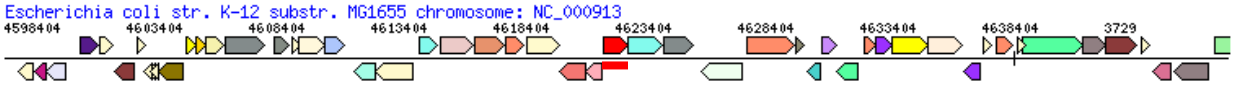


Figure 13. *E.coli serB* and *Mrub_0125* have the same highly conserved amino acids; Both genes code for same domain, HAD – haloacid dehalogenase-like hydrolase (PF12710). Panel A demonstrations the pairwise alignment for *E.coli b4388/serB* (#SEQ) against the consensus sequence (#HMM). Panel B displays the pairwise alignment for *Mrub_0125* (#SEQ) against the consensus sequence (#HMM). The conserved glycine and aspartic acids are indicated by the red asterisks. These alignments were made by Pfam at http://pfam.sanger.ac.uk/search.

PDB, or the protein database, shows the similarities to a protein whose crystal structure has been determined. For *E.coli serB*, the PDB code was 3N28 with the name "Crystal structure of probable phosphoserine phosphatase from vibrio cholerae, unliganded form." The E-value for this result was 4.39845E-81. *Mrub_0125* has a PDB code of 3VFF and was named "The crystal structure of the protein with unknown function from Bordetella pertussis Tohama I." It had an E-value of 5.17358E-6. The result from *E.coli serB* is consistent with the predicted annotation. *Mrub_0125*, on the other hand, had a strange result. Because the gene matched with a protein that has an unknown function, it does not tell us much about the protein. The outcome of PDB for these two genes is not helpful in determining if they are orthologs or are related.

The images from Figure 14 illustrate the ortholog neighborhoods of *E.coli serB* and *Mrub_0125* from the IMG/EDU gene finder tool. The *E.coli serB* gene is underlined in red in Panel A and is next to genes going in the same direction. The genes that are next to it are not the same color and do not have the same function so, we can conclude that the genes are not in an operon. Results are the same for *Mrub_0125.* The *Mrub_0125* gene, underlined in red in Panel B, is next to a few genes that are also going in the same direction. The genes going in the same direction are not similar colors to the *Mrub_0125* gene. This means that we can infer they do not have similar functions and therefore are not a part of an operon.

Panel A



Panel B

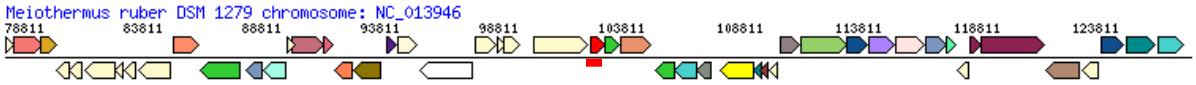
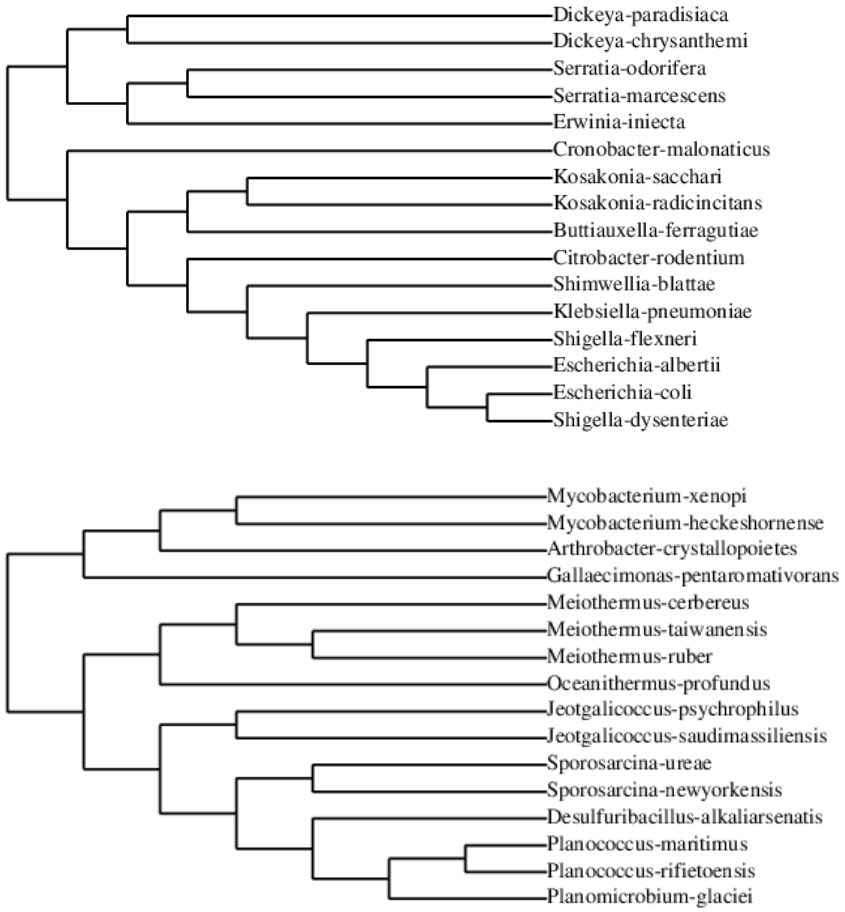


Figure 14. *E.coli serB* and *Mrub_0125* genes are not a part of an operon. Panel A shows the *E.coli b4388/serB* Chromosome Viewer; Panel B is presenting the *Mrub_0125* Chromosome Viewer. Red lines indicate the location of each gene on their respective viewer. Neighborhood region images from IMG/EDU at https://img.jgi.doe.gov/.

Phylogeny.fr uses T-Coffee multiple sequence alignments to provide evidence for or against horizontal gene transfer (HGT). Figure 15 depicts the *E.coli serB* tree (Panel A) that shows that all of the surrounding species are from the phylum Proteobacteria. This result means that it is unlikely that HGT transpired for this gene. Panel B in figure 15 shows the *Mrub_0125* tree. It shows that the four most directly related proteins to *Mrub_0125* are within the Deinococcus-Thermus phylum, but the rest of the tree contains less related species. There are other species from the Aquificae, Dictyoglomus, Proteobacteria, and Firmicutes phylum on the tree, but they aren't directly surrounding the *M. ruber* species. Consequently, HGT is possible to have occurred for the *Mrub_0125* gene.



Panel A

Panel B

Figure 15. Horizontal gene transfer is possible because there are numerous species in the *M.ruber* tree that are not within its respective phylum. Panel A is the phylogenetic tree for *E.coli b4388/serB*. Panel B is the phylogenetic tree for *Mrub_0125*. T-Coffee multiple sequence alignments were used for each gene to build the phylogenetic trees. The phylogenetic trees are from Phylogeny.fr at http://www.phylogeny.fr/

Table 3 is a rundown of the results from the bioinformatics tools comparing *E. coli b2551*(*glyA*) to *Mrub_2910*. The initial BLASTp outcome that we completed is shown in the first row of Table 3. The two amino acid sequences were about the same length, so the bit score is more understandable in this comparison. The bit score was 446 and the E-value of the BLASTp is 1e-154, which is close to zero. We can presume that these two sequences have similar amino acid sequences, suggesting functional similarities, and do not align simply by chance. A search of the CDD database gave the same COG number (COG0112), GlyA, for the two proteins. Both gave significantly small E-values, which indicate that they could be the same enzyme in the glycine biosynthesis/degradation pathway. A combination of TMH, SignalP, LipoP, and PSORT-B suggest that both of the proteins are localized to the cytoplasm of the cell. The lack of a cleavage site indicates they are not membrane-bound nor traverse a membrane. In addition, the TIGRfam tool did not pull any type of result for either gene. This is a strange outcome, but because it happened for both genes it does not harm our comparison. Moreover, Pfam outcomes showed that the proteins have the same domains, SHMT – Serine hydroxymethyltransferase (PF00464). The protein database (PDB) pulled two different names and numbers for each sequence, but again this may be because the database pulls organisms that are closely related to the input sequence. *E. coli b2551* (*glyA*) and *Mrub_2910* both had the same enzyme commission number, E.C.2.1.2.1. They were both predicted to be involved in the same step of glycine biosynthesis and degradation which is a sub-pathway of methane metabolism.

Table 3: *E. coli b2551*(*glyA*) is orthologous to *Mrub_2910*

Bioinformatics tool used	***E. Coli b2551* gene (*glyA*)**	***M. ruber Mrub_2910* gene**
BLAST *E.coli* against *M.ruber*	Score: 446 bits E-value: 1e-154	
CDD Data (COG category)	COG Number: COG0112 GlyA	
	E-value: 0.0	E-value: 1.73e-98
Cellular Localization	Cytoplasm of the cell	
TIGRfam – protein family	None	
	None	None
Pfam – protein family	1) PF00464 (SHMT - Serine hydroxymethyltransferase)	
	E-values: 1) 5e-192	E-values: 1) 3.3e-164
Protein Database	1DFO – Crystal Structure at 2.4 angstrom resolution of E. coli serine hydroxymethyltransferase in complex with glycine and 5 formyl tetrahydrofolate	2DKJ – Crystal Structure of T.th.HB8 Serine Hydroxymethyltransferase
	E-value: 0.0	E-value: 2.97584E-167
Enzyme commission number	E.C. 2.1.2.1 – Glycine hydroxymethyltransferase	
KEGG pathway map	Pathway ID: 00680 Methane Metabolism	

The image in Figure 16 is the results of the preliminary BLAST search of *E.coli glyA* against *Mrub_2910*. The figure shows that 55% of the amino acids were identical between the sequences and 293 amino acids were characteristically similar. With a close to zero E-value of 1e-154, we can determine that these two sequences were not aligned by chance and represent structural and functional similarity. We can start to see that *E.coli glyA* and *Mrub_2910* might share some important structural resemblances. This BLAST search is the first sign that these genes might be orthologs.

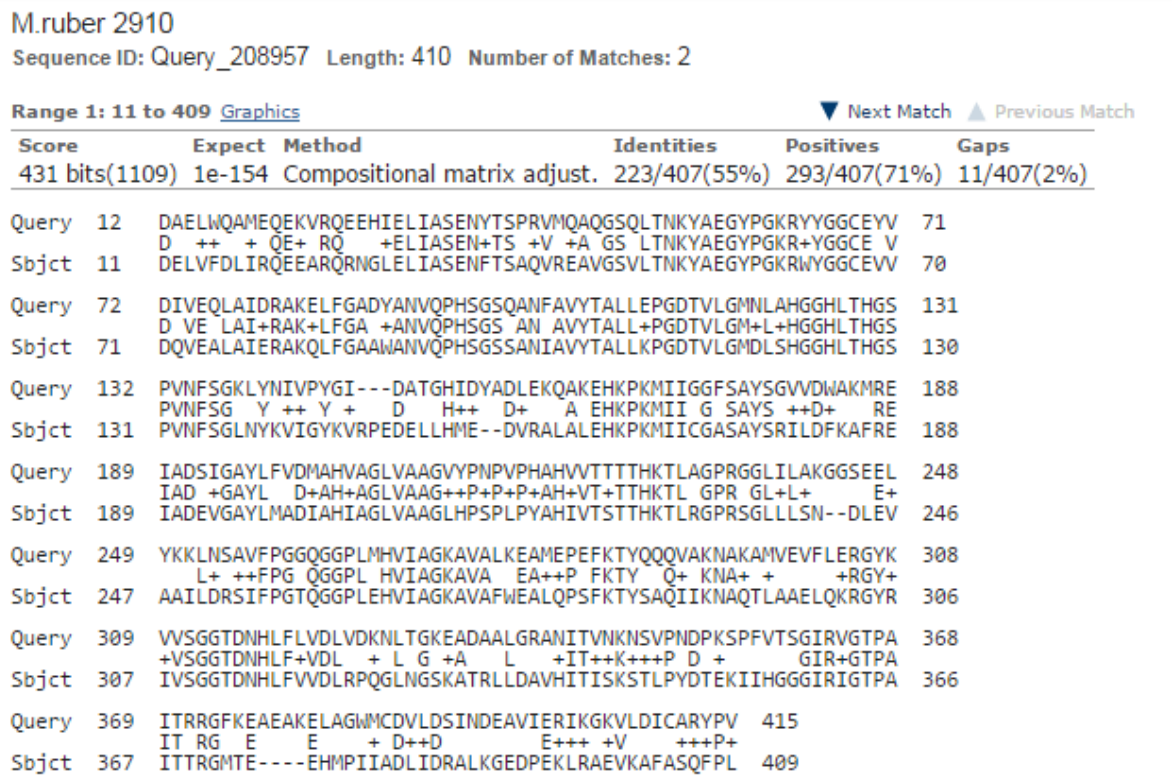
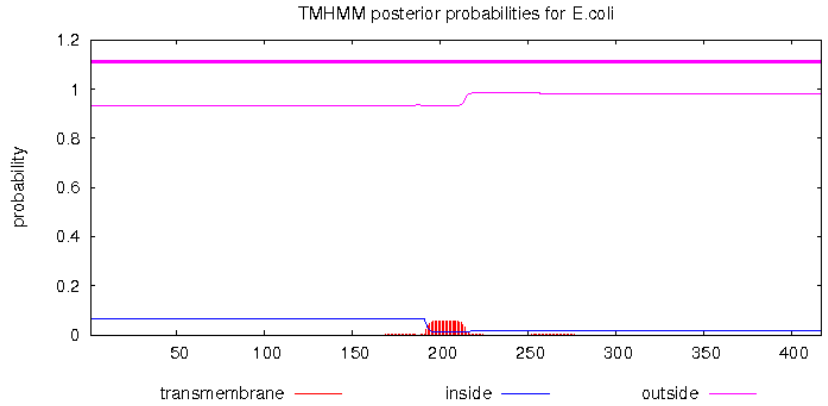


Figure 16. *E.coli glyA* and *Mrub_2910* have similar amino acid sequences. Query sequence: *E. coli glyA;* Subject sequence: *Mrub_2910.* Analysis was performed using the NCBI BLAST bioinformatics tool at http://www.ncbi.nlm.nih.gov.

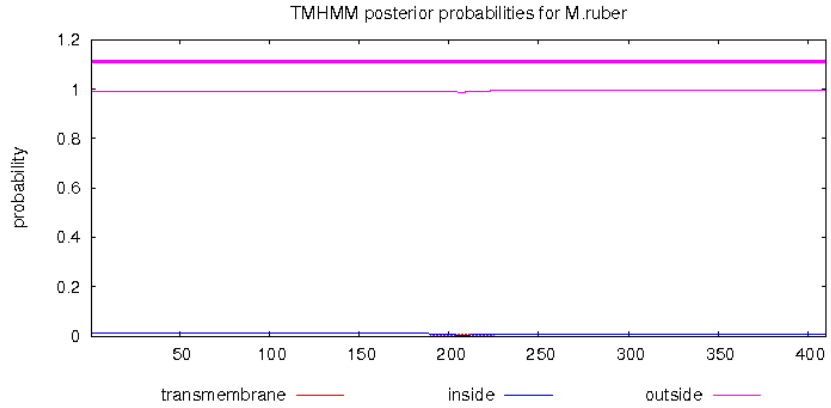
Figure 17 displays the TMHMM hydropathy charts for *E.coli glyA* and *Mrub_2551*. The red in a TMHMM that rise above a certain point represent the presence of transmembrane helicies. Panel A shows a very short red peak, but the height of the peak is not nearly high enough to be significant. Consequently, both of the TMHMM hydropathy plots show that the proteins coded by these genes are in the cytoplasm and not in the membrane of the cell.

```
# E.coli Length: 417
# E.coli Number of predicted TMHs:  0
# E.coli Exp number of AAs in TMHs: 1.22075
# E.coli Exp number, first 60 AAs:  0
# E.coli Total prob of N-in:        0.06537
E.coli  TMHMM2.0        outside      1   417
```



Panel A

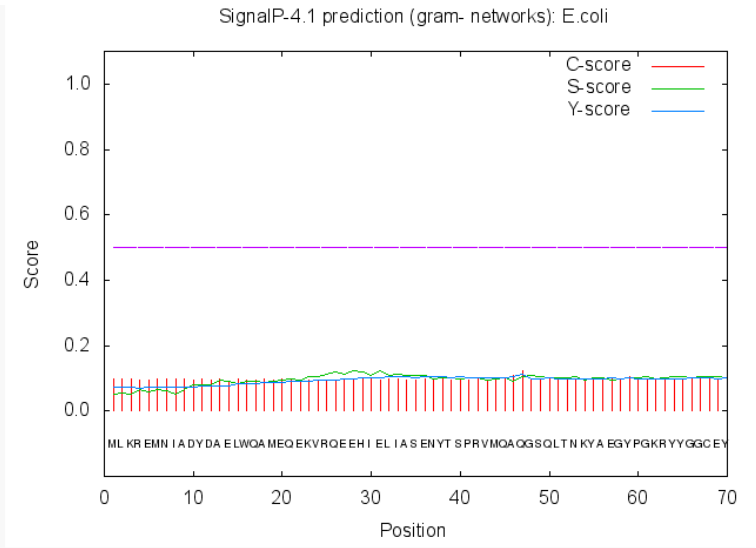
```
# M.ruber Length: 410
# M.ruber Number of predicted TMHs:  0
# M.ruber Exp number of AAs in TMHs: 0.15115
# M.ruber Exp number, first 60 AAs:  0
# M.ruber Total prob of N-in:        0.00997
M.ruber TMHMM2.0        outside      1   410
```



Panel B

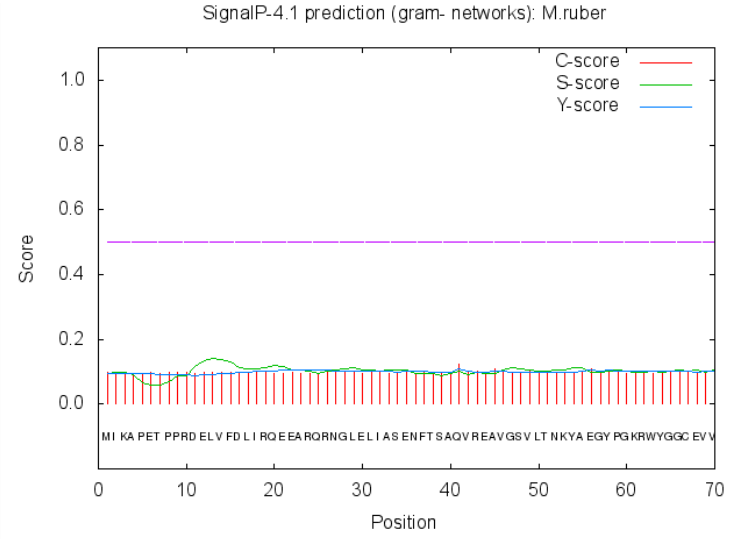
Figure 17. *E.coli glyA* and *Mrub_2910* do not contain TMH regions; both are predicted to be located in cytoplasm. Panel A shows the TMHMM for *E.coli b2551/glyA;* Panel B shows the TMHMM for *Mrub_2910*. TMHMM Server v 2.0 http://www.cbs.dtu.dk/services/TMHMM was used to make the hydropathy charts.

Figure 18 are the SignalP graphs for *E.coli glyA* and *Mrub_2910.* Again, the SignalP tool is used in order to predict protein cleavage sites for proteins that might be bound to the cell membrane or pass through the membrane. If the D-value is lower than the cutoff value, represented by the purple line, then the protein does not have any cleavage sites. *E.coli glyA* (Panel A) has a D-value of 0.103 that is not above the cutoff value of 0.570. *Mrub_2910* (Panel B) has a D-value of 0.107, which is also below the cutoff. This data indicates that both of the proteins do not contain cleavage sites, and therefore remain in the cytoplasm.




```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    47       0.123
  max. Y    47       0.111
  max. S    31       0.124
  mean S     1-46    0.092
       D     1-46    0.103   0.570   NO
Name=E.coli     SP='NO' D=0.103 D-cutoff=0.570 Networks=SignalP-noTM
```

Panel A




```
# Measure  Position  Value    Cutoff   signal peptide?
  max. C    41       0.122
  max. Y    41       0.110
  max. S    13       0.142
  mean S     1-40    0.102
       D     1-40    0.106   0.570   NO
Name=M.ruber    SP='NO' D=0.106 D-cutoff=0.570 Networks=SignalP-noTM
```

Panel B

Figure 18. *E.coli glyA* and *Mrub_2910* do not have cleavage sites; D values were below cutoff values. Panel A shows the plot for *E.coli b2551/glyA*; Panel B shows the plot for *Mrub_2910*. Graphs made by Signal P server v. 4.1 http://www.cbs.dtu.dk/services/SignalP.

Two other bioinformatics tools were also used to determine cellular localization of the proteins. The LipoP tool predicted that both *E.coli glyA* and *Mrub_2910* are found in the cytoplasm of the cell and do not have any cleavage sites. The results for *E.coli glyA* on PSORT-B was a score of 10.0 for cytoplasmic location with the rest of the scores being 0.0. PSORT-B showed cytoplasmic score of 9.97 and cytoplasmic membrane and periplasmic scores of 0.01 for *Mrub_2910.* The 0.01 scores for the cytoplasmic membrane and periplasmic locations for *Mrub_2910* are insignificant. The final prediction from PSORT-B for both genes' cellular locations was that they are in the cytoplasm. Because all of our results for the cellular localization tools do not indicate any cleavage sites, we do not use the Phobius tool. All bioinformatics tools predict that both *E.coli glyA* and *Mrub_2910* are in the cytoplasm (Table 3).

The pathways in Figure 19 show the serine sub-pathway of methane metabolism. The green colored enzymes are thought to be present in the organism. We can see that the *E.coli glyA* and *Mrub_2910* are predicted to be a part of singular, reversible step in glycine synthesis/degradation and code for the same enzyme, glycine hydroxymethyltransferase. This is even more confirmation that the two genes are orthologous to one another.

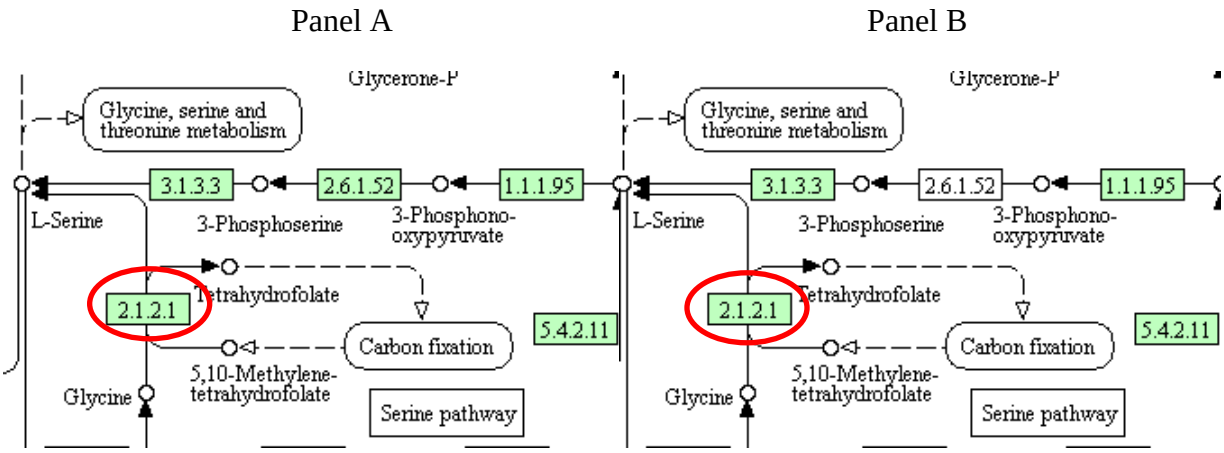
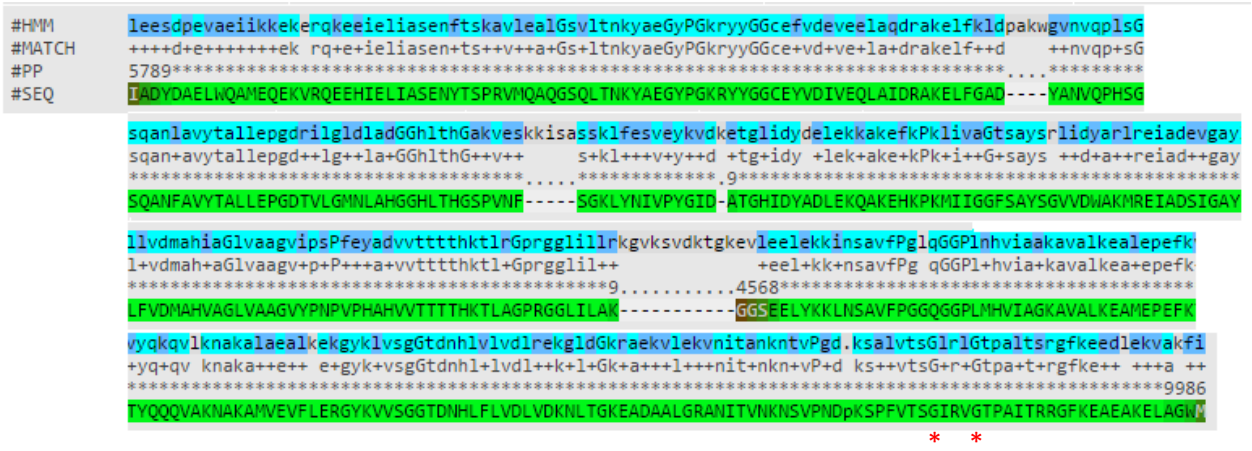


Figure 19. *E.coli glyA* and *Mrub_2910* are in the same biochemical pathway. Panel A shows the KEGG pathway when selecting *Escherichia coli.* Panel B shows the KEGG pathway after focusing on *Meiothermus ruber.* These methane metabolism pathway maps are from The Kyoto Encyclopedia of Genes and Genomes (KEGG) database at http://www.genome.jp/kegg/pathway.html.

To determine the structural similarities between *E.coli glyA* and *Mrub_2910* we utilized TIGRFAM, Pfam, and PDB. TIGRFAM determines similar protein structures. For both genes, TIGRFAM did not give any results. This outcome may be because of our cut off value we chose and because not every known protein has a protein family that is described in TIGRfam. This result does not tell us much about the two genes orthologous nature, but because both genes gave us nothing, it is not contradictory evidence for our claims (Lori Scott, personal communication). While this gives us little information, the other bioinformatics tools that were utilized for structural evidence do give us good data.

A search of the Pfam database identified the same protein family and number for each gene (Figure 6). The result of the first hit was PF00464 known as SHMT – serine hydroxymethyltransferase. *E.coli glyA* and *Mrub_2910* had significant E-values of 5e-192 and 3.3e-164 respectively. *E.coli glyA* had a score of 637.7 and *Mrub_2910* had a score of 546.1. The pairwise alignments in Figure 20 tell us that both *E.coli glyA* and *Mrub_2910* have the extremely conserved glycine at the end of the amino acid sequence. The alignment compares each sequence to a consensus sequence from other proteins. *E.coli glyA* and *Mrub_2910* sequences compare against the same exact consensus sequence. This is even more of an indication that the two genes are orthologs.

Panel A



Panel B

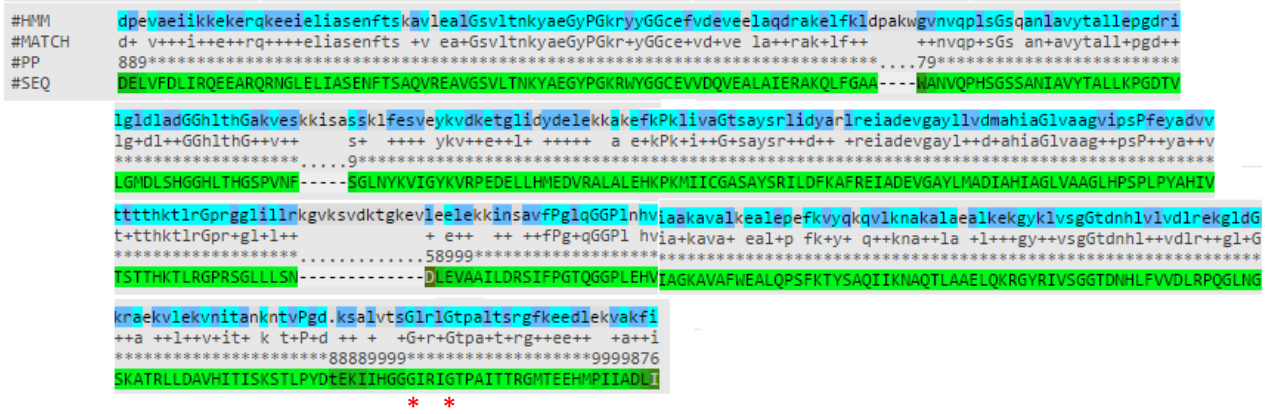
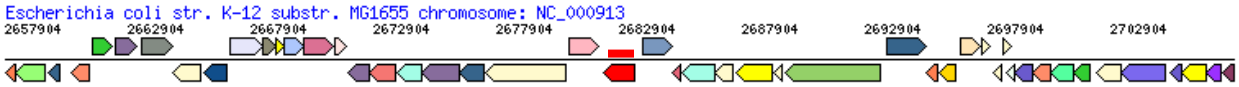


Figure 20. *E.coli glyA* and *Mrub_2910* have the same highly conserved amino acids; Both genes code for same family, SHMT – serine hydroxymethyltransferase (PF00464). Panel A presents the pairwise alignment for *E.coli b2551/glyA* (#SEQ) against the consensus sequence (#HMM). Panel B shows the pairwise alignment for *Mrub_2910* (#SEQ) against the consensus sequence (#HMM). The highly conserved glycine amino acids are indicated by the red asterisks. These alignments were made by Pfam at http://pfam.sanger.ac.uk/search.

A PDB hit reveal sequence similarity to a protein whose crystal structure has been determined. For *E.coli glyA*, the PDB code was 1DFO and the name was "Crystal Structure at 2.4 angstrom resolution of E. coli serine hydroxymethyltransferase in complex with glycine and 5 formyl tetrahydrofolate." The resulting E-value for this was 0.0. *Mrub_2910* has a PDB code of 2DKJ with the name being "Crystal Structure of T.th.HB8 Serine Hydroxymethlytransferase." It had an E-value of 2.97584E-167. While the results were different for the two genes, we see that both of the names include serine hydroxymethyltransferase. We may infer that the results are not identical because the PDB is pulling different specie's serine hydroxymethyltransferase for *E.coli* and *M.ruber.*

The images from Figure 21 display the ortholog neighborhoods of *E.coli glyA* and *Mrub_2910* from the IMG/EDU tool. The *E.coli glyA* gene and *Mrub_2910* gene are underlined in red in Panel A and B respectively. Panel A, the *E.coli glyA*, shows that the gene is not next to any other gene going in the same direction. Because there are no genes that are going in the same direction as our gene, we can conclude that *E.coli glyA* is not a part of an operon. This is the same result for *Mrub_2910.* The *Mrub_2910* gene is also not surrounded by any genes that are going in the same direction as it. Therefore, *Mrub_2910* is not a part of an operon. The similar evidence for the two genes is another confirmation that the two genes are orthologs.

Panel A



Panel B

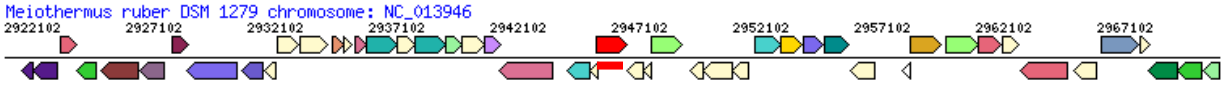
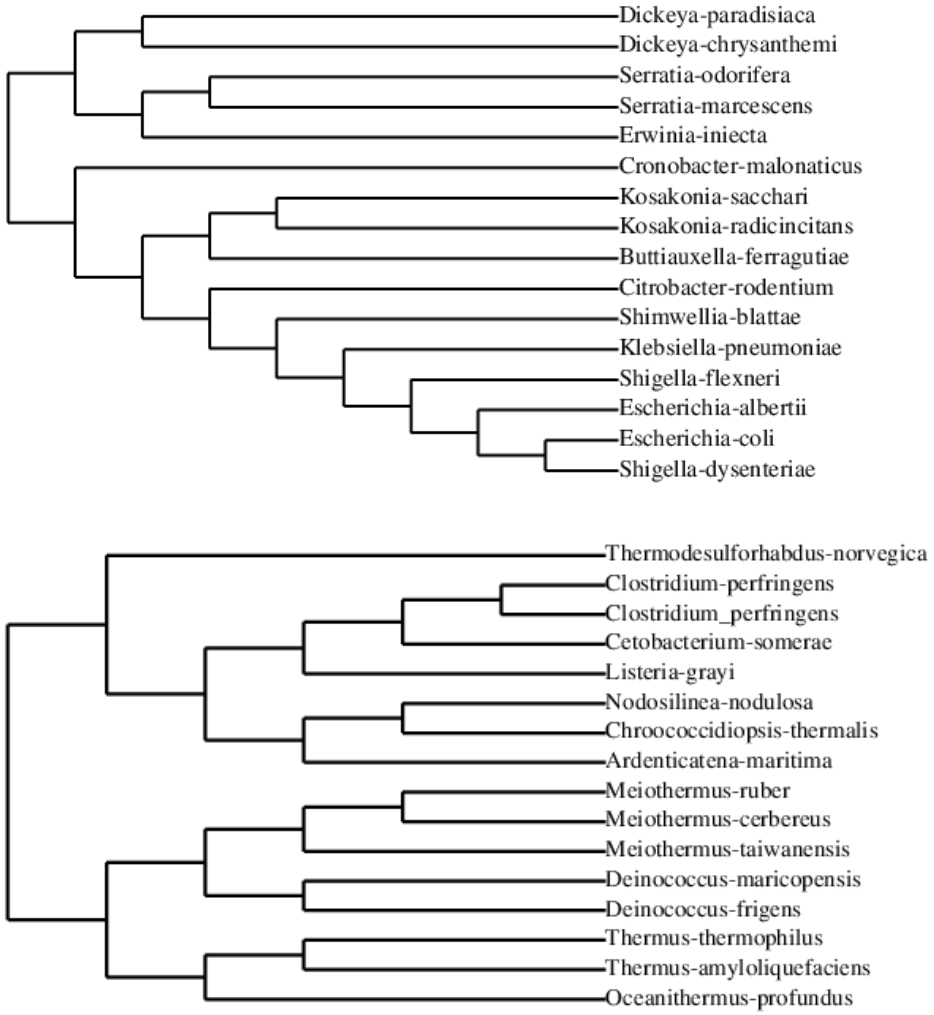


Figure 21. *E.coli glyA* and *Mrub_2910* genes are not a part of an operon. Panel A shows the *E.coli b2551/glyA* Chromosome Viewer; Panel B is showing the *Mrub_2910* Chromosome Viewer. Red lines indicate the location of each gene on their respective viewer. Neighborhood region images from IMG/EDU at https://img.jgi.doe.gov/.

Again, Phylogeny.fr provides evidence for or against horizontal gene transfer (HGT) being an option in for these particular genes. Figure 22 illustrates the *E.coli glyA* tree (Panel A) and shows that all of the nearby species are from the phylum Proteobacteria. This means that it is unlikely that HGT occurred for this gene. Panel B in figure 22 shows the *Mrub_2910* tree. It indicates that the proteins most closely related to the protein encoded by *Mrub_2910* are from the Deinococcus-Thermus and Cholorflexi phylum. There are other species from the Cyanobacter, Fusobacteria, Proteobacteria, and Firmicutes phylum on the map. These species aren't directly surrounding the *M. ruber* species. While it is not very likely that HGT occurred for the *Mrub_2910* gene, because the Cholorflexi species is close to the *M. ruber* species, it is still a possibility.



Panel A

Panel B

Figure 22. Horizontal gene transfer is not likely because each query is surrounded by numerous species within its respective phylum. Panel A is the phylogenetic tree for *E.coli b2551/glyA*. Panel B is the phylogenetic tree for *Mrub_2910*. Multiple sequence alignments from T-Coffee were used for each gene to build the respective phylogenetic trees. The phylogenetic trees are from Phylogeny.fr at http://www.phylogeny.fr/

Conclusion

Results from this bioinformatics study indicate that *Mrub_0173* and *serA* (Table 1), *Mrub_0125* and *serB* (Table 2), and *Mrub_2910* and *glyA* (Table 3) show similarities in sequence, cellular location, and structure. Evidence of these orthologous pairs was first indicated by BLAST analyses that compared the amino acid sequences of each pair that showed low E-values. Further support of the results came from TIGRfam and Pfam matching the protein sequences to the correct proteins and domains. The programs matched *Mrub_0173* and *E.coli b2913* to phosphoglycerate dehydrogenase, *Mrub_0125* and *E.coli b4388* to phosphoserine phosphatase, and *Mrub_2910* and *E.coli b2551* to serine hydroxymethyltranserase. More confirmation came from cellular location bioinformatics tools THM, SignalP, LipoP, and PSORT-B. For each pair, the programs revealed that the proteins for each gene are in the cytoplasm. They phylogenetic trees for all of the *M. ruber* genes are consistent with the traditional phylogenetic tree because the surrounding phyla are from the Deinococcus-Thermus phylum. Other bioinformatics tools used for the program displayed the same results for each gene pair. There was no refuting data for any of our gene pairs. Based on the consistency at which the bioinformatics data matched up between each gene pair, we can conclude that *Mrub_0173* and *E.coli b2913 (serA), Mrub_0125* and *E.coli b4388 (serB),* and *Mrub_2910* and *E.coli b2551 (glyA)* are all orthologous pairs.

If we were to study the *Mrub_0125* gene, putative *serB*, we would attempt to do a site-directed mutagenesis. We would choose to change the highly conserved glycine at position 178 to an alanine (Figure 23). According to Betts and Russell (2003), glycine plays a distinct functional role. Changing a conserved glycine to any other amino acid could result in a big change in function of the protein. We would change the GGC codon to GCC, turning the glycine to alanine. The forward primer for this mutagenesis would be Q5SDM_2/8/2017_F and would have the sequence AGAGCCTATGcCGACAGCCTG. The reverse primer for the mutagenesis would be Q5SDM_2/8/2017_R and the sequence would be ATAAAGCACCTCGCCATC. Figure 24 shows the outcome of the NEBaseChanger to determine the primers for this experiment (NEBaseChanger). The recommended annealing temperature for both would be 63°C. We would use these primers to change the glycine to alanine and then observe any changes in serine synthesis.

Figure 23. Glycine at spots 178 and Aspartic Acid at 179 are highly conserved for *Mrub_0125*. HMM logo was made by Pfam at http://pfam.sanger.ac.uk/search.

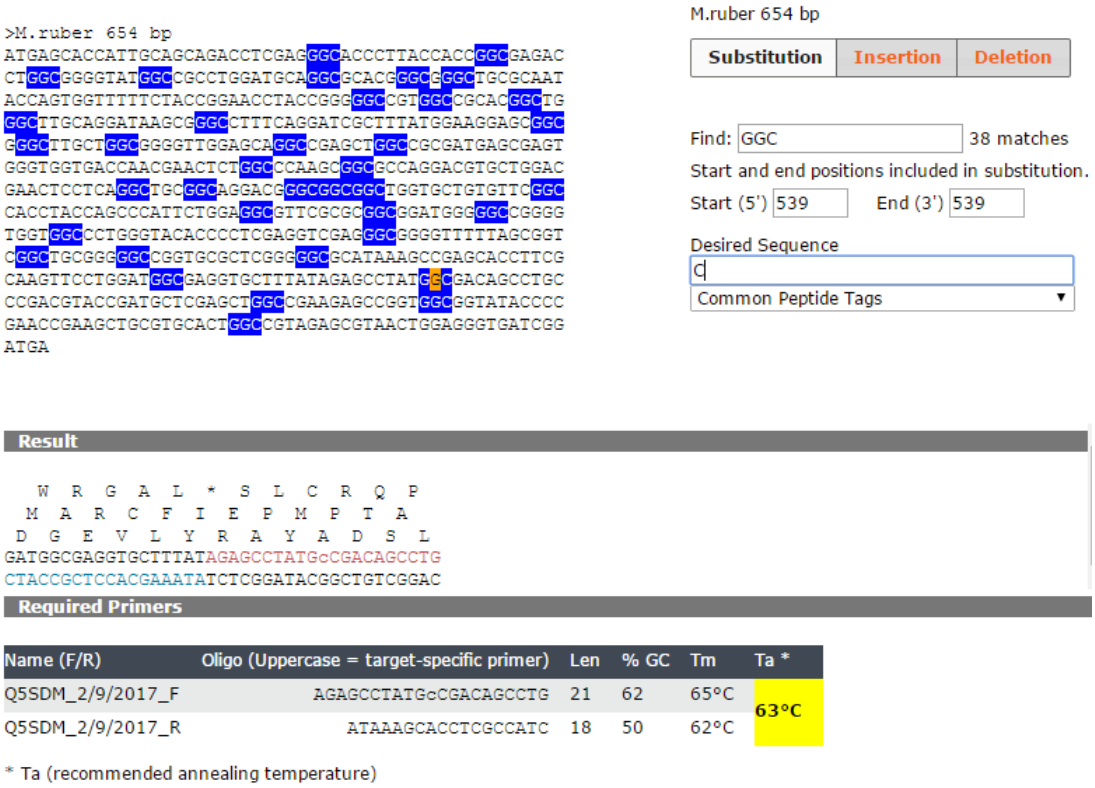


Figure 24. One nucleotide, G, at position 539 identified to be changed to a C nucleotide in order to change glycine to alanine. Forward primer: Q5SDM_2/8/2017_F with oligo AGAGCCTATGcCGACAGCCTG. Reverse: Q5SDM_2/8/2017_R with the olgio ATAAAGCACCTCGCCATC. Primers determined by NEBaseChanger that is available at http://nebasechanger.neb.com/.

Literature Cited

Arora G, Tiwari P, Mandal RS, Gupta A, Sharma D, Saha S, Singh R. High Throughput Screen Identifies Small Molecule Inhibitors Specific for Mycobacterium tuberculosis Phosphoserine Phosphatase. Journal of Biological Chemistry. 2014 [accessed 2017 Feb 5];289(36):25149–25165.

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 M. J., and Russell, R. B. (2003) In Bioinformatics for Geneticists (Barnes, M. R., and Gray, I. C., Eds.). [accessed 2017 Feb 5]; 289–316, Wiley, England.

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/

Dereeper A., Guignon V., Blanc G., Audic S., Buffet S., Chevenet F., Dufayard J.F., Guindon S., Lefort V., Lescot M., Claverie J.M., Gascuel O. 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res, 1;36 (Web Server issue):W465-9. [accessed 2016 Dec 6]. Available at http://phylogeny.fr

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

Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O. 2001. TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res 29(1):41-3. [accessed 2016 Dec 6]. http://blast.jcvi.org/web-hmm/

Jacob J, Duclohier H, Cafiso, DS. 1999. The role of proline and glycine in determining the backbone flexibility of a channel-forming peptide. Biophys J. [accessed 2017 Feb 5]; 76: 1367–1376.

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

Kanehisa M, Sato Y, 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. EcoCyc: fusing model organism databases with systems biology *Nucleic Acids Research* 41:D605-612. [Accessed 2016 Dec 6]. Available from: https://ecocyc.org/

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

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. [Accessed 2016 Dec 6]. Available from: http://nar.oxfordjournals.org/content/40/D1/D115.full

Metabocard for Glycine (HMDB00123). Human Metabolome Database. [accessed 2017 Feb 5]. Available at: http://www.hmdb.ca/metabolites/HMDB00123

Metabocard for L-Serine (HMDB00187). Human Metabolome Database. [accessed 2017 Feb 5]. Available at: http://www.hmdb.ca/metabolites/HMDB00187

NCBI BLAST [Internet] National Center for Biotechnology Information; [accessed 2017 Feb 5]; Available from http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=FAQ#expect

NEBaseChanger. New England BioLabs Inc. [accessed 2017 Feb 7]. 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 [accessed 2016 Dec 6]. Available from: http://www.ebi.ac.uk/Tools/msa/tcoffee/

Persidis A. 1999. Bioinformatics. Nature Biotechnology. [accessed 2017 Feb 5]; 17(8): 828-830.

Petersen T.N, S, Brunak, G. von Heijne, H. Nielsen. 2011. Discriminating signal peptides from transmembrane regions. Nature Methods, 8:785-786. [accessed 2016 Dec 6]. Available from: http://www.cbs.dtu.dk/services/SignalP

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.

Scott LR. Meiothermus ruber Genome Analysis Project. [Internet]. GENI-ACT; [accessed 2017 Feb 5]; Available from: http://geni-science.org/secure/projects/view/

Tindall et al. 2010. Complete genome sequence of Meiothermus ruber type strain. Stand Genomic Sci [accessed 2017 Feb 5]; 3(1): 26-36.

Tom KJ, Snell K, Duran M, Berger R, Poll-The B-T, Surtees R. L-Serine in Disease and Development. Biochemical Journal. 2003 [accessed 2017 Feb 5];371(3):653–661.

Yu N.Y., 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**.** [Accessed 2016 Dec 6]. Available from: http://www.psort.org/psortb/