2-2016

Bioinformatics indicates that *Meiothermus ruber* genes Mrub_1710 and Mrub_1712 are homologs of the *Escherichia coli* genes b2903 (P-protein; glycine dehydrogenase) and b2905 (T-protein; aminomethyltransferase), respectively

Malory J. Groen
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Bioinformatics indicates that *Meiothermus ruber* genes Mrub_1710 and Mrub_1712 are homologs of the *Escherichia coli* genes b2903 (P-protein; glycine dehydrogenase) and b2905 (T-protein; aminomethyltransferase), respectively

Malory Groen

Introduction

Glycine is the simplest amino acid and is composed of only two carbons and one amino group (Kikuchi et al. 2008). The major pathway for the catabolism of glycine is a reversible reaction catalyzed by the glycine cleavage system (Kikuchi et al. 2008). The *Escherichia coli* gene b2903, located from base-pair 3049041 to 3046168, encodes the enzyme glycine dehydrogenase, also known as P-protein, which contains 957 amino acids. And, the *Escherichia coli* gene b2905, located from base-pair 3050667 to 3049573, encodes the enzyme aminomethyltransferase, also known as T-protein, which contains 364 amino acids. These two enzymes, along with another enzyme (L-protein) and a carrier protein (H-protein), make the glycine cleavage system (Motokawa and Kikuchi 1974). Together, these four proteins form a multi-enzyme complex (Okamura-Ikeda et al. 1993). The glycine cleavage reaction, which is completely reversible, is partitioned into three partial reactions (Kikuchi et al. 2008). The first partial reaction of glycine cleavage is catalyzed by P-protein, with H-protein serving as a co-substrate (Kikuchi et al. 2008). In this partial reaction, the carboxyl carbon of glycine is converted to carbon dioxide while the remnant of the glycine molecule is transferred to one of the sulfhydryl groups formed by the reductive cleavage of disulfide in lipoate attached to H-protein (Fujiwara and Motokawa 1983). This H-protein, which is involved in all partial reactions of glycine cleavage, has a lipoyllysine arm that shuttles the reaction intermediate and reducing equivalents between the active sites of the components of the glycine cleavage system (Kikuchi et al. 2008). The second partial reaction of glycine cleavage is catalyzed by T-protein. This reaction requires H_4folate and yields ammonia, N^5,N10-methylene-H_4folate, and H-protein with a reduced lipoate (Fujiwara et al. 1984). The third and final partial reaction of glycine cleavage is catalyzed by L-protein. In this reaction, the reduced lipoate attached to H-

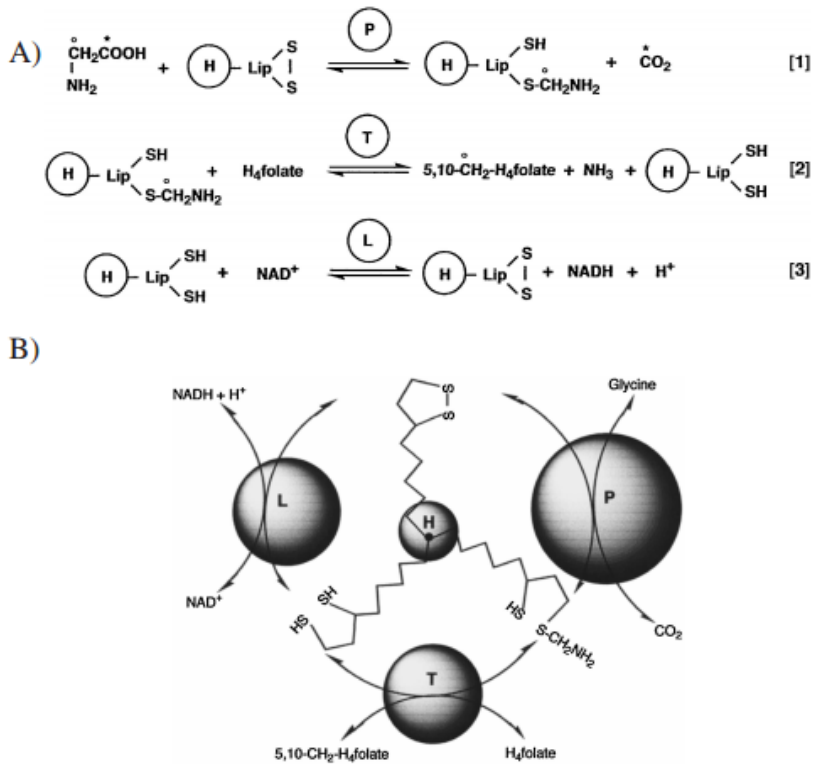


Figure 1. A) Mechanism of the glycine cleavage reaction. B) Schematic presentation of the glycine cleavage reaction. P, H, T and L in the circles represent respective proteins. Lip, H4folate, 5,10-CH2-H4folate represent lipoyl moiety, tetrahydrofolate, and N5,N10-methylene-H4folate, respectively (Kikuchi et al. 2008).

protein is reoxidated by the enzymatic function of L-protein, which regenerates the original H-protein enzyme (Kikuchi et al. 2008). It is important to note that the L-protein is not a glycine cleavage reaction-specific enzyme and is not a part of the *Escherichia coli* gcv operon, which encodes for the glycine cleavage system enzymes P-protein, H-protein, and T-protein (Okamura-Ikeda et al. 1993). An overview of this reaction can be seen in Figure 1 (Kikuchi et al. 2008).

Functional evidence supports that the b2903 and b2905 genes encode the glycine dehydrogenase and aminomethyltransferase proteins, respectively, in *E. coli* which are key components of the glycine degradation system complex. In the experiment that confirmed the functional evidence of the b2903 and b2905 genes, *E. coli* BL21 (DE3)pLysS cells were lysed and affinity chromatography was used to separate the proteins in *E. coli* (Okamura-Ikeda et al. 1993). The T-protein was eluted off of the column a 0.16-0.18 M NaCl while the P-protein was eluted off of the column at 0.20-0.21 M NaCl. The fractions for each of these proteins were pooled and concentrated. The amino acid sequences of these proteins were determined in an Applied Biosystems 477 sequencer and used to design oligonucleotide probes for gene cloning. The nucleotide sequence of a fragment of *E. coli* DNA containing the genes for the components of the glycine cleavage system was determined. Then, a plasmid containing the glycine cleavage system operon (gcv operon) was transfected into *E. coli* cells, and as a result, the gene products P-protein and T-protein were overexpressed. This directly links the b2903 and b2905 genes to their protein products, P-protein and T-protein, respectively. The activity of P-protein was assayed by the glycine/$^{14}CO_2$ exchange reaction. This confirmed that P-protein, along with the assistance of H-protein as a carrier molecule, did indeed convert the carboxyl carbon of glycine to carbon dioxide. The activity of T-protein, along with the assistance of H-protein as a carrier molecule, was assayed by measuring the conversion of N^5,N10-methylene-H_4folate to H_4folate by measuring the decrease in absorbance at 290 nm (Okamura-Ikeda, Fujiwara, Motokawa 1987). The lipoyl prosthetic group of H-protein employed in the experiment was previously reduced since the intermediate formation catalyzed by T-protein requires the reduced H-protein. While this measures the occurrence of the reverse reaction, it is applicable to glycine degradation because it is a completely reversible process, as mentioned earlier. So, this assay confirmed that T-protein can functionally displace the amino group from glycine, forming N^5,N10-methylene-H_4folate.

Studying glycine cleavage is particularly important because its cleavage produces a readily usable form of energy in the form of NADH and because one of the products of glycine degradation, N^5,N10-methylene-H_4folate, is used for the biosynthesis of various cellular substances, such as purines, thymidylate, and methionine (Kikuchi et al. 2008). However, the study of glycine degradation is particularly important when studying *Meiothermus ruber* because the percentage of glycine in a particular protein plays a role in that protein's stability. A substitution of glycine into a protein increases the distance between the alpha carbon and the amino group of the amino acid residue (Jacob et al. 1999). Because this bond length is increased, the steric hindrance is decreased and the overall flexibility of the bond is increased. So, a higher percentage of glycine in a protein would increase the overall flexibility of that protein. In the reverse perspective, the removal or replacement of glycine in a protein domain dramatically reduces the high-amplitude hinge motion of this helix (Jacob et al. 1999). So, a lower percentage of glycine in a protein would decrease the overall flexibility of that protein and increase its rigidity. This concept of glycine ratios is principally important when studying *Meiothermus ruber* because it is a thermophilic bacterium. These bacteria grow at temperatures around 60^{O}C (Garrity et al. 2001). Because of this, their proteins need to be more rigid, or stable, in order to

retain their structure and function at higher temperatures. And, more thermodynamically stable proteins can be made by decreasing the overall ratio of glycine within that protein. So, glycine cleavage and the enzyme associated with this process are particularly important when studying *M. ruber*.

In order to further study P-protein and T-protein, and with regards to the Genomic Encyclopedia of Bacteria and Archaea (GEBA) Project, the *Meiothermus ruber* Mrub_1710 and Mrub_1712 genes and the *Escherichia coli* b2903 and b2905 genes were studied with respect to one another. Because Mrub_1710 and Mrub_1712 have been predicted to serve the same functions and be homologous to b2903 and b2905, respectively, it may be of particular use when studying the cleavage of glycine in *M. ruber*. The GEBA project began as a collaboration with the Joint Genome Institute (JGI), which is a division of the US Department of Energy (Phylogenetic …2015). The goal of the GEBA project is to sequence thousands of bacterial and archaeal genomes from many different branches of the Tree of Life, and specifically, the genomes of obscure microbes that have the prospect to provide insight into biologically important processes (Wu *et al* 2009). Very little is currently known about the species within the Deinococcus-Thermus phylum, which contains the organism *Meiothermus ruber*, so studying this organism can provide some insights into this understudied branch of the Tree of Life, and possibly enhance our understanding of the evolution, physiology, and metabolic capacity of microbes (Phylogenetic …2015).

One of the objectives of the GEBA project is to utilize different bioinformatics programs in the research. Bioinformatics is already well-developed with a lot of different data contained in different databases and it is anticipated that someday, bioinformatics will affect everyone and play an even larger role in scientific research. For this project, bioinformatics systems were used to determine if the *Meiothermus ruber* genes Mrub_1710 and Mrub_1712 were homologs of the *Escherichia coli* genes b2903 (P-protein; glycine dehydrogenase) and b2905 (T-protein; aminomethyltransferase), respectively, by determining their structural, and therefore functional, similarity. *E. coli* is a very well-studied organism with a lot of data to compare to our genes of interest, Mrub_1710 and Mrub_1712. The BLAST program was one bioinformatics tool that was used to determine the sequence similarity and between Mrub_1710 and b2903 and between Mrub_1712 and b2905. Because the amino acid sequence, or primary structure, of a protein determines its subsequent tertiary structure and function, the sequence similarities are good indications of whether or not two proteins can be homologous. The BLAST program can align a sequence of interest with many different parameters to find similar genes in different organisms. And, other selected organisms can be used to compare to the genes of interest, which in the case of this research were Mrub_1710 and Mrub_1712. An alignment between the b2903 and Mrub_1710 sequences and an alignment between the b2905 and Mrub_1712 sequences are what ultimately generated the E-values associated with the alignment of the two sets of gene sequences. This E-value is indicative high sequence similarity that is unlikely due to chance if the value is low or zero.

So, the purpose of this project was to use a variety of bioinformatics tools to gather information about b2903, b2905, Mrub_1710 and Mrub_1712. With that information, it was then possible to compare the two sets of genes from the different organisms. These aspects of the research facilitated the overarching goal of the project, which was to compare the two sets of genes from *E. coli* and *M. ruber* in the hopes of determining the extent of their structural similarity and possible homology. So, this research was conducted with regards to the hypothesis that the *Meiothermus ruber* gene Mrub_1710 is a homolog of the *Escherichia coli* gene b2903

(P-protein; glycine dehydrogenase) and that the *Meiothermus ruber* gene Mrub_1712 is a homolog of the *Escherichia coli* gene b2905 (T-protein; aminomethyltransferase).

Methods

Bioinformatics programs within the GENI-ACT lab notebook were used. A listing and short description of all programs used in GENI-ACT can be found at http://www.geni-act.org/education/main/. This research deviated slightly from the GENI-ACT protocol in the following ways: the sequences used for the alignment of one the two lists of top ten hits for *E. coli* in the T-Coffee program excluded the use of multiple strains of *E. coli* and other organisms in the *Escherichia* genus from the list of sequences; MetaCyc was used for *M. ruber* instead of EcoCyc, which was used for *E. coli*; and *E. coli* was blasted against *M. ruber*, which is not a step outlined in the GENI-ACT lab notebook.

Results

P-protein: (b2903 and Mrub_1710)

Sequence-Based Similarity Data

Sequence-based similarity data indicate that b2903 and Mrub_170 have very high structural similarity. A protein BLAST comparison between b2903 and Mrub_1710 produced an E-value of 0.0 (Table 1, Figure 2). This indicates that the two sequences are highly similar and that there is very little chance that the two sequences are related solely by chance. A protein BLAST was also performed on both b2903 and Mrub_1710 separately. Two of the top hits for b2903 were sequences from *Trichuri trichiura* (1988 bits, E-value=0.0) and *Escherichia fergusonii* (1983 bits, E-value=0.0), both of which are predicted to be glycine dehydrogenase proteins (Table 1). Two of the top hits for Mrub_1710 were *Meiothermus cerbereus* (1751 bits, E-value=0.0) and *Meiothermus rufus* (1554 bits, E-value=0.0), both of which are predicted glycine dehydrogenase proteins as well (Table 1). This is important because both proteins, independently of one another, are characterized by being related to glycine dehydrogenase proteins. A conserved domain database (CDD) search assigned the same cluster of orthologous groups (COG) to both b2903 and Mrub_1710. The protein sequences were labeled as COG1003, or glycine cleavage system protein P (pyridoxal-binding), C-terminal domain, and COG0403, or glycine cleavage system protein P (pyridoxal-binding), N-terminal domain (Table 1). The E-values for b2903 and Mrub_1710 were 0e+00 for both COG categories. This result is significant and indicates that b2903 and Mrub_1710 likely have the same or similar function that validates their structural similarity. So, all three databases producing the sequence-based similarity data indicate that b2903 and Mrub_1710 are structurally similar and likely perform a similar function.

Table 1: *E. coli* b2903 and Mrub_1710 are orthologs

Description of Evidence	***E. coli* (b2903)**	***M. ruber* (Mrub_1710)**
BLAST *E. coli* against *M. ruber*	Score: 1098 bits; E-value: 0.0	
Protein BLAST	*Trichuris trichiura: 1988 bits, E-value: 0.0 (glycine dehydrogenase)*	*Meiothermus cerbereus: 1751 bits, E-value: 0.0 (glycine dehydrogenase)*
	Escherichia fergusonii: 1983 bits, E-value: 0.0 (glycine dehydrogenase)	*Meiothermus rufus: 1554 bits, E-value: 0.0 (glycine dehydrogenase)*
CDD (COG) Category	COG1003 Glycine cleavage system protein P (pyridoxal-binding), C-terminal domain	
	E=0e+00	E=0e+00
	COG0403 Glycine cleavage system protein P (pyridoxal-binding), N-terminal domain	
	E=0e+00	E=0e+00
Cellular Localization	Cytoplasmic	
Pfam: Protein Family	PF02347, CL0061, glycine cleavage system P-protein	
	E=5.7e-194	E=2.4e-183
TIGRfam: Protein Family	TIGR00461 gcvP: glycine dehydrogenase	
	E=0	E=0
PDB	4LGL Glycine decarboxylase P-protein	
	E=0.0	E=0.0
E.C. Number	E.C.1.4.4.2. Glycine dehydrogenase	
KEGG Pathway	Glycine, Serine, and Threonine Metabolism	

Score	Expect	Method	Identities	Positives	Gaps
1098 bits(2839)	0.0	Compositional matrix adjust.	583/950(61%)	701/950(73%)	15/950(1%)

```
Query  11   SGAFIERHIGPDAAQQQEMLNAVGAQSLNALTGQIVPKDIQLATPPQVGAPATEYAALAE  70
            S  F  RHIGP  A  ++ML AVG  SL  L  Q VP  I+ A P  +G   +E   LA
Sbjct  9    STDFSRRHIGPTPADIEQMLQAVGVSSLEELIQQTVPASIREAEPLNIGPGLSETEMLAR  68

Query  71   LKAIASRNKRFTSYIGMGYTAVQLPPVILRNMLENPGWYTAYTPYQPEVSQGRLEALLNF  130
            ++AIAS+N+ FTS IG GY    LPPVI RN+LENP WYTAYTPYQPE+SQGRLEALLNF
Sbjct  69   MRAIASKNQVFTSLIGQGYYGTILPPVIQRNLLENPAWYTAYTPYQPEISQGRLEALLNF  128

Query  131  QQVTLDLTGLDMASASLLDEATAAAEAMAMAKRVSKLKNANRFFVASDVHPQTLDVVRTR  190
            Q +  DLTGLD+A+ASLLDEATAAAEAMA+A+R  K      FFV  +VHPQTL V++TR
Sbjct  129  QTMVADLTGLDIANASLLDEATAAAEAMAVAQRSCKAPR-TAFFVDRNVHPQTLAVLQTR  187

Query  191  AETFGFEVIVDDAQKVLDHQDVFGVLLQQVGTTGEIHDYTALISELKSRKIVVSVAADIM  250
            AE  G++++V D +  L+ + +FG + Q   T G+I D    I  +K+   +  VAAD++
Sbjct  188  AEPLGWQLVVGDPETDLNPEGLFGAIFQYPDTLGQIRDLRGPIERVKAAGGLAVVAADLL  247

Query  251  ALVLLTAPGKQGADIVFGSAQRFGVPMGYGGPHAAFFAAKDEYKRSMPGRIIGVSKDAAG  310
            AL LLT PG+ GADI  GSAQRFGVPMGYGGPHA + A +D  KR++PGR++GVS D+ G
Sbjct  248  ALTLLTPPGELGADIAVGSAQRFGVPMGYGGPHAGYMAIRDALKRALPGRLVGVSIDSRG  307

Query  311  NTALRMAMQTREQHIRREKANSNICTSQVLLANIASLYAVYHGPVGLKRIANRIHRLTDI  370
            N A R+ +QTREQHIRREKA SNICT+QVLLA IAS+YAVYHGP GL+ IA  +H  T +
Sbjct  308  NRAYRLTLQTREQHIRREKATSNICTAQVLLAVIASMYAVYHGPEGLREIAQSVHDKTAV  367

Query  371  LAAGLQQKGLKLRHAHYFDTLCVEVADKAG-VLTRAEAAEINLRSDILNAVGITLDETTT  429
            LAAGL++ G +  +AH+FDTL +    +   +L  A    INLR      +GI LDETTT
Sbjct  368  LAAGLKRLGFERVNAHFFDTLTIRAEGRVEEILAEARRRRINLRKVDEAHIGIALDETTT  427

Query  430  RENVMQLFNVLLGDNHGLDIDTLDKDVAHDSRSIQPAMLRDDEILTHPVFNRYHSETEMM  489
             E +  ++    GD    D    +         +  A+ R    LTHPVFNRYHSETE+M
Sbjct  428  PEVIEAVWQAFGGDFRYQDFVPENH--------LPAALQRTSPYLTHPVFNRYHSETELM  479

Query  490  RYMHSLERKDLALNQAMIPLGSCTMKLNAAAEMIPITWPEFAELHPFCPPEQAEGYQQMI  549
            RYM  L  +DLAL++AMIPLGSCTMKLNAAAEMIPI+WPEFA LHPF P EQA+GY ++
Sbjct  480  RYMRRLADRDLALDRAMIPLGSCTMKLNAAAEMIPISWPEFANLHPFAPAEQAQGYHELF  539

Query  550  AQLADWLVKLTGYDAVCMQPNSGAQGEYAGLLAIRHYHESRNEGHRDICLIPASAHGTNP  609
              L+ WL ++TGYDAV +QPNSGAQGEYAGLLAIR YH SR EGHR++CLIP+SAHGTNP
Sbjct  540  ETLSRWLCEITGYDAVSLQPNSGAQGEYAGLLAIRAYHRSRGEGHRNVCLIPSSAHGTNP  599

Query  610  ASAHMAGMQVVVVACDKNGNIDLTDLRAKAEQAGDNLSCIMVTYPSTHGVYEETIREVCE  669
            ASAHMAGM+VVVVACD  G +DL DL AKA+Q  D L+ +MVTYPSTHGV+EE IRE+CE
Sbjct  600  ASAHMAGMEVVVVACDAQGYVDLNDLEAKAKQHADRLAAVMVTYPSTHGVFEEKIRELCE  659

Query  670  VVHQFGGQVYLDGANMNAQVGITSPGFIGADVSHLNLHKTFCIPHGGGGPGMGPIGVKAH  729
            +VH++GGQVYLDGAN+NAQVG+  PG  GADVSHLNLHKTF IPHGGGGPGMGPI VKAH
Sbjct  660  IVHRYGGQVYLDGANLNAQVGLAKPGHYGADVSHLNLHKTFAIPHGGGGPGMGPIAVKAH  719

Query  730  LAPFVPGHSVVQIEGMLTRQGAVSAAPFGSASILPISWMYIRMMGAEGLKKASQVAILNA  789
            LAPF+PGH V  ++G     G V+AAP+GSASILPIS+ YI MMG+EGLK+A++VAILNA
Sbjct  720  LAPFLPGHPV--LDGGTAPVGPVAAAPYGSASILPISFAYIWMMGSEGLKRATEVAILNA  777

Query  790  NYIASRLQDAFPVLYTGRDGRVAHECILDIRPLKEETGISELDIAKRLIDYGFHAPTMSF  849
            NYIASRL+  FPVLY    GRVAHECILD R  K    I+  DIAKRLID+GFHAPTMSF
Sbjct  778  NYIASRLEPYFPVLYKNEKGRVAHECILDPRAFKARCDITAEDIAKRLIDFGFHAPTMSF  837

Query  850  PVAGTLMVEPTESESKVELDRFIDAMLAIRAEIDQVKAGVWPLEDNPLVNAPHIQSELVA  909
            PVAGTLMVEPTESESK ELDRFI+AM+AIR EI Q++ G    E++PL +APH   +L
Sbjct  838  PVAGTLMVEPTESESKHELDRFINAMIAIRQEIAQIERGELTAEESPLRHAPHTVLDLAD  897

Query  910  E-WAHPYSREVAVFPAGVA--DKYWPTVKRLDDVYGDRNLFCSCVPISEY  956
            E W   Y+R    FPAG +  DKYW  V R+D+VYGDRN+ CSC P+ EY
Sbjct  898  ENWNRKYTRAQGCFPAGQSAMDKYWSPVNRVDNVYGDRNIVCSCPPVEEY  947
```

Figure 2. A protein BLAST between Mrub_1710 and b2903 indicates a very low E-value—0.0—which is indicative of high sequence similarity that is not likely due solely to chance. NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) generated this protein BLAST.

Localization Data

All of the localization data indicates that the b2903 gene and the Mrub_1710 gene sequences code for cytoplasmic proteins. TMHMM predicted 0 transmembrane helices for both gene products, as indicated by the transmembrane topology graphs depicted in Figure 3. With no transmembrane helices, it is most likely that the gene product would be found in the cytoplasm, and not inserted into a membrane. SignalP determined signal peptide probabilities for both *E. coli* b2903 and Mrub_1710 to be 0.094 and 0.139, respectively. These probabilities are very low, which indicates that there is likely no signal peptide coded for in the gene sequences. Because of this, it is unlikely that there is a signal peptide attached to these genes that would indicate their migration to a different part of the cell. So, these proteins are likely made in the cytoplasm and then remain there to fulfill their roles. The SignalP prediction graphs are represented below in Figure 4. PSORT-B also indicated that both b2903 and Mrub_1710 are cytoplasmic proteins as well, as described in Table 2. Along with this program, LipoP suggested that the best prediction for cellular location of both was in the cytoplasm. While the posterior probability graphs generated by Phobius, as seen in Figure 6, indicate some small regions of hydrophobicity, these hydrophobic regions are likely used for the attachment of the multi-enzyme glycine dehydrogenase system complex. Therefore, these small hydrophobic regions likely do not correlate to transmembrane regions. So, all five of these programs corroborate one another and indicate that both b2903 and Mrub_1710 are cytoplasmic proteins.

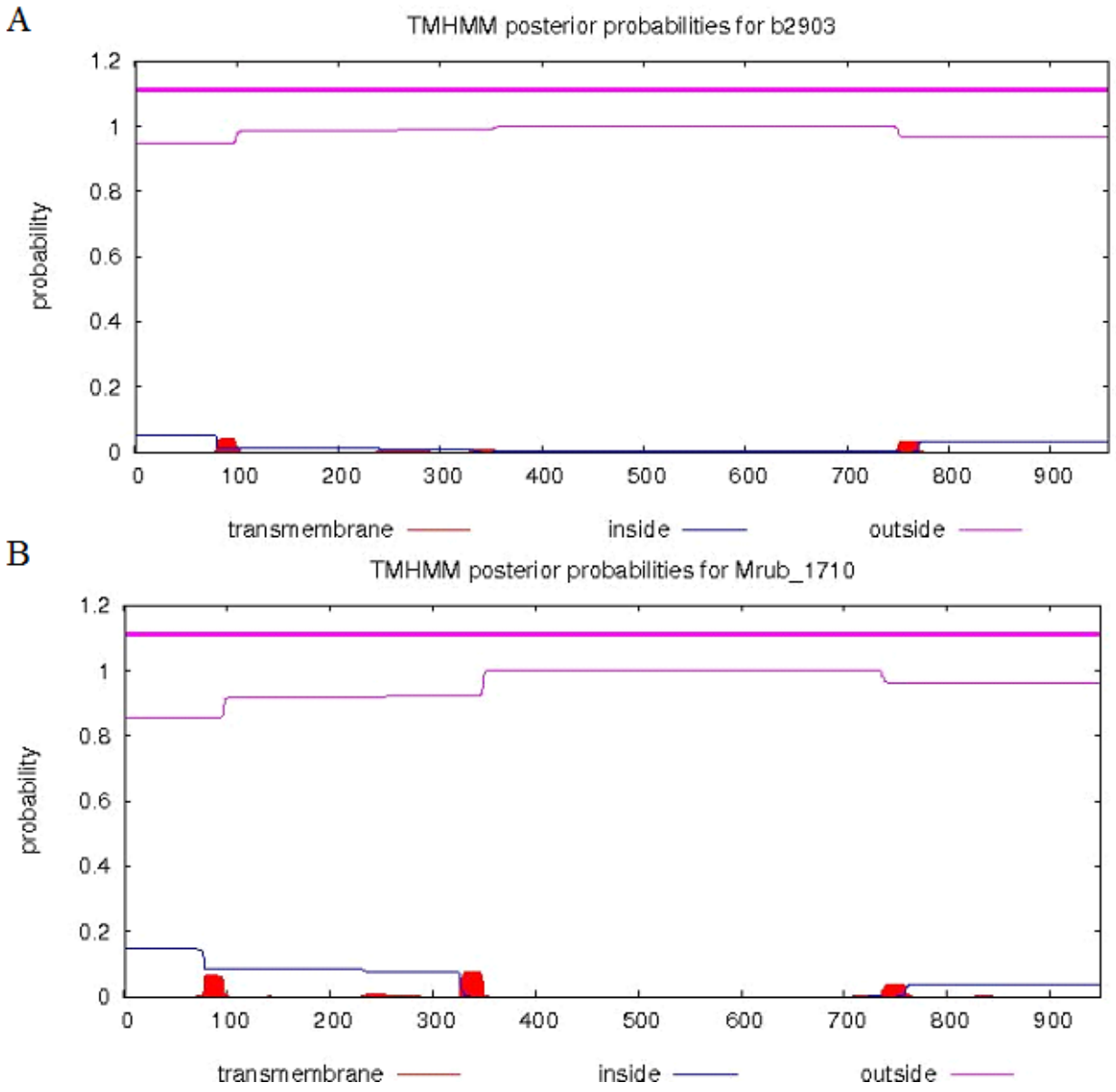


Figure 3. b2903 and Mrub_1710 do not contain TMH regions; a cytoplasmic location is predicted. Server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) created this hydrophobicity plot depicting the TMHMM posterior probabilities for A: b2903 and B: Mrub_1710.

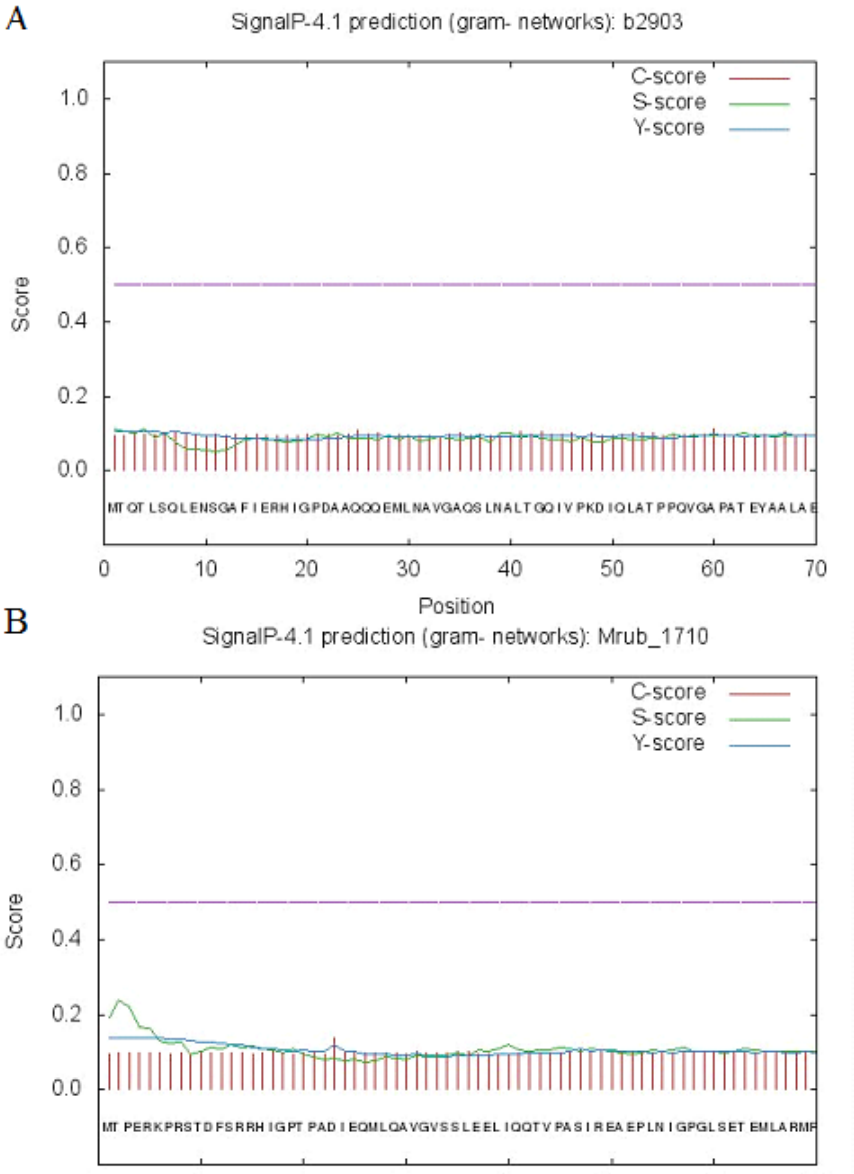


Figure 4. b2903 and Mrub_1710 do not contain regions indicating the presence of a signal peptide; a cytoplasmic cellular location is predicted. SignalP 4.1 server (http://www.cbs.dtu.dk/services/SignalP) created the SignalP-4.1 prediction graphs for A: b2903 and B: Mrub_1710.

Table 2: *E. coli* b2903 and Mrub_1710 are cytoplasmic proteins

Cellular Location	*E. coli* b2903	*M. ruber* Mrub_1710
Cytoplasm	9.97	9.97
Cytoplasmic Membrane	0.01	0.01
Periplasm	0.01	0.01
Outer Membrane	0.00	0.00
Extracellular	0.00	0.00

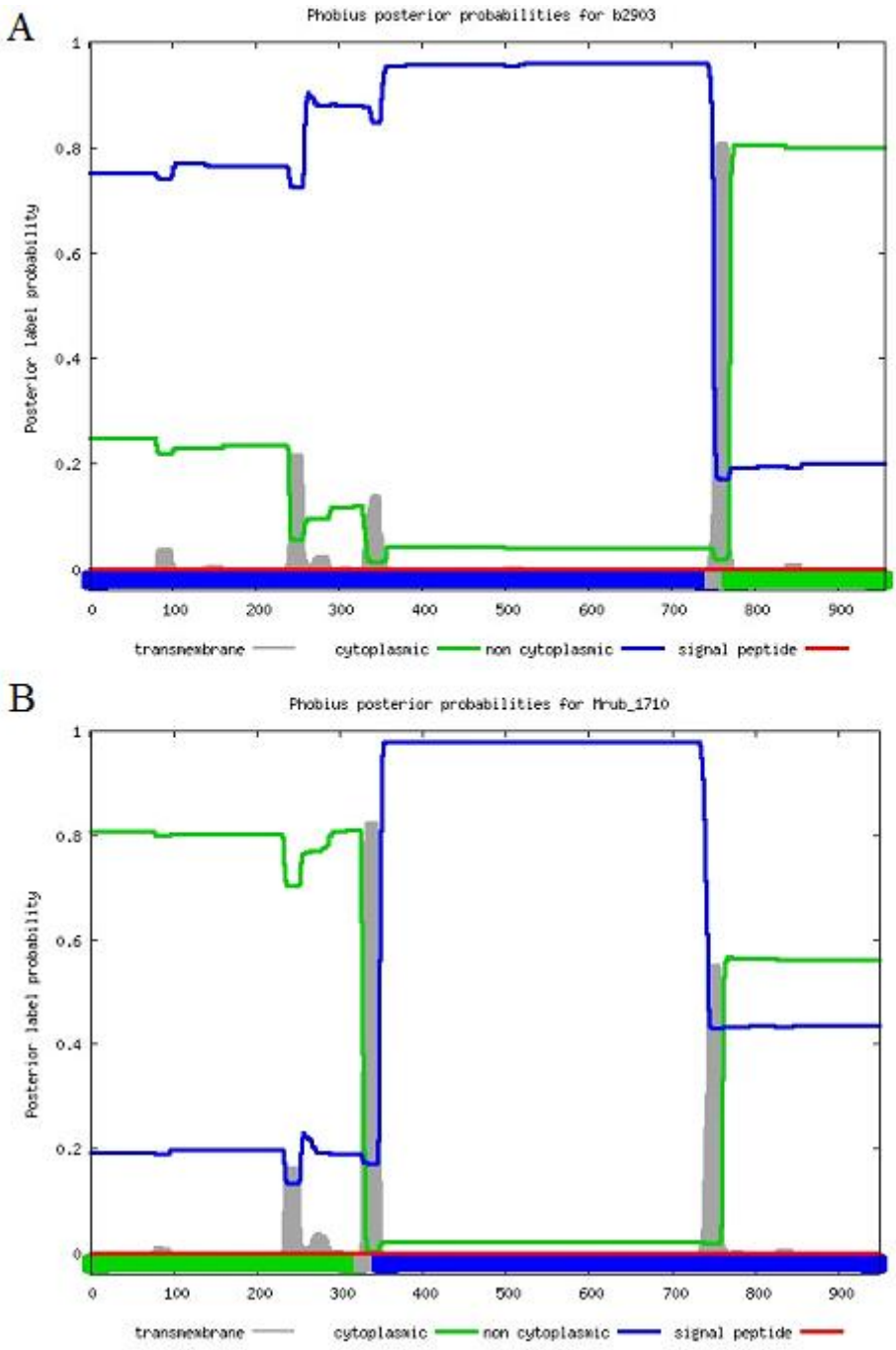


Figure 5. b2903 and Mrub_1710 do not contain large hydrophobic regions correlating to transmembrane regions—the small hydrophobic regions likely correlate to the regions off attachment of the multi-enzyme glycine cleavage system complex; a cytoplasmic location is predicted. Phobius (http://phobius.sbc.su.se/) created the posterior probability graphs for A: b2903 and B: Mrub_1710.

Structure-Based Evidence

Pfam indicates that the sequences from both b2903 and Mrub_1710 correlate to PF02347, or glycine cleavage system P-protein with E-values of 5.7e-194 and 2.4e-183, respectively (Table 1). These E-values are significant and indicate that there is little chance that the *M. ruber* sequence aligns with the consensus sequence correlating to the glycine dehydrogenase protein simply due to chance. The pair-wise alignments for these two gene family sequences from both *E. coli* and *M. ruber* indicate that there are 39 conserved amino acid residues (marked by the capital letters) between the consensus sequence of the group I pyridoxal-dependent decarboxylases family and the two genes of interest, as seen in Figure 8. These key amino acids are in nearly the same location of the protein and therefore are likely important to the protein's overall structure and function. The pair-wise alignments also indicate there are many conservative substitutions (marked by the plus signs). These conservative substitutions are the replacement of an amino acid in the gene of interest by another amino acid with similar characteristics. Because of this, these substitutions likely have little or no effect on the structure and function of the protein. So, because there is a high degree of similarity (including both conserved amino acids and conservative substitutions) between the consensus sequence of the group I pyridoxal-dependent decarboxylases family and the sequences of the two genes of interest, it is likely that the Mrub_1710 gene and the b2903 gene have a similar structure and function, indicating possible homology.

TIGRFAM indicates that the sequences from both b2903 and Mrub_1710 correlate to TIGR00461, or gcvP: glycine dehydrogenase, with E-values of 0 (Table 1). These E-values are significant and indicate that there is little chance that the *M. ruber* sequence aligns with the P-protein, simply due to chance. This was corroborated by the findings using the PDB program. Both sequences correspond to a PDB code, 4LGL, or glycine decarboxylase P-protein with E-values of 0.00 (Table 1). This is especially important because there are not very many crystal structures in the PDB database, so the fact that Mrub_1710 correlated to the same crystal structure is a strong indication that it is homologous to the b2903 protein. So, all three of these databases suggest that Mrub_1710 is a glycine cleavage system P-protein, just as b2903 is.

A:

B:

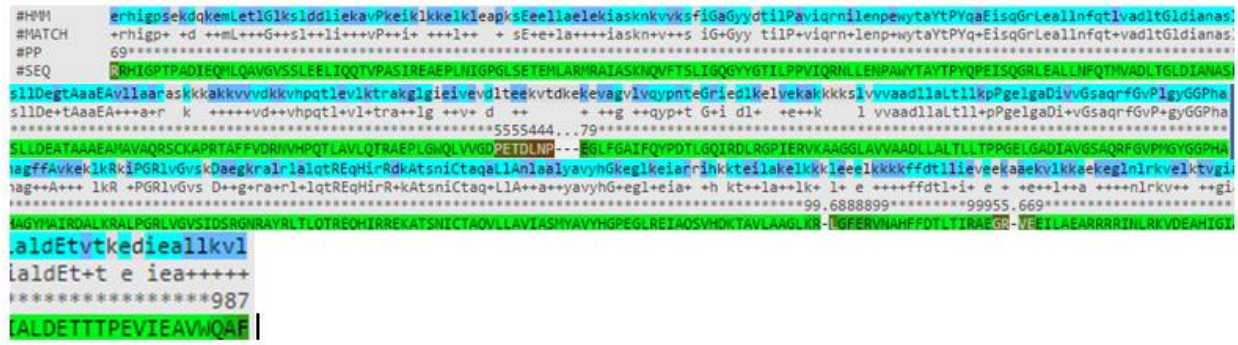


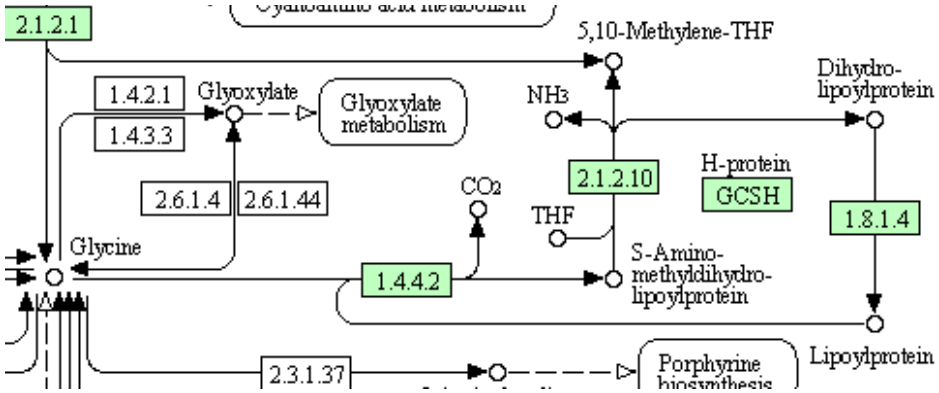
Figure 6. A pairwise alignment between the consensus sequence and both sequences of the genes of interest indicate that there are 39 conserved amino acid residues and numerous conservative substitutions among the regions of the sequence corresponding to the glycine cleavage system P-protein family. *E. coli* b2903 and *M. ruber* Mrub_1710 have very similar structures and likely the same function. The pairwise alignments were generated by Pfam (http://pfam.xfam.org/search) for the consensus sequences of the amino acid kinase family versus A: b2903 and B: Mrub_1710.

Enzymatic Function Data

Both b2903 and Mrub_1710 were indicated to have an E. C. number of E.C.1.4.4.2, which is characterized as a glycine dehydrogenase enzyme (Table 1). Pathways generated by KEGG for both *E. coli* and *M. ruber* both depict b2903 and Mrub_1710 performing the same function within their glycine, serine, and threonine metabolism pathways, as seen in Figure 7. Because these proteins have the same enzymatic function, as predicted by KEGG, and because of the structural similarities between the two proteins as noted in previous sections, it is likely that b2903 and Mrub_1710 are homologs.

`

A:



B:

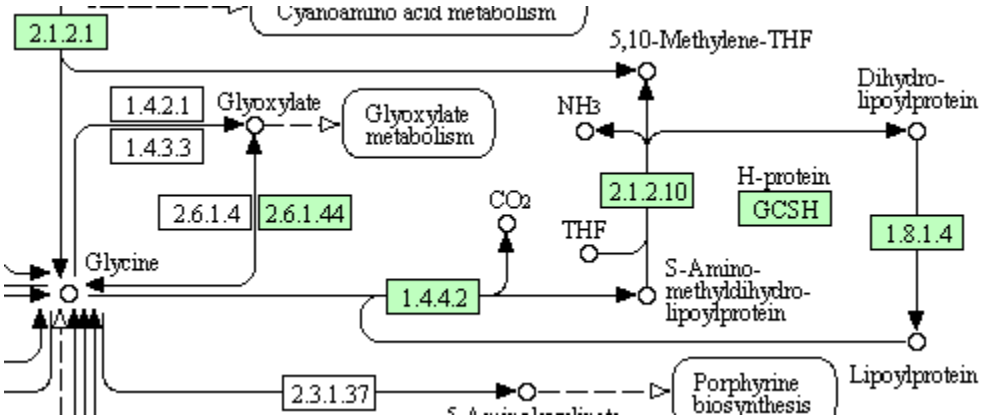
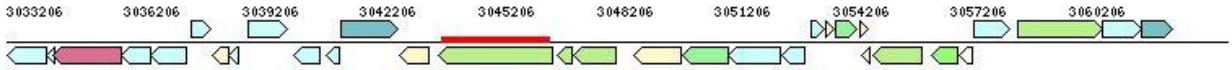


Figure 7. b2903 and Mrub_1710 perform the same function within the glycine degradation pathway (E.C. 1.4.4.2). KEGG (http://www.genome.jp/kegg/kegg2.html) generated the enzymatic function pathways for A: *Escherichia coli* and B: *Meiothermus ruber*.

Ortholog Neighborhood Region Data

The ortholog neighborhood regions for both b2903 and Mrub_1710 indicate that both genes are likely part of a three-gene operon (Figure 8). Two additional gene sequences line up with the b2903 sequence for *E. coli* and its equivalent, Mrub_1710, in *M. ruber*, and all three genes are transcribed in the same direction. Also, all are involved in amino-acid metabolism. So, because the gene sequences have the same transcription direction and the same type of function, they are likely part of a three-gene operon. As mentioned earlier, it is known that b2903 is part of a three-gene operon. So, because Mrub_1710 appears to be so as well according to the ortholog neighborhood region map, this is further evidence that b2903 and Mrub_1710 are homologous.

A:



B:

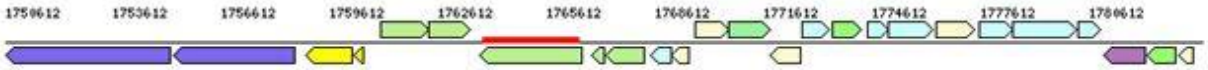


Figure 8. b2903 and Mrub_1710 are both part of three-gene operons involved in glycine degradation, further corroborating their structural and functional similarity. IMG/EDU (https://img.jgi.doe.gov/cgi-bin/edu/main.cgi) generated ortholog neighborhood region maps for A: b2903 and B: Mrub_1710.

T-protein: (b2905 and Mrub_1712)

Sequence-Based Similarity Data

Sequence-based similarity data indicate that b2905 and Mrub_1712 are very structurally similar. A protein BLAST comparison between b2905 and Mrub_1712 produced an E-value of 6e-79 (Table 3, Figure 9). This indicates that the proteins are very structurally similar and that there is very little chance that the two sequences are related solely by chance. A protein BLAST was also performed on both b2905 and Mrub_1712 separately. Two of the top hits for b2905 were sequences from *Escherichia fergusonii* (751 bits, E-value=0.0) and *Shigella flexneri* (751 bits, E-value=0.0), both of which are predicted to be aminomethyltransferase, or glycine cleavage system protein T-proteins, as well (Table 3). Two of the top hits for Mrub_1712 were *Meiothermus cerbereus* (622 bits, E-value=0.0) and *Meiothermus rufus* (594 bits, E-value=0.0), both of which are predicted glycine cleavage system protein T-proteins. This is important because both proteins, independently of one another, are characterized by being related to glycine cleavage system protein T-proteins. A conserved domain database (CDD) search assigned the same cluster of orthologous groups (COG) to both b2905 and Mrub_1712. The protein sequences were labeled as COG0404, or glycine cleavage system T-protein (aminomethyltransferase) (Table 3). The E-values for b2905 and Mrub_1712 were E=3.27e-154 and E=4.71e-143, respectively. This result is significant and indicates that b2905 and Mrub_1712 likely have the same or similar function that validates their structural similarity. So, all three databases producing the sequence-based similarity data indicate that b2905 and Mrub_1712 are highly structurally similar and therefore likely perform a similar function.

Table 3: *E. coli* b2905 and Mrub_1712 are orthologs

Description of Evidence	***E. coli* (b2905)**	***M. ruber* (Mrub_1712)**
BLAST *E. coli* against *M. ruber*	Score: 1098 bits; E-value: 6e-79	
Protein BLAST	*Escherichia fergusonii: 751 bits, E-value: 0.0 (glycine cleavage system protein T)* *Shigella Flexneri: 751 bits, E-value: 0.0 (aminomethyltransferase)*	*Meiothermus cerbereus: 622 bits, E-value: 0.0 (glycine cleavage system protein T)* *Meiothermus rufus: 594 bits, E-value: 0.0 (glycine cleavage system protein T)*
CDD (COG) Category	COG0404 Glycine cleavage system T-protein (aminomethyltransferase)	
	E= 3.27e-154	E=4.71e-143
Cellular Localization	Cytoplasmic	
Pfam: Protein Family	PF01571, CL0289, Aminomethyltransferase folate-binding domain	
	E=5.3e-90	E=1.9e-41
	PF08669, CL 0289, Glycine cleavage T-protein C-terminal barrel domain	
	E=6.2e-18	E=1e-20
TIGRfam: Protein Family	TIGR00528 gcvT: glycine cleavage system T-protein	
	E=6.8e-274	E=1.2e-138
PDB	1VLO Aminomethyltransferase (T-protein)	
	E=0.00	E=1.4501E-83
E.C. Number	E.C.2.1.2.10 Aminomethyltransferase	
KEGG Pathway	Glycine, Serine, and Threonine Metabolism	

Score	Expect	Method	Identities	Positives	Gaps
237 bits(605)	6e-79	Compositional matrix adjust.	155/363(43%)	204/363(56%)	19/363(5%)

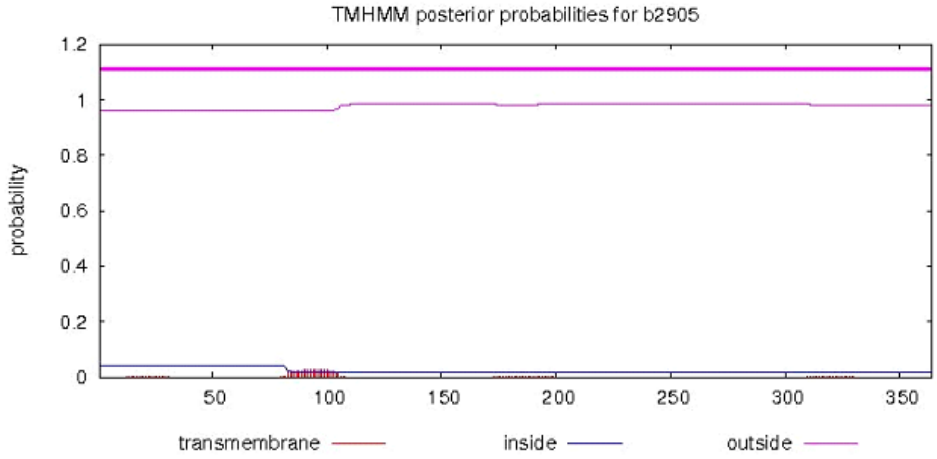
```
Query  3    QQTPLYEQHTLCGARMVDFHGWMMPLHYGSQIDEHHAVRTDAGMFDVSHMTIVDLRGSRT  62
            + TPL++ H   GARMV F G+ MP+ Y S   EH AVR   GMFDVSHM    ++G
Sbjct  2    KTTPLHQAHLALGARMVPFAGYEMPIQYTSITSEHLAVRGLVGMFDVSHMGEFWIKGPGA  61

Query  63   REFLRYLLANDVAKLTKSGKALYSGMLNASGGVIDDLIVYYFTEDFFRLVVNSATREKDL  122
             EFL+Y   NDV KL K G+A YS + NA GGV+DD+ +Y   E+ + +VVN+A  EKD
Sbjct  62   LEFLQYATLNDVTKL-KVGRAHYSMLPNAQGGVVDDIYLYRTGEEEYLMVVNAANIEKDW  120

Query  123  SWITQHAEPFGIEIT-VRDDLSMIAVQGPNAQAKAATLFNDAQRQAVEGMKPFFGVQAGD  181
             + + AE F + +    D  ++IAVQGP A A    L  D    + +    F G  AG
Sbjct  121  EHLQRLAEGFEVRLEDASDFFALIAVQGPQAVAVLQKL-CDTDLVSRKKNDTFMGKLAGK  179

Query  182  -LFIATTGYTGEAGYEIALPNEKAADFWRALVEAGVKPCGLGARDTLRLEAGMNLYGQEM  240
             +  A TGYTGE GYE+ +  ++A   W AL+EAGV PCGLGARDTLRLEAG  LYG E+
Sbjct  180  WVRFARTGYTGEDGYEVFVAPDEAPAVWAALLEAGVTPCGLGARDTLRLEAGFPLYGHEL  239

Query  241  DETISPLAANMGWTIAWEPADRDFIGREA-LEVQREHGTEKLVGLVMTEKGVLRNELPVR  299
             +T +PL     W +      ++F G++A L+   E    +LVGL + E G+ R    V
Sbjct  240  TDTTNPLCTPFDWVVK---GQKEFFGKQAMLDAACER---RLVGL-LVEGGIPREGYRV-  291

Query  300  FTDAQGNQHEGIITSGTFSPTLGYSIALARVPE---GIGETAIVQIRNREMPVKVTKPVF  356
                 G +  GI+TSGT SP L   I LA V     G+G    V+IR R  P  V +  F
Sbjct  292  ---LGGGKEVGILTSGTHSPVLKKGIGLAYVQSDWAGVGTALEVEIRGRAAPAAVVETPF  348

Query  357  VRN  359
            V+
Sbjct  349  VKR  351
```

Figure 9. A protein BLAST between Mrub_1712 and b2905 indicates a very low E-value—6e-79 — which is indicative of structural similarity. NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi) generated this protein BLAST.

Localization Data

All of the localization data indicates that the b2905 gene and the Mrub_1712 gene sequences code for cytoplasmic proteins. TMHMM predicted 0 transmembrane helices for both gene products, as indicated by the transmembrane topology graphs depicted in Figure 10. With no transmembrane helices, it is most likely that the gene product would be found in the cytoplasm, and not inserted into a membrane. SignalP determined signal peptide probabilities for both b2905 and Mrub_1712 to be 0.115 and 0.358, respectively. These probabilities are very low, which indicates that there is likely no signal peptide coded for in the gene sequences. Because of this, it is unlikely that there is a signal peptide attached to these genes that would indicate their migration to a different part of the cell. So, these proteins are likely made in the cytoplasm and then remain there to fulfill their roles. The SignalP prediction graphs are represented below in Figure 11. PSORT-B also indicated that both b2905 and Mrub_1712 are cytoplasmic proteins as well, as described in Table 4. Along with this program, LipoP suggested that the best prediction for cellular location of both was in the cytoplasm. These findings were also indicated by the posterior probability graphs generated by Phobius, as seen in Figure 12. These graphs do not note any large areas of hydrophobicity, which would be indicative of a transmembrane protein. So, all five of these programs corroborate one another and indicate that both b2905 and Mrub_1712 are cytoplasmic proteins.

A



B

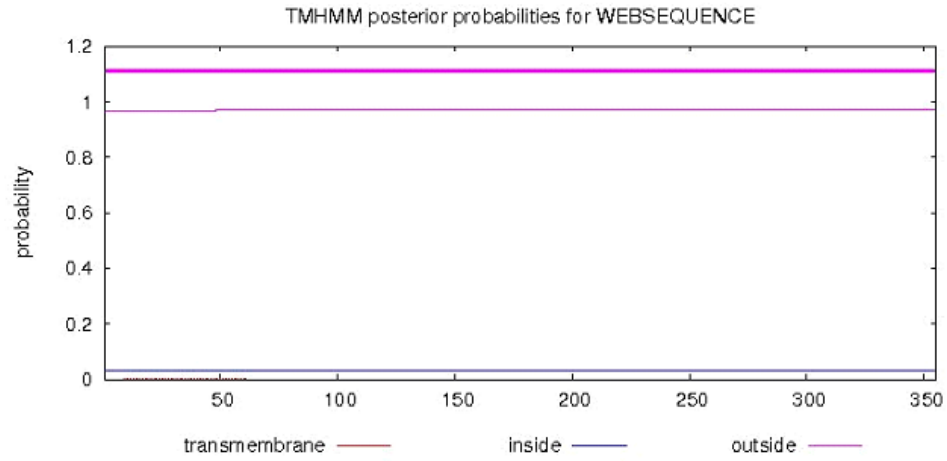


Figure 10. b2905 and Mrub_1712 do not contain TMH regions; a cytoplasmic location is predicted. Server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) created this hydrophobicity plot depicting the TMHMM posterior probabilities for A: b2905 and B: Mrub_1712.

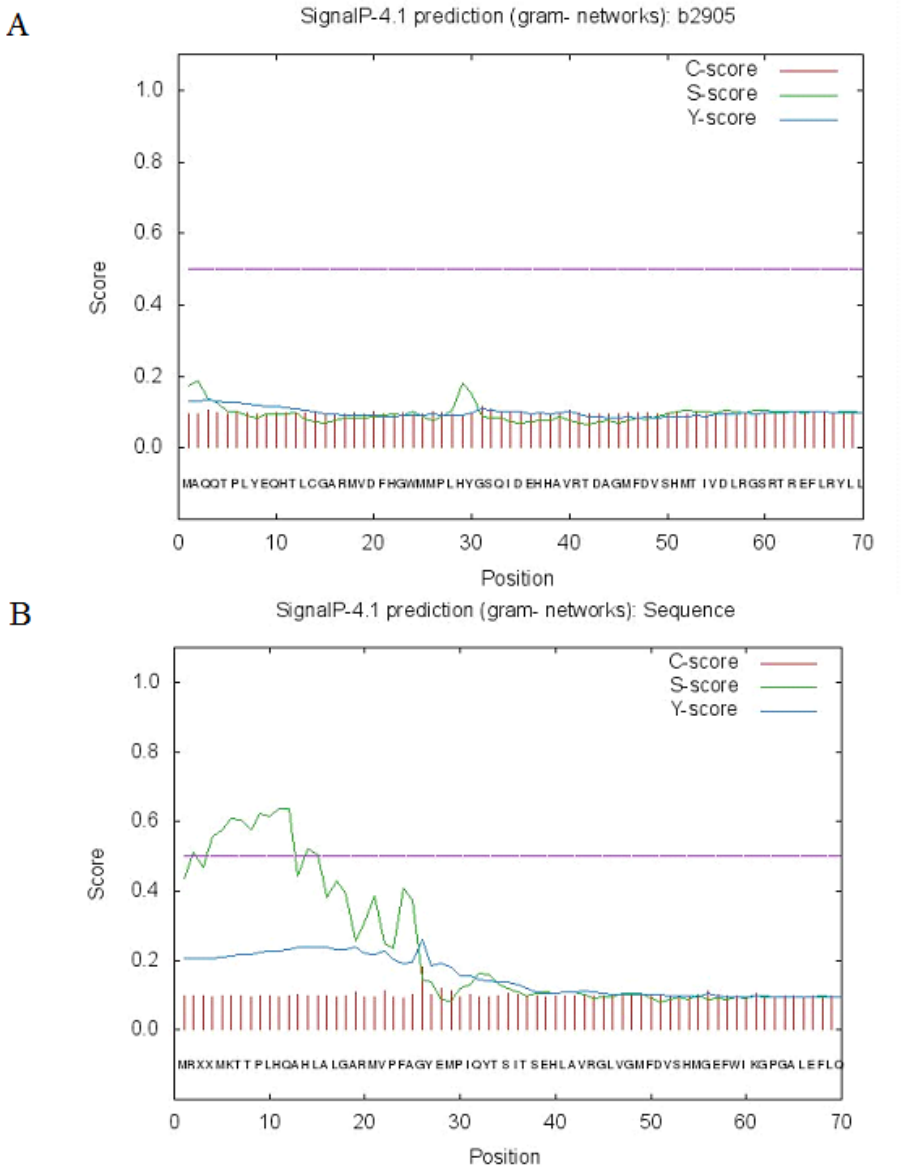
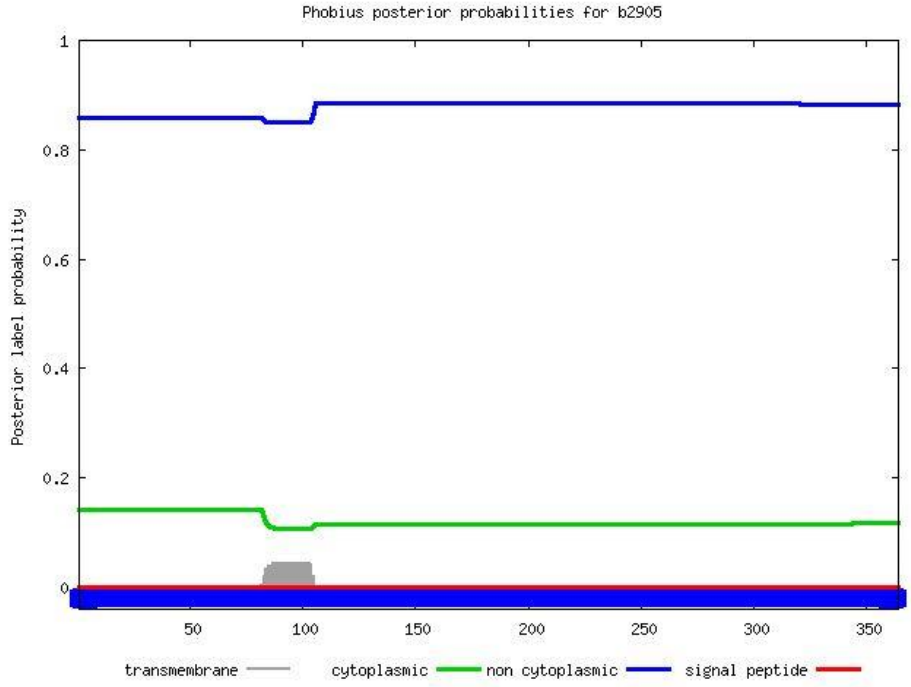


Figure 11. b2905 and Mrub_1712 do not contain regions indicating the presence of a signal peptide; a cytoplasmic cellular location is predicted. SignalP 4.1 server (http://www.cbs.dtu.dk/services/SignalP) created the SignalP-4.1 prediction graphs for A: b2905 and B: Mrub_1712.

Table 4: *E. coli* b2905 and Mrub_1712 are cytoplasmic proteins

Cellular Location	*E. coli* B2905	*M. ruber* Mrub_1712
Cytoplasm	9.97	9.97
Cytoplasmic Membrane	0.01	0.01
Periplasm	0.01	0.01
Outer Membrane	0.00	0.00
Extracellular	0.00	0.00

A



B

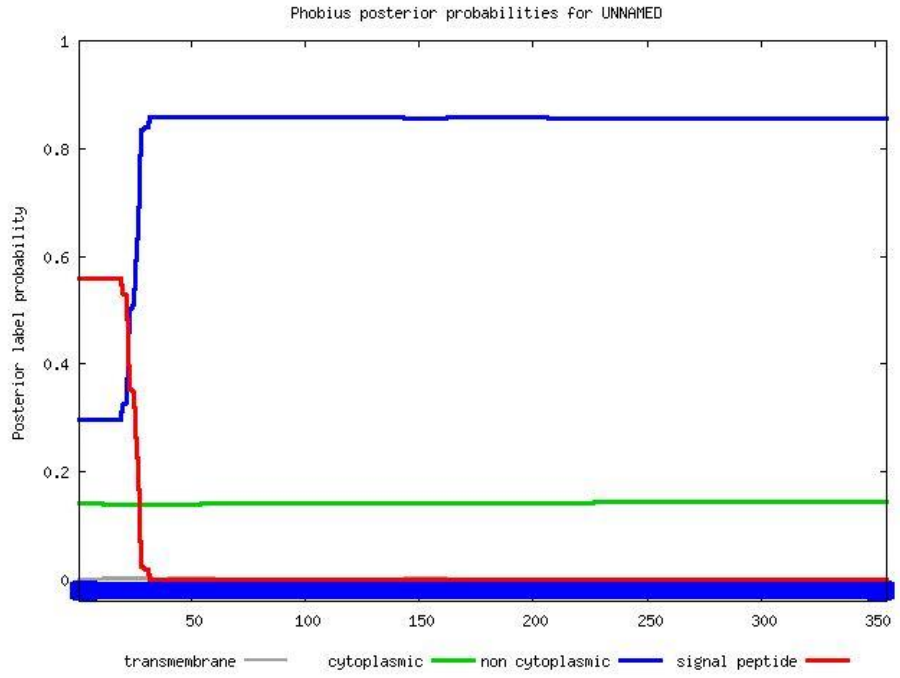


Figure 12. b2905 and Mrub_1712 do not contain large hydrophobic regions; a cytoplasmic location is predicted. Phobius (http://phobius.sbc.su.se/) created the posterior probability graphs for A: b2905 and B: Mrub_1712.

Structure-Based Evidence

Pfam indicates that the sequences from both b2905 and Mrub_1712 correlate to PF01571, CL0289, or aminomethyltransferase folate-binding domain with E-values of 5.3e-90 and 1.9e-41, respectively (Table 3). These E-values are significant and indicate that there is little chance that the *M. ruber* sequence aligns with the consensus sequence corresponding to the aminomethyltransferase protein, simply due to chance. The pair-wise alignments for these two gene family sequences from both *E. coli* and *M. ruber* indicate that there are 17 conserved amino acid residues for the glycine cleavage system family among the two genes of interest and the consensus sequence, as seen in Figure 13. These key amino acids are in nearly the same location of the protein and therefore are likely important to the protein's overall structure and function. The pair-wise alignments also indicate there are many conservative substitutions (marked by the plus signs). These conservative substitutions are the replacement of an amino acid in the gene of interest by another amino acid with similar characteristics. Because of this, these substitutions likely have little or no effect on the structure and function of the protein. So, because there is a high degree of similarity (including both conserved amino acids and conservative substitutions) between the consensus sequence of the group I pyridoxal-dependent decarboxylases family and the sequences of the two genes of interest, it is likely that the Mrub_1712 gene and the b2905 gene have a similar structure and function, indicating possible homology.

Pfam also indicated that both sequences were correlated PF08669, CL0289 with E-scores of 6.2e-18 and 1e-20 for b2905 and Mrub_1712, respectively. These E-values are significant and indicate that there is little chance that the *M. ruber* sequence aligns with the consensus sequence correlating to the aminomethyltransferase protein, simply due to chance. The pair-wise alignments for these two gene family sequences from both *E. coli* and *M. ruber* indicate that there are 3 completely conserved amino acid residues for the amino acid kinase family among the two genes of interest and the consensus sequence as seen in Figure 14. These key amino acids are in nearly the same location of the protein and therefore are likely important to the protein's overall structure and function. The pair-wise alignments also indicate there are many conservative substitutions (marked by the plus signs). These conservative substitutions are the replacement of an amino acid in the gene of interest by another amino acid with similar characteristics. Because of this, these substitutions likely have little or no effect on the structure and function of the protein. So, because there is a high degree of similarity (including both conserved amino acids and conservative substitutions) between the consensus sequence of the group I pyridoxal-dependent decarboxylases family and the sequences of the two genes of interest, it is likely that the Mrub_1712 gene and the b2905 gene have a similar structure and function, indicating possible homology.

TIGRFAM indicates that the sequences from both b2905 and Mrub_1712 correlate to TIGR00528, or gcvT: glycine cleavage system T-protein, with E-values of 6.8e-274 and 1.2e-138, respectively (Table 1). The E-values are significant and indicate that there is little chance that the *M. ruber* sequence aligns with the aminomethyltransferase, or T-protein, simply due to chance. This was corroborated by the findings using the PDB program. Both sequences correspond to a PDB code, 1VLO, or aminomethyltransferase (T-protein) with E-values of 0.00 and 1.4501E-83 for *E. coli* and *M. ruber*, respectively (Table 1). This is especially important because there are not very many crystal structures in the PDB database, so the fact that Mrub_1712 correlated to the same crystal structure is a strong indication that it is homologous to the b2905 protein. So, all three of these databases suggest that Mrub_1712 is aminomethyltransferase, just as b2905 is.

A:

B:

Figure 13. A pairwise alignment between the consensus sequence and both sequences of the genes of interest indicate that there are 17 conserved amino acid residues and numerous conservative substitutions among the regions of the sequence corresponding to the glycine cleavage system P-protein family. *E. coli* b2905 and *M. ruber* Mrub_1712 have very similar structures and likely the same function. The pairwise alignments were generated by Pfam (http://pfam.xfam.org/search) for the consensus sequences of the amino acid kinase family versus A: b2905 and B: Mrub_1712.

A:

```
#HMM      gdfiGkeallrqkaegvarrrlvgleleltaplregepvya...ddgekvGevtsgtySptlgknialayvrteharigteveveirgkriratvvk
#MATCH    +dfiG+eal+ q+++g+   +lvgl ++++++lr   pv     ++++  G +tsgt+Sptlg +iala v++    ig++  v+ir+++++++v+k
#PP       59**************8..7*****99988*********99998888**********************54...4677777888********998
#SEQ      RDFIGREALEVQREHGT--EKLVGLVMTEKGVLRNELPVRFtdaQGNQHEGIITSGTFSPTLGYSIALARVPE---GIGETAIVQIRNREMPVKVTK
```

B:

```
#HMM      dfiGkeallrqkaegvarrrlvgleleltaplregepvyaddgekvGevtsgtySptlgknialayvrteharigteveveirgkriratvvk
#MATCH    +f+Gk+a+l    e    rrlvgl +e+ +++reg+ v+  +g++vG +tsgt+Sp l+k i+layv  + a +gt++eveirg++ +a vv+
#PP       7*********9998....9*****9998.9**********.**************************************************97
#SEQ      EFFGKQAMLDAACE----RRLVGLLVEG-GIPREGYRVLG-GGKEVGILTSGTHSPVLKKGIGLAYVQSDWAGVGTALEVEIRGRAAPAAVVE
```

Figure 14. A pairwise alignment between the consensus sequence and the sequences of the gene of interest indicate that there are 3 conserved amino acid residues and numerous conservative substitutions among the regions of the sequence corresponding to the glycine cleavage system P-protein family. *E. coli* b2905 and *M. ruber* Mrub_1712 have very similar structures and likely the same function. The pairwise alignments were generated by Pfam (http://pfam.xfam.org/search) for the consensus sequences of the amino acid kinase family versus A: b2905 and B: Mrub_1712.

Enzymatic Function Data

Both b2905 and Mrub_1712 were indicated to have an E. C. number of E.C.2.1.2.10, which is characterized as an aminomethyltransferase enzyme (Table 1). Pathways generated by KEGG for both *E. coli* and *M. ruber* both depict b2905 and Mrub_1712 performing the same function within their glycine, serine, and threonine metabolism pathways, as seen in Figure 15. Because these proteins have the same enzymatic function, as predicted by KEGG, and because of the structural similarities between the two proteins as noted in previous sections, it is likely that b2905 and Mrub_1712 are homologs.

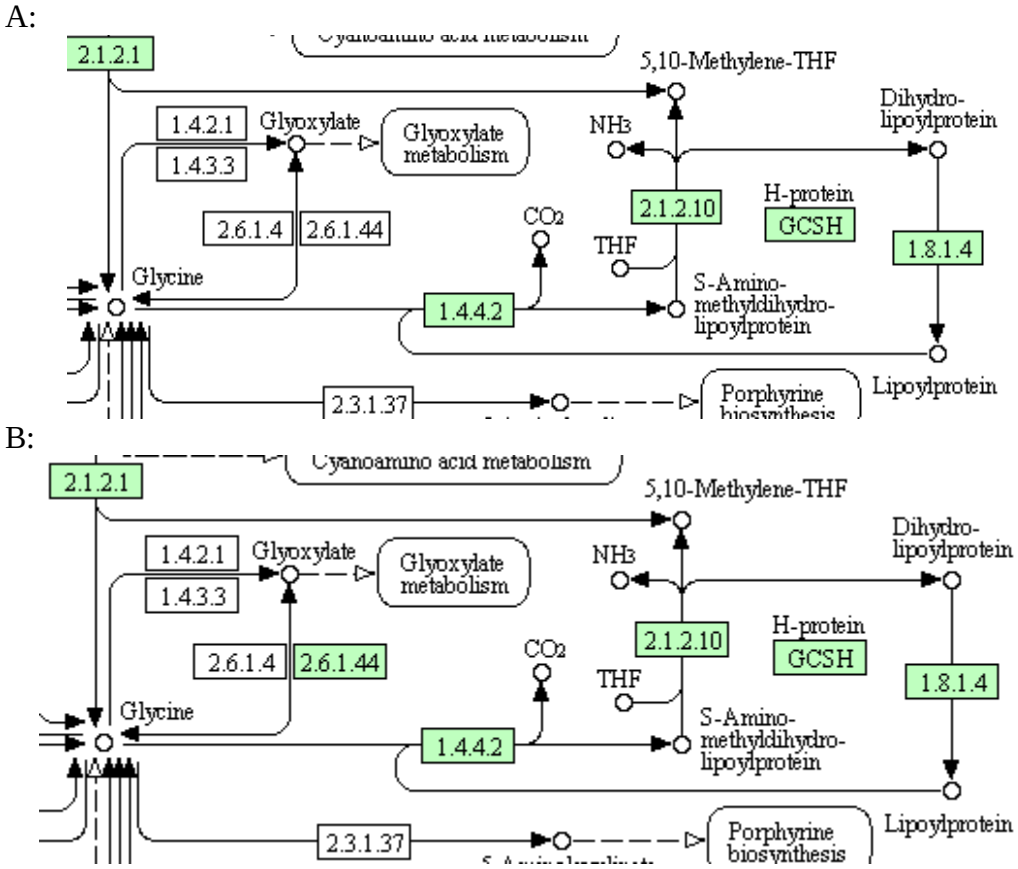
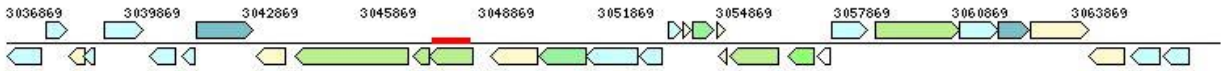


Figure 15. b2905 and Mrub_1712 perform the same function within the glycine degradation pathway (E.C. 2.1.2.10). KEGG (http://www.genome.jp/kegg/kegg2.html) generated the enzymatic function pathways for A: *Escherichia coli* and B: *Meiothermus ruber*.

Ortholog Neighborhood Region Data

The ortholog neighborhood regions for both b2905 and Mrub_1712 indicate that both genes are likely part of a three-gene operon (Figure 16). Two additional gene sequences line up with the b2905 sequence for *E. coli* and its equivalent, Mrub_1712, in *M. ruber*, and all three are transcribed in the same direction. Also, all are involved in amino-acid metabolism. So, because the gene sequences have the same transcription direction and the same type of function, they are likely part of a three-gene operon. As mentioned earlier, it is known that b2905 is part of a three-gene operon. So, because Mrub_1712 appears to be so as well according to the ortholog neighborhood region map, this is further evidence that b2905 and Mrub_1712 are homologous.

A:



B:

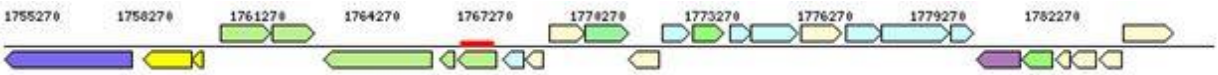


Figure 16. b2905 and Mrub_1712 are both part of three-gene operons involved in glycine degradation, further corroborating their structural and functional similarity. IMG/EDU (https://img.jgi.doe.gov/cgi-bin/edu/main.cgi) generated ortholog neighborhood region maps for A: b2905 and B: Mrub_1712.

Conclusions

The hypotheses that the *Meiothermus ruber* gene Mrub_1710 is a homolog of the *Escherichia coli* gene b2903 and that the *Meiothermus ruber* gene Mrub_1712 is a homolog of the *Escherichia coli* gene b2905 were supported by the evidence collected using the GENI-ACT system. A BLAST comparison between the *M. ruber* genes and their hypothesized homologous *E. coli* genes yielded very low E-values, indicating that the two sets of enzymes have high structural similarity and therefore likely have the same function and are feasibly homologous to one another.

The protein family names and categories generated by each of the programs used for each of the sets of proteins (Pfam, CDD (COG), TIGRfam, E.C. number, and PDB category names) also corroborated one another. The protein names generated for b2903 and Mrub_1710 included glycine dehydrogenase, glycine cleavage system, Group I pyridoxal-dependent decarboxylases, P-protein, and glycine decarboxylase. Although the name glycine dehydrogenase is the accepted name for this protein, P-protein and glycine decarboxylase are accepted alternative terms. Also because this protein is one of the components of the glycine cleavage system, this term accurately describes these proteins as well. So, all names are applicable to the b2903 protein and indicate a strong likelihood that Mrub_1710 is also the glycine dehydrogenase protein. The protein names generated for b2905 and Mrub_1712 included aminomethyltransferase, glycine cleavage system, and T-protein. Although the name aminomethyltransferase is the accepted name for this protein, T-protein is an accepted alternative term. Also because this protein is one of the components of the glycine cleavage system, this term accurately describes these proteins as well. So all names are applicable to the b2905 protein and indicate a strong likelihood that Mrub_1712 is also the aminomethyltransferase protein.

Also, all programs used predicted that both the *E. coli* genes b2903 and b2905 as well as the *M. ruber* genes Mrub_1710 and Mrub_1712 are cytoplasmic proteins. Along with this data, the KEGG pathways generated for each of the proteins for both *E. coli* and *M. ruber* indicated that the two sets of proteins (b2903/Mrub_1710 and b2905/Mrub_1712) perform the same enzymatic function within their respective pathways. Lastly, the ortholog neighborhood regions for b2903 and b2905 indicate that these genes are a part of a three-gene operon. The same was indicated by the ortholog neighborhood regions for Mrub_1710 and Mrub_1712, which further supports their homology to the b2903 and b2905 genes, respectively.

References

Fujiwara K and Motokawa Y. 1983. Mechanism of the glycine cleavage reaction. Steady state kinetic studies of the P-protein-catalyzed reaction. J Biol Chem 258(13):8156-62.

Fujiwara K, Okamura-Ikeda K, Motokawa Y. 1984. Mechanism of the glycine cleavage reaction. Further characterization of the intermediate attached to H-protein and of the reaction catalyzed by T-protein. J Biol Chem 259(17):10664-8.

Garrity G, Boone D, Castenholz R, editors. 2001. Bergey's manual of systematic bacteriology. 2nd ed. Baltimore, MD: Bergey's Manual Trust and Williams & Wilkins; p417.

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 76: 1367–1376.

Kikuchi G, Motokawa Y, Yoshida T, and Hiraga K. 2008. Glycine cleavage system: reaction mechanism, physiological significance, and hyperglycemia. Proc Jpn Acad 84(7):246-263.

Motokawa Y, Kikuchi, G. 1974. Glycine metabolism by rat liver mitochondria. Reconstitution of the reversible glycine cleavage system with partially purified protein components. Arch Biochem 164(2):624-33.

Okamura-Ikeda K, Fujiwara K, Motokawa Y. 1987. Mechanism of the glycine cleavage reaction: Properties of the reverse reaction catalyzed by T-protein. J Biol Chem 262(14):6746-49.

Okamura-Ikeda K, Ohmura Y, Fujiwara K, Motokawa Y. 1993. Cloning and nucleotide sequence of the gcv operon encoding the *Escherichia coli* glycine-cleavage system. Eur J Biochem 216(2):539-48.

Phylogenetic Diversity. [Internet]. 2015. Joint Genome Institute; [2015 December 15]. Available from: http://jgi.doe.gov/our-science/science-programs/microbial-genomics/phylogenetic-diversity.

Wu, D et al. 2009. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. Nature 462(7276):1056-60.