2017

Mrub_2052, Mrub_0628, and Mrub_2034 genes are predicted to be orthologous to b0688, b2039, and b3789 genes found in *Escherichia coli*, which are involved in streptomycin biosynthesis

James P. Hartnett
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Mrub_2052, Mrub_0628, and Mrub_2034 genes are predicted to be orthologous to b0688, b2039, and b3789 genes found in *Escherichia coli*, which are involved in streptomycin biosynthesis

James P. Hartnett
Dr. Lori R. Scott Laboratory
Biology Department, Augustana College
639 38th Street, Rock Island IL 61201

Introduction

Project Overview

In this project three genes believed to be involved in streptomycin biosynthesis were analyzed. These genes include Mrub_2052, Mrub_0628, and Mrub_2034, which are found in the bacteria *Meiothermus ruber* (*M. ruber*). A verity of different bioinformatics tools were utilized throughout this process including BLAST, T-Coffee, WebLogo, TMHMM, SignalP, LipoP, PSORTb, Phobius, IMG/EDU gene finder, TIGRFAM, Pfam, RCSB PDB, Kegg Pathway database, MetaCyc, ExPASy, and Phylogeny. These tools assisted in providing information that could help determine the location, function, structure, and identity of these genes. Three orthologous genes to those of interest were also analyzed. They included b0688, b2039, and b3789, which are found in the bacteria *Escherichia coli* (*E. coli*) (Kanehisa et al., 2016). *E. coli* has been studied in detail, so much is known about these orthologous genes (Blatterner,1997). This information assisted in confirming the predictions made by the bioinformatics tools in regards the *M. ruber* genes of interest. Overall, this project has not only helped gain a better understanding of streptomycin biosynthesis, but it has also provided some much needed insight on *M. ruber* as a whole.

Why is it Important to Study *Meiothermus Ruber*?

Meiothermus ruber was initially isolated in a Russian hot spring where it was concluded to be a gram negative, obligate aerobic bacterial species, but since then very little has been published in regards to *M. ruber* suggesting very little is known about it (Tindal et al., 2010). Considering this, back in 2009, the *Meiothermus ruber* Genome Analysis Project was created (Scott, 2016). This project is in collaboration with Joint Genome Institute (JGI) as part of its Genomic Encyclopedia of Bacteria and Archaea (GEBA) project. The goal of this project is to study organisms that have been deemed obscure in hopes to discover new genes and processes researchers have not yet stumbled upon (http://jgi.doe.gov/). New discoveries such as these can lead to several breakthroughs in fields such as energy production and pathogenesis (http://jgi.doe.gov/). *M. ruber* is one of the "obscure" organisms that was chosen to be studied. It is part of the Deinococcus-Thermus phylum, a collection of organisms that typically live in high temperature environments (35ºC-70ºC) (Tindall et al., 2010). In 2010 its genome was sequenced showing that it has a total of about 3,105 genes and 71.8% of them code for proteins that have been assigned a presumed function (Tindall et al., 2010). This is one of the major reasons why *M. ruber* was chosen to be studied. Through studying these proteins found in *M. ruber* a more diverse plethora of knowledge can be gained in regards to protein function.

***Escherichia coli* as a Model Organism**

Escherichia coli (*E. coli*) is a gram negative bacterial species that has been highly studied (Blatterner, 1997). It was initially chosen to have its entire genome sequenced due to how easily it can be grown in a laboratory setting (Blatterner,1997). Over 4288 protein coding genes in *E. coli* have been analyzed (Blattner, 1997). A protein BLAST alignment was conducted and confirmed that *M. ruber* genes Mrub_2052, Mrub_0628, and Mrub_2034 genes have very similar sequences to b0688, b2039, and b3789 genes found in *E. coli*. This suggests that the proteins that these genes code for may have similar function. b0688 has been determined to code for the protein phosphoglucomutase, which is involved in streptomycin biosynthesis pathway. b2039 and b3789 have been determined to be paralogs that code for the protein glucose-1-phosphate thymidylyltransferase, which is also involved in the streptomycin biosynthesis pathway. These genes are not part of the same operon.

Streptomycin Biosynthesis

Just like penicillin, streptomycin is an antibiotic (Schatz, 1944). Antibiotics help organisms defend against invading bacterial species (Schatz, 1944). According to the bioinformatics tool KEGG, it has been determined that *E. coli* has a streptomycin biosynthesis pathway (Kanehisa et al., 2016). This is particularly interesting because *E. coli* itself is a gram negative bacterial species. In Figure 3 the streptomycin biosynthesis pathway for *E. coli* can be seen (Kanehisa et al., 2016). The genes of interest in this pathway include b0688, which is represented by E.C. number 5.4.2.2, as well as b2039 and b3789, which are represented by E.C. number 2.7.7.24. These genes are not a part of an operon. Enzyme 5.4.2.2 (b0688) codes for phosphoglucomutase. Phosphoglucomutase converts D-Glucose- 6P (also known as D-glucopyranose 6-phosphate) to D-Glucose-1P (also known as α-D-glucopyranose 1-phosphate). A visual for this can be seen in Figure 1.

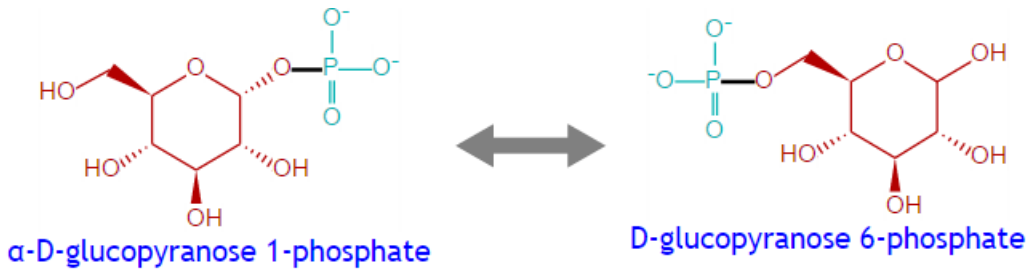


Figure 1. Products and reactants of phosphoglucomutase reaction. Image taken from https://ecocyc.org/

Maximum activity of phosphoglucomutase is obtained in the presence of alpha-D-glucose 1,6-bisphosphate (http://www.expasy.ch). Phosphoglucomutase is involved in several different pathways along with streptomycin biosynthesis (Kanehisa et al., 2016). Some include glycolysis, the pentose phosphate pathway, and amino sugar and nucleotide sugar metabolism (Kanehisa et al., 2016). Enzyme 2.7.7.24 (b2039 and b3789) codes for glucose-1-phosphate thymidylyltransferase. Glucose-1-phosphate thymidylyltransferase converts D-Glucose-1P to dTDP-glucose (also known as dTDP-α-D-glucose). A visual representation of this can be seen on Figure 2.

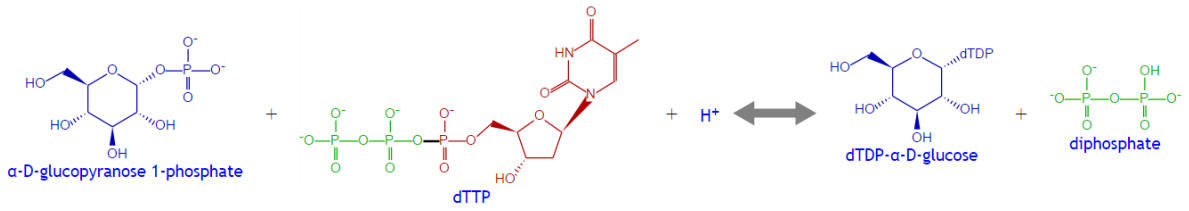


Figure 2. Glucose-1-phosphate thymidylyltransferase reaction. α-D-glucopyranose 1-phosphate is reacted with dTTP to form dTDP-α-D-glucose by Glucose-1-phosphate thymidylyltransferase. Images taken from https://ecocyc.org/

Glucose-1-phosphate thymidylyltransferase is involved in several other pathways then streptomycin biosynthesis (Kanehisa et al., 2016). Some include polyketide sugar unit biosynthesis and acarbose and validamycin biosynthesis (Kanehisa et al., 2016). These are just a couple of steps in the streptomycin biosynthesis pathway. Gaining a better understanding about this pathway may help to gain a better understanding of why bacterial species produce antibiotics.

Purpose/Hypothesis

In this project a wide variety of bioinformatics tools were utilized to analyze the genes Mrub_2052, Mrub_0628, and Mrub_2034 in *Meiothermus ruber*. These genes and their protein products were compared to the genes b0688, b2039, and b3789 from *Escherichia coli* to determine if the *M. ruber* genes are orthologs of the *E. coli* genes. One of the tools used was the Basic Local Alignment Search Tool (BLAST). Protein BLAST compares the amino acid sequences of two or more enzymes and analyzes their sequence similarity. According to pBLAST, the Mrub genes and the *E.coli* genes have very similar sequences (see the Results section). This suggested that the M. rub genes have a similar function to their respective E. coli counterparts. Consequently, I hypothesize that Mrub_2052, Mrub_0628, and Mrub_2034 are orthologous to b0688, b2039, and b3789.

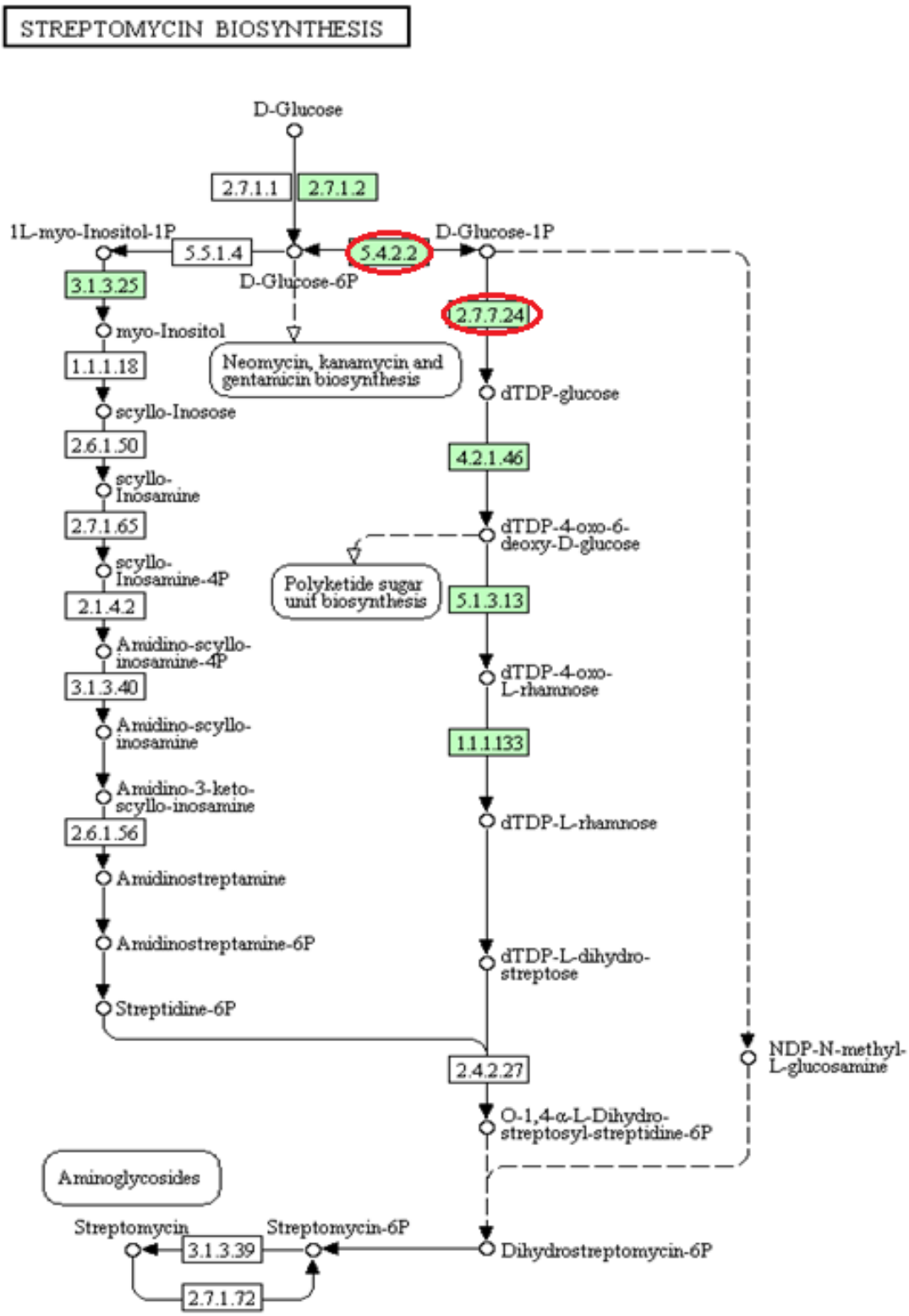


Figure 3. Streptomycin biosynthesis pathway in *Escherichia coli* showing reactants, products and enzymes involved. E.C. number 5.4.2.2 circled to identify gene of interest b0688. E.C. number 2.7.7.24 circled to identify paralog genes of interest b2039 and b3789. Image taken from http://www.genome.jp/kegg-bin/show_pathway?eco00521+b2039.

Methods

The first step to confirming if Mrub_2052, Mrub_0628, and Mrub_2034 are orthologous to b0688, b2039, and b3789, respectively, was to use the program KEGG (Kanehisa et al., 2016) to determine if both *E. coli* and *M. ruber* were predicted to have the same streptomycin biosynthesis pathway components. The next step was to determine how similar the putative ortholgs were to each other using the program the Basic Local Alignment Search Tool (BLAST) (Madden, 2002). BLAST forms a pairwise alignment with these known sequences and shows their degree of similarity based on the measure called an E-value. E-values are used to compare the alignment of two sequences. If an E-value if very small that means that the sequences are very similar and the two genes most likely code for proteins of similar function. Small E-value suggest that the alignment did not occur due to chance. If an E- value is smaller than 10^{-100}, it is sometimes given as 0.0 (Madden, 2002). The cutoff for E-values was chosen by Dr. Scott as 0.001. The next tool used is the Conserved Domain Database Search (CDD) (Marchler-Bauer et al. 2016). The CDD is used to find Clusters of Orthologous Groups (COGs). If the *E. coli* gene and the *M. ruber* gene of interest both belong to the same COG group it suggests that they contain the same protein domain, and most likely are orthologous to one another. Next, the Tree-based Consistency Objective Function for Alignment Evaluation tool (T-Coffee) was used (Notredame et al.,200). T-Coffee is a multiple sequence alignment tool that compares the amino acid sequence of interest to multiple similar/homologous sequences. One can analyze these alignments to determine which amino acids are conserved across all the different sequences. If an amino acid is conserved across different species it suggests it is important to the proteins function. WebLogo is then uses the data from T-Coffee to create a visual representation of these results (Crooks ., 2004). The next step is finding what is known about the cellular location of these proteins. The tool Transmembrane Helices Hidden Markov Models (TMHMM) compares the amino acid sequence of interest to known helices that typically cross the cell membrane (Krogh et al., 2016). The results predict if the protein has any transmembrane helices and if it is predicted to be outside or inside the cell membrane. SignalP (Petersen et al., 2011) predicts if a transmembrane helicies at the N-terminus of the protein are actually that or signal peptides. Signal peptides can be confused with transmembrane helices. LipoP was a tool used to give an overall prediction of the location of the protein of interest based on its amino acid sequence (Juncker et al., 2003). PSORT-B was also used to help determine the location of the protein of interest (N.Y. Yu et al., 2010). It gives a series of scores that predict if the protein is in the cytoplasmic region, periplasmic region, cytoplasmic membrane, outer membrane, or extracellular region. Phobius is the next tool used and it creates a visual representation of the results of TMHMM and SignalP (Kall et al., 2004). The IMG/EDU Gene Finder tool was used to determine if there were any other possible start codons in the amino acid sequence of interest (Markowitz et al. 2012). It displayed three reading frames and the possible other alternative start codons. If the proposed alternative was the correct distance from the Shine-Dalgarno region and in the correct reading frame it might be an alternative start codon. TIGRFAM was the next tool used (Haft et al., 2001). It is a collection of protein families constructed from full-length protein sequences. It compares the amino acid sequence of interest to these protein families and helps predict the name of the protein based on bit scores and E-values. Recall, the smaller the E-value the more likely the sequences are similar. Pfam identifies if a protein belongs to a particular protein family, or it might identify a particular protein domain in the query sequence. (Finn et al.). Protein Data Bank (PDB) (Berman et. al., 2000) is a curated collection of crystalized proteins.If a PDB hit is obtained for a query sequence, then 3-D structure neighbors,

crystallization coordinates, and atomic coordinates and a sequence alignment can help determine the identity of the gene of interest. Next step is to observe what the protein of interest is involved in. It displays the E.C. numbers of the proteins involved as well as their reactants and products. Next, ExPASy was used to confirm the E.C. number for the enzymes. (http://www.expasy.ch). The next step was to determine if the protein had any paralogs. This was done using the KEGG pathway map, which has information on the paralog if there is one.

If the two putative orthologs are part of an operon that has similar components, then this is strong evidence of their functional similarity. To determine of our genes of interest were part of an operon, we used the IMG/EDU Gene Neighborhood tool, which is linked to the Gene Details page through the JGI's IMG platform. Numerous species can be viewed and if there is what appears to be a common order of linked genes conserved across species, then it can be assumed that the gene of interest is part of an operon.

To determine if horizontal gene transfer has occurred in the evolutionary history of our genes of interest, we applied several programs in the final GENI-ACT module. Horizontal gene transfer is the transfer of a gene from on organism to another that is not its offspring (http://www.gene-act.org). When this occurs, it can give a new cellular ability to an organism that didn't previously have it or it can cause the new gene to evolve a different function over time (Podell, 2007). The Phylogeny.fr (citation) tool was used to determine if horizontal gene transfer has occurred or not as well as phylums. If the phylum is the same across the different species most closely related to the M. ruber gene, then it suggests that horizontal gene transfer did not occur. Another way to observe whether or not horizontal gene transfer has occurred is to observe the guanine-cytosine (GC) map. This compares the average guanine-cytosine percentage across various species to the guanine-cytosine percentage of the gene of interest (http://www.gene-act.org). If they are vastly different it suggests horizontal gene transfer has occurred. This information can be collected using the IMG/EDU Gene Finder tool yet again.

The question of this project was, are *M. ruber* genes Mrub_2052, Mrub_0628, and Mrub_2034 orthologous to b0688, b2039, and b3789, respectively? Based on the sequence similarity between the *M. ruber* genes and the *E. coli* genes a hypothesis was formed in regards to this question. The next step was to carry out the same analysis as was done to the *E. coli* genes on the *M. ruber* genes using the various bioinformatics tools explained previously in this section. The results can then be interpreted and compared to confirm or deny if the hypothesis was true. The results of this finding are shared with the *M. ruber* community through this paper (http://www.geni-science.org).

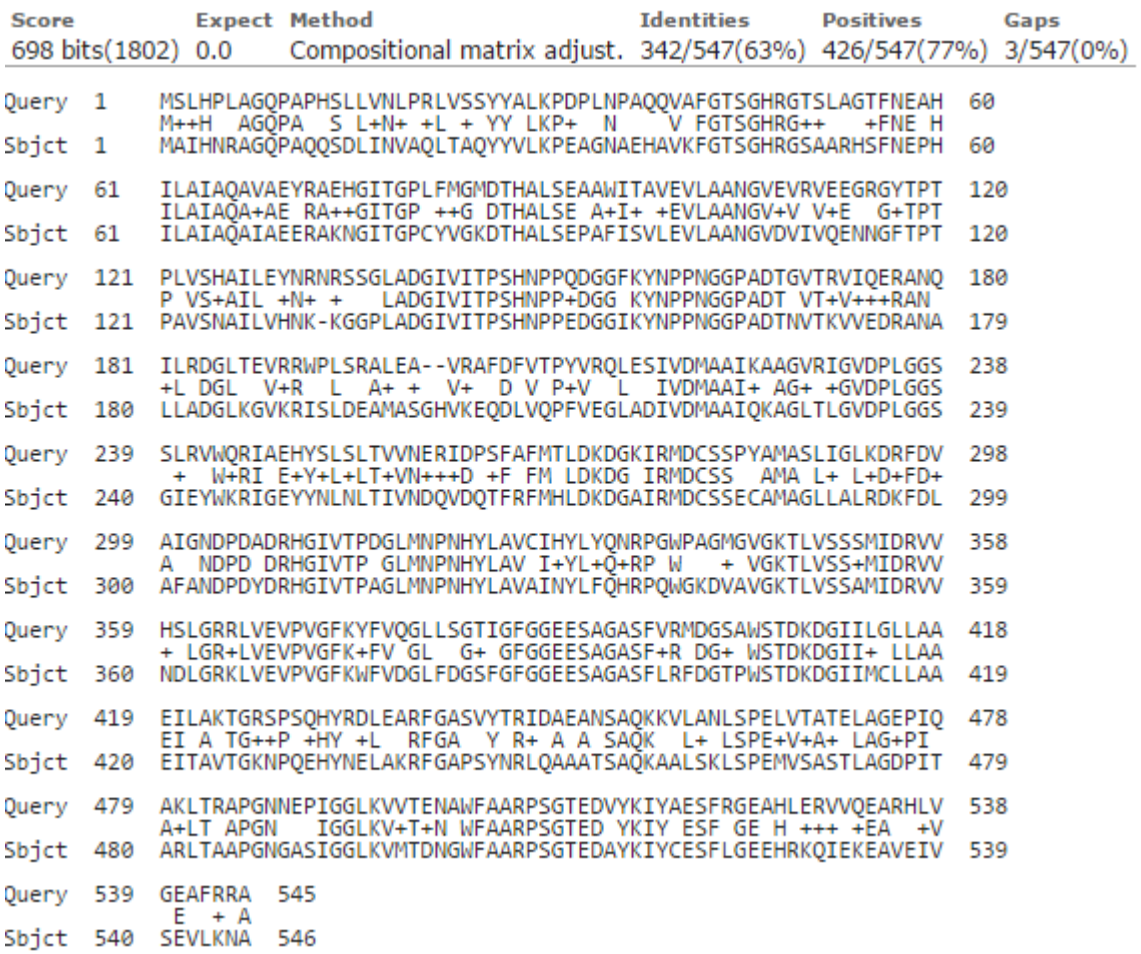
Results

The first two genes to be compared to determine whether or not they are orthologous were Mrub_2052 and b0688. Table 1 is a summary of the results of a variety of different bioinformatics tools that conclude that the *M. ruber* gene Mrub_2052 and *E. coli* gene b0688 are orthologous and report the final prediction of the identity of the *M. ruber* gene.

Table 1: Mrub_2052 and b0688 might be orthologous to one another

Bioinformatics tools used	***M. ruber* Mrub_2052**	***E. coli* b0688**
BLAST *E.coli* against *M. ruber*	Score: 698 bits E-value: 0.0	
CDD Data (COG category)	COG number: COG0033 Phosphoglucomutase	
	E-value: 0.0	E-value: 0.0
Cellular Localization	Cytoplasm of the cell	
TIGRFAM – protein family	TIGRFAM number: TIGR01132 (phosphoglucomutase, alpha-D-glucose)	
	Score: 1303.4 E-value: 0.0	Score: 1631.1 E-value: 0.0
Pfam – protein family	Pfam number: PF02878 (Phosphoglucomutase/phosphomannomutase, alpha/beta/alpha domain I) Pfam number: PF02880 (Phosphoglucomutase/phosphomannomutase, alpha/beta/alpha domain III)	
	E- values: 1.4e-36 4.4e-29	E- values: 9.2e-36 6.1e-31
PDB – protein database	2FUV phosphoglucomutase	
	E-value: 0.0	E-value: 0.0
Enzyme commission number – E.C. number	E.C. 5.4.2.2 phosphoglucomutase (alpha-D-glucose-1,6-bisphosphate-dependent)	
KEGG pathway map	Streptomycin Biosynthesis	
Identity	Phosphoglucomutase	

The first bioinformatics tool listed in the table is the protein BLAST of *E.coli* against *M. ruber*. A bit score and E-value of the alignment is listed. Recall, a large bit score and small E-value suggests that the two sequences are very similar. As can be seen in table 1, the BLAST alignment of Mrub_2052 against b0688 shows a high bit score and a low E-value (If an E- value is smaller than 10^{-100}, it is sometimes given as 0.0). This means the two genes have a very similar amino acid sequence and that the alignment isn't due to chance. A more detailed description of the BLAST results can be seen in Figure 4.



Score	Expect	Method	Identities	Positives	Gaps
698 bits(1802)	0.0	Compositional matrix adjust.	342/547(63%)	426/547(77%)	3/547(0%)

```
Query  1    MSLHPLAGQPAPHSLLVNLPRLVSSYYALKPDPLNPAQQVAFGTSGHRGTSLAGTFNEAH  60
            M++H  AGQPA  S L+N+ +L + YY LKP+  N    V FGTSGHRG++   +FNE H
Sbjct  1    MAIHNRAGQPAQQSDLINVAQLTAQYYVLKPEAGNAEHAVKFGTSGHRGSAARHSFNEPH  60

Query  61   ILAIAQAVAEYRAEHGITGPLFMGMDTHALSEAAWITAVEVLAANGVEVRVEEGRGYTPT  120
            ILAIAQA+AE RA++GITGP ++G DTHALSE A+I+ +EVLAANGV+V V+E  G+TPT
Sbjct  61   ILAIAQAIAEERAKNGITGPCYVGKDTHALSEPAFISVLEVLAANGVDVIVQENNGFTPT  120

Query  121  PLVSHAILEYNRNRSSGLADGIVITPSHNPPQDGGFKYNPPNGGPADTGVTRVIQERANQ  180
            P VS+AIL +N+ +   LADGIVITPSHNPP+DGG KYNPPNGGPADT VT+V+++RAN
Sbjct  121  PAVSNAILVHNK-KGGPLADGIVITPSHNPPEDGGIKYNPPNGGPADTNVTKVVEDRANA  179

Query  181  ILRDGLTEVRRWPLSRALEA--VRAFDFVTPYVRQLESIVDMAAIKAAGVRIGVDPLGGS  238
            +L DGL  V+R  L  A+ +  V+  D V P+V  L  IVDMAAI+ AG+ +GVDPLGGS
Sbjct  180  LLADGLKGVKRISLDEAMASGHVKEQDLVQPFVEGLADIVDMAAIQKAGLTLGVDPLGGS  239

Query  239  SLRVWQRIAEHYSLSLTVVNERIDPSFAFMTLDKDGKIRMDCSSPYAMASLIGLKDRFDV  298
             +  W+RI E+Y+L+LT+VN+++D +F FM LDKDG IRMDCSS  AMA L+ L+D+FD+
Sbjct  240  GIEYWKRIGEYYNLNLTIVNDQVDQTFRFMHLDKDGAIRMDCSSECAMAGLLALRDKFDL  299

Query  299  AIGNDPDADRHGIVTPDGLMNPNHYLAVCIHYLYQNRPGWPAGMGVGKTLVSSSMIDRVV  358
            A  NDPD DRHGIVTP GLMNPNHYLAV I+YL+Q+RP W   + VGKTLVSS+MIDRVV
Sbjct  300  AFANDPDYDRHGIVTPAGLMNPNHYLAVAINYLFQHRPQWGKDVAVGKTLVSSAMIDRVV  359

Query  359  HSLGRRLVEVPVGFKYFVQGLLSGTIGFGGEESAGASFVRMDGSAWSTDKDGIILGLLAA  418
            + LGR+LVEVPVGFK+FV GL  G+ GFGGEESAGASF+R DG+ WSTDKDGII+ LLAA
Sbjct  360  NDLGRKLVEVPVGFKWFVDGLFDGSFGFGGEESAGASFLRFDGTPWSTDKDGIIMCLLAA  419

Query  419  EILAKTGRSPSQHYRDLEARFGASVYTRIDAEANSAQKKVLANLSPELVTATELAGEPIQ  478
            EI A TG++P +HY +L  RFGA  Y R+ A A SAQK  L+ LSPE+V+A+ LAG+PI
Sbjct  420  EITAVTGKNPQEHYNELAKRFGAPSYNRLQAAATSAQKAALSKLSPEMVSASTLAGDPIT  479

Query  479  AKLTRAPGNNEPIGGLKVVTENAWFAARPSGTEDVYKIYAESFRGEAHLERVVQEARHLV  538
            A+LT APGN   IGGLKV+T+N WFAARPSGTED YKIY ESF GE H +++ +EA  +V
Sbjct  480  ARLTAAPGNGASIGGLKVMTDNGWFAARPSGTEDAYKIYCESFLGEEHRKQIEKEAVEIV  539

Query  539  GEAFRRA  545
             E  + A
Sbjct  540  SEVLKNA  546
```

Figure 4. Blast amino acid alignment of *M. ruber* Mrub_2052 and *E.coli* b0688. Mrub_2052 is the query sequence and b0688 is the subject sequence. This analysis was performed using NCBI BLAST bioinformatics tool at http://blast.ncbi.nlm.nih.gov.

Out of the total 547 amino acids 342 of them are conserved (63%). The bit score of 698 is fairly high suggesting the two sequences are related. E-values smaller than 10^{-100}, it is sometimes given as 0.0. This alignment has an E-value of 0.0 which is very small suggesting that the sequence is not conserved due to chance. This was the first piece of information suggesting the genes are orthologous.

The next bioinformatics tool listed was the CDD. Recall, if the *E. coli* gene and the *M. ruber* gene of interest both belong to the same COG group it suggests that they most likely are orthologous to one another. As can be seen in the table both genes belong to the COG family COG0033. Both *M. ruber* and *E. coli* have very small E-values in relation to the family suggesting that they aren't related due to chance. This was another piece of evidence suggesting the two genes are orthologous to one another.

The next column on the table represents the predicted cellular location of the proteins coded by the *M. ruber* and *E. coli* gene. This prediction was made by using bioinformatics tools such as

TMHMM, SignalP, LipoP, PSORT-B, and Phobius. Figure 5 is a visual representation of the TMHMM results for both Mrub_2052 and b0688.

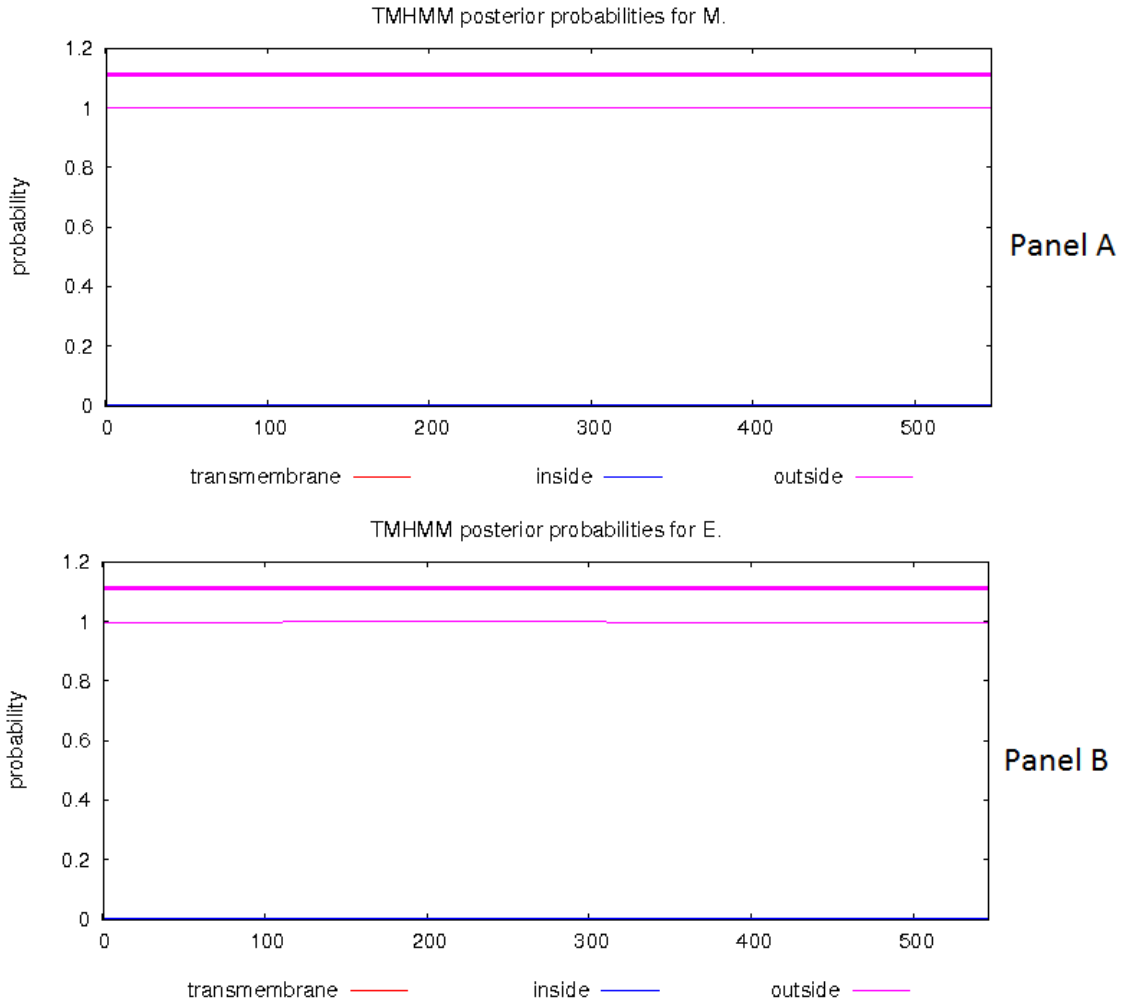


Figure 5. TMHMM transmembrane helices graph comparison of *M. ruber* Mrub_2052 and *E. coli* b0688 suggesting there are no transmembrane helices for either protein and that the protein is most likely located outside of the cell (cytoplasm). Panel A is the TMHMM transmembrane helices graph for *M. ruber* Mrub_2052 and panel B is the TMHMM transmembrane helices graph for *E. coli* b0688. TMHMM Server v. 2.0 found at http://www.cbs.dtu.dk/services/TMHMM was used to create these graphs.

In Figure 5 it can be seen that there are no red peaks in either panel A or B (panel A representing Mrub_2052 and panel B representing b0688). This suggests that there are no transmembrane helices for either protein these genes code for. What these graph do tell us though is that there is a high probability that the proteins that these genes code for are located outside of the membrane, which would be the cytoplasm. This was one of the pieces of information suggesting that both proteins are located in the cytoplasm and that the genes that code for them are orthologous. The next tool used was SignalP. Recall SignalP is used to determine whether or not the predicted

transmembrane helices are actually that or if they are signal peptides. A visual representation of this data can be seen in Figure 6.

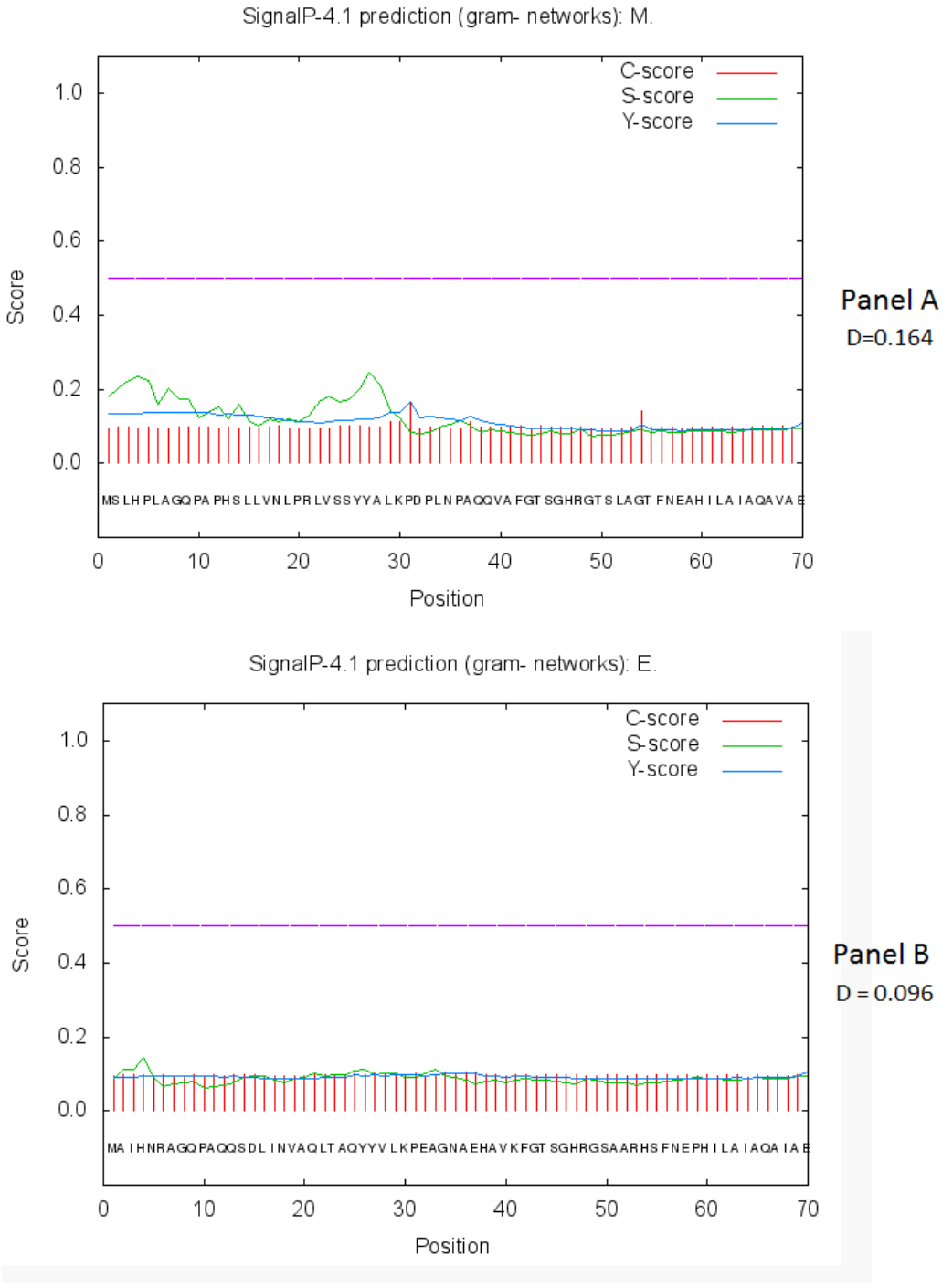


Figure 6. SignalP graphical representation of *M. ruber* Mrub_2052 and *E. coli* b0688 suggesting there are no transmembrane helices or signal peptides. Panel A represents *M. ruber* Mrub_2052 and panel B represents *E. coli* b0688. D values are located under the panel names on the graph. SignalP server v. 4.1 http://www.cbs.dtu.dk/services/SignalP was used to create these plots.

In Figure 6, there are four notable pieces of information that can be used to determine signal peptides. Those include the C-score, which distinguishes signal peptide cleavage sites from everything else, the S-score, which distinguishes the signal peptide position, the Y-score, which is a combined score of the C and S score, and the D value which it the probability that there is a signal peptide. On both panels A and B (panel A representing Mrub_2052 and panel B representing b0688) the C, S, and Y scores are very low. This makes sense because, as was suggested in Figure 5, there are no transmembrane helices that might be a signal peptide. The D value for both panel A and B is very low also confirming that there are no signal peptides. The fact that both Mrub_2052 and b0688 do not have transmembrane helices or signal peptides suggest that they may be orthologous.

After the bioinformatics tools TMHMM and SignalP were used LipoP, PSORT-B, and Phobius were also used. LipoP predicted that the proteins coded by Mrub_2052 and b0688 are found in the cytoplasm. PSORT-B confirmed this by attributing its highest score to the cytoplasm. Since Phobius is a visual representation of the results of TMHMM and SignalP and there were no results for either of these tools, the phobius was not useful. These tools suggested that both the protein coded for by Mrub_2052 and b0688 are found in the cytoplasm of the cell. This was another piece of evidence suggesting these genes were orthologous.

The next column displays what TIGRFAM protein family (domain) they are most closely related to. Notice that both *M. ruber* and *E. coli* are related to the same family with very low E-values. This family suggests that the identity of both proteins that Mrub_2052 and b0688 code for is most likely phosphoglucomutase. This again suggests they are orthologous to one another. The next column is the Pfam protein family. On the table it can be seen that both Mrub_2052 and b0688 are related to two separate protein families with low E-values. A visual representation of the comparison the top Pfam protein family to the gene sequences can be seen in Figure 7.

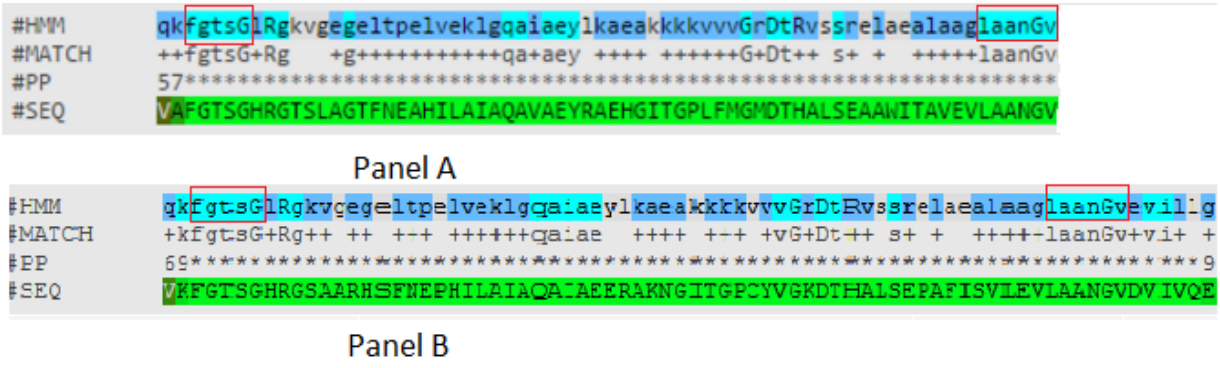


Figure 7. Pfam protein family PF02878 comparison to *M. ruber* Mrub_2052 and *E. coli* b0688 displaying conserved amino acids. Panel A represents *M. ruber* Mrub_2052 and panel B represents *E. coli* b0688. The red boxes display the similar conserved amino acids to PF02878. This pairwise alignment was created using the Pfam website http://pfam.sanger.ac.uk/search.

Figure 7 shows the comparison of both Mrub_2052 and b0688 to the protein family amino acid sequence. As can be noted by the red boxes in the Figure there are a number of similarities in conserved amino acids. This further concludes that they are highly related and also suggests that

the protein these genes code for is phosphoglucomutase. The next column is the PDB. Both genes were suggested to be phosphoglucomutase. The E.C. number also confirmed this identity prediction.

The second to last column on the table is the KEGG pathway data, which tell us what processes the proteins are involved in. Both are involved in streptomycin biosynthesis. A graphic of this process can be seen in Figure 8.

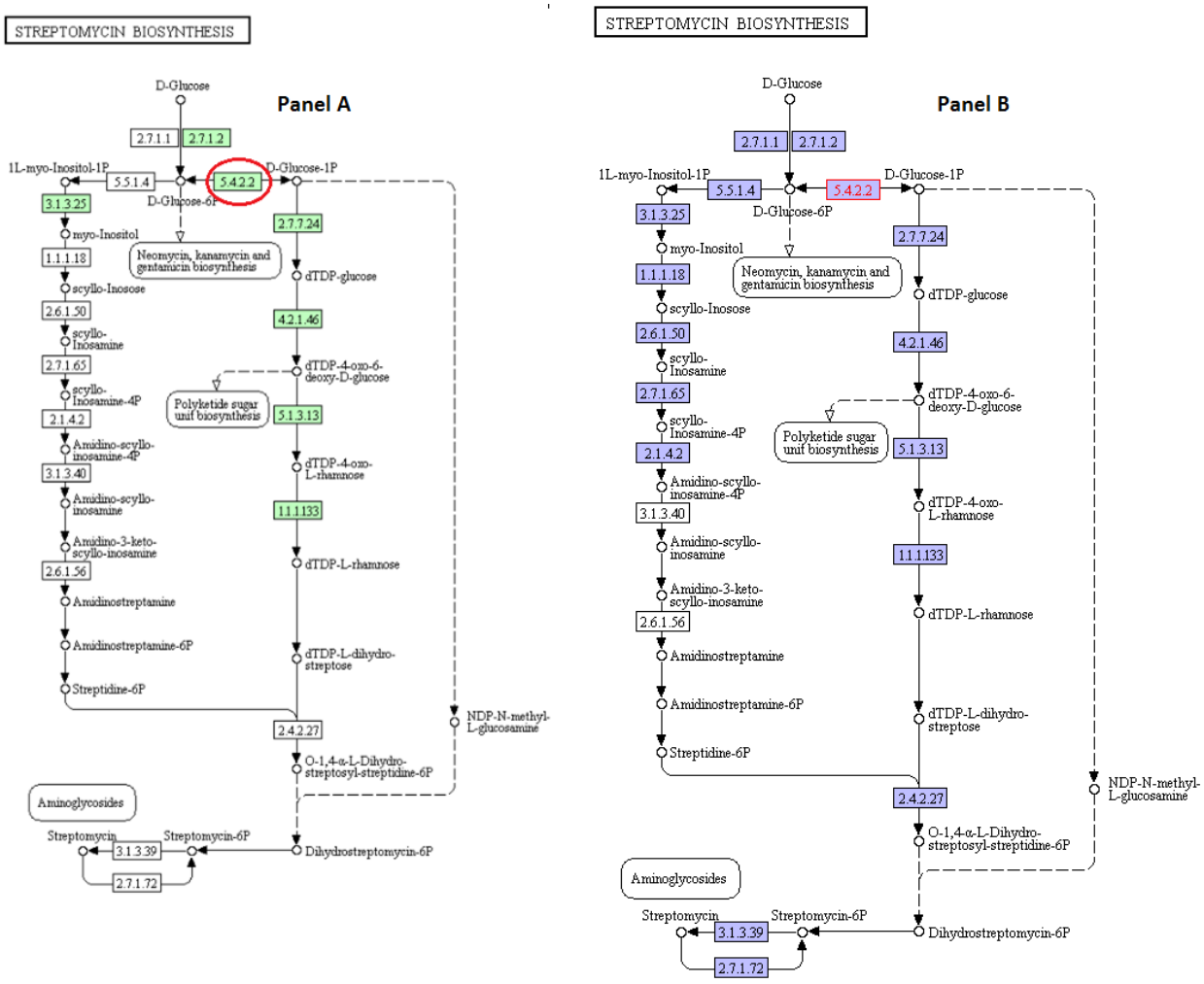


Figure 8. Enzyme E.C. number 5.4.2.2 presence in the streptomycin biosynthesis pathway. The protein coded by Mrub_2052 and b0688 is represented by E.C. number 5.4.2.2. The Kyto Encyclodpedia of Genes and Genomes (KEGG) data base was used to create this map at http://www.genome.jp/kegg/pathway.html

Figure 8 displays the entirety of the streptomycin biosynthesis pathway. The enzymes (E.C numbers) colored green can be found in *E. coli* and *M. ruber*. The red circle highlights E.C. number 5.4.2.2 which represents Mrub_2052 and b0688. As can be seen in the Figure this enzyme is suggested to convert converts D-Glucose- 6P (also known as D-glucopyranose 6-phosphate) to D-Glucose-1P (also known as α-D-glucopyranose 1-phosphate). The keg pathway map was also used to determine if there are any known paralogs for the enzyme in both *M. ruber* and *E. coli*. It was determined that there are no paralogs for either. This evidence also suggests Mrub_2052 and b0688 are orthologous.

Some additional information that was collected that is not included in table 1 is a comparison of the Mrub_2052 and b0688 gene neighborhood maps. In Figure 9 this comparison can be seen.

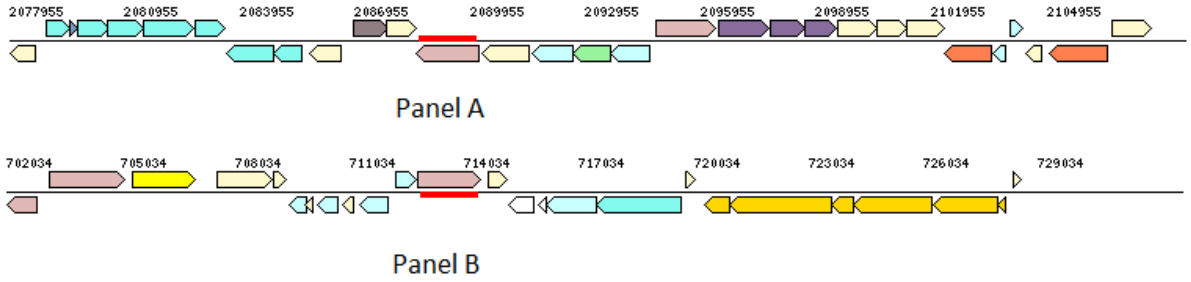


Figure 9. Comparison of *M. ruber* Mrub_2052 and *E. coli* b0688 gene neighborhood maps and the unlikelihood of being in an operon. Panel A represents *M. ruber* Mrub_2052 and panel B represents *E. coli* b0688. Both are underlined by a red line. Images were taken from http://img.jgi.doe.gov/.

Each gene in a gene neighborhood map is represented by an arrow. The function of that gene is represented by the color of the arrow. If there are numerous genes that are the same color pointing in the same direction it means that they are part of an operon. The genes in both *M. ruber* and *E. coli* (underlined by a red line) don't appear to be a part of an operon. The color of both genes is the same though. This suggests that both have similar function, further proving that the two are orthologous to one another.

Another piece of additional information that was collected that is not included in table 1 is a comparison of the Mrub_2052 and b0688 phylogenetic trees. In Figure 10 this comparison can be seen.

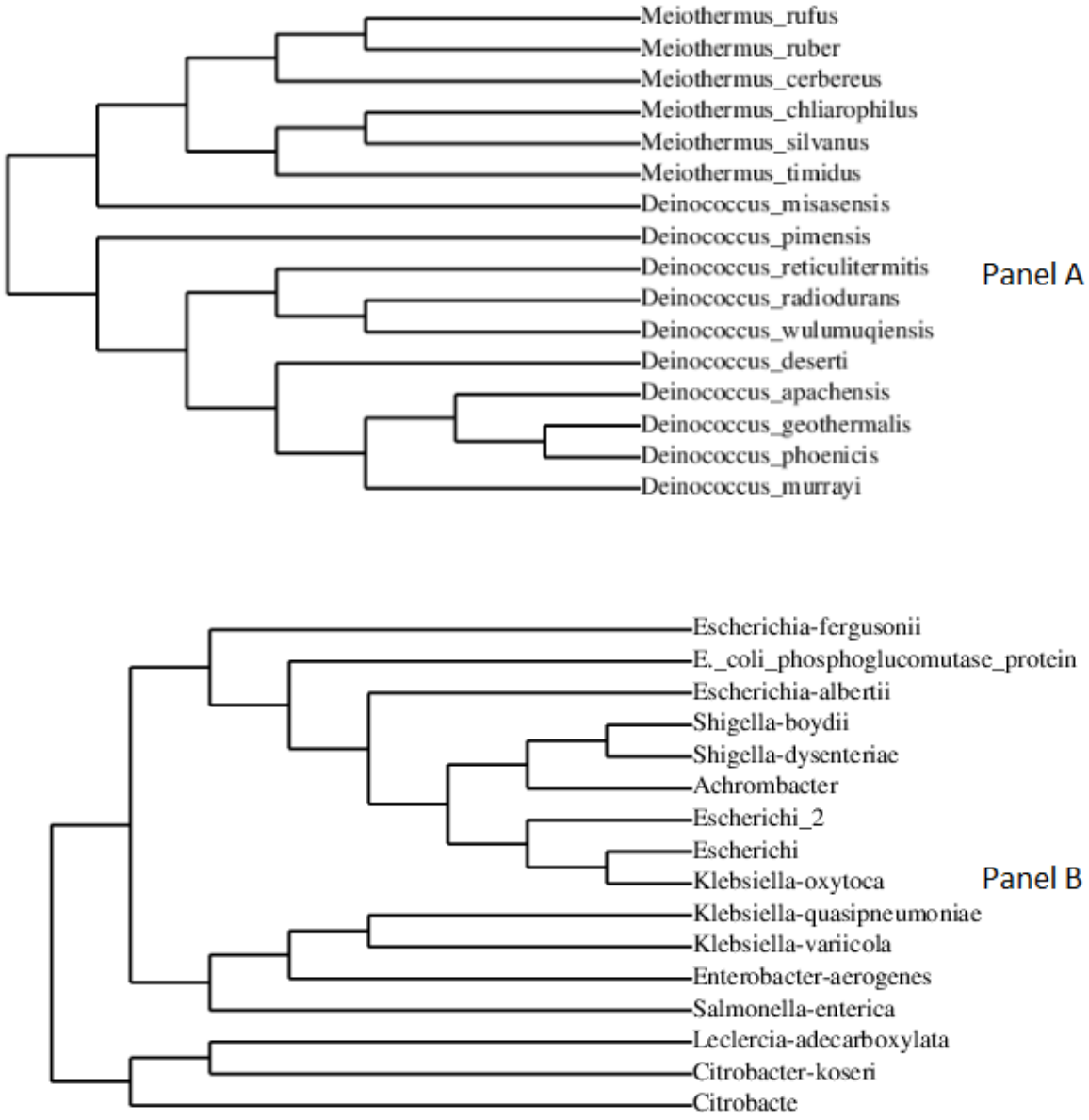


Figure 10. Comparison of *M. ruber* Mrub_2052 and *E. coli* b0688 phylogenic trees suggest horizontal gene transfer did not occur. Panel A represents M. ruber Mrub_2052 and panel B represents E. coli b0688. Images were created using http://www.phylogeny.fr

The images in Figure 10 show the phylogenetic trees of *M. ruber* and *E. coli*. Panel A represents *M. ruber* and panel B represents *E. coli.* All the organisms in panel A are part of the Deinococcus-Thermus phylum. All the organisms in panel B are part of the proteobacteria phylum. Since the organisms close to *M ruber* and *E. coli* in each panel belong to the same phylum there is no evidence that horizontal gene transfer occurred.

Based on the vast amount of information provided by these bioinformatics tools it can be confirmed that the two genes are orthologous and code for the protein phosphoglucomutase.

The next two genes to be compared to determine whether or not they are orthologous were Mrub_0628 and b2039. Table 2 is a summary of the results of a variety of different bioinformatics tools that predict that the *M. ruber* gene Mrub_0628 and *E. coli* gene b2039 might be orthologous and report the final prediction of the identity of the *M. ruber* gene.

Table 2: Mrub_0628 and b2039 are orthologous to one another

Bioinformatics tools used	***M. ruber* Mrub_0628**	***E. coli* b2039**
BLAST *E.coli* against *M. ruber*	Score: 122 bits E-value: 5e-37	
CDD Data (COG category)	COG number: COG1209 glucose-1-phosphate thymidyltransferase	
	E-value: 1.19e-117	E-value: 1.12e-175
Cellular Localization	Cytoplasm of the cell	
TIGRFAM – protein family	TIGRFAM number: TIGR01208 (glucose-1-phosphate thymidylyltransferase)	
	Score: 614.5 E-value: 1.4e-181	Score: 658.1 E-value: 1e-194
Pfam – protein family	Pfam number: PF00483 (NTP_transferase)	
	E- values: 3.2e-48	E- values: 5.1e-74
PDB – protein database	3HL3 Glucose-1-Phosphate Thymidylyltransferase from *Bacillus anthracis* in Complex with a Sucrose	1H5R thymidylyltransferase complexed with thymidine and glucose-1-phospate
	E-value: 4.21e-39	E-value: 4.41e-173
Enzyme commission number – E.C. number	E.C. 2.7.7.24 Glucose-1-phosphate thymidylyltransferase	
KEGG pathway map	Streptomycin Biosynthesis	
Identity	Glucose-1-phosphate thymidylyltransferase	

The first bioinformatics tool listed in the table is the protein BLAST of *E.coli* against *M. ruber*. A bit score and E-value of the alignment is listed. Recall, a large bit score and small E-value suggests that the two sequences are very similar. As can be seen in table 2, the BLAST alignment of Mrub_0628 against b2039 shows a high bit score and a low E-value. This means the two genes have a very similar amino acid sequence and that the alignment isn't due to chance. A more detailed description of the BLAST results can be seen in Figure 11.

Score	Expect	Method	Identities	Positives	Gaps
122 bits(307)	5e-37	Compositional matrix adjust.	81/236(34%)	126/236(53%)	6/236(2%)

```
Query  6    KGLILAAGRGTRLRPLTHTRPKPVIRLAGKPIIRYAVDNLLEAGITEIGVVVSPDTIEDI  65
            KG+ILA G GTRL P+T    K ++ +  KP+I Y +  L+ AGI +I ++ +P
Sbjct  5    KGIILAGGSGTRLYPVTMAVSKQLLPIYDKPMIYYPLSTLMLAGIRDILIISTPQDTPRF  64

Query  66   KLALKDCS--GVQITYIVQEEALGIAHAVGTAKDWLGQSPFVLYLGDNLFQ-KGVKSFVE  122
            +  L D S  G+ + Y VQ    G+A A    ++++G     L LGDN+F    +   +E
Sbjct  65   QQLLGDGSQWGLNLQYKVQPSPDGLAQAFIIGEEFIGGDDCALVLGDNIFYGHDLPKLME  124

Query  123  AYQPGIS-AVIALVRVPDPRQFGVAVLEE-GRIVKLLEKPKNPPSDLAVAGVYVFGPVIM  180
            A     S A +    V DP ++GV   ++ G  + L EKP  P S+ AV G+Y +   ++
Sbjct  125  AAVNKESGATVFAYHVNDPERYGVVEFDKNGTAISLEEKPLEPKSNYAVTGLYFYDNDVV  184

Query  181  DIIANLKPSARGEYEITDAIQALVDRGHTVLGQEIAGW-WKDTGRPADLLDANRLL  235
             +  NLKPSARGE EITD  +  +++G   +     G+ W DTG    L++A+  +
Sbjct  185  QMAKNLKPSARGELEITDINRIYLEQGRLSVAMMGRGYAWLDTGTHQSLIEASNFI  240
```

Figure 11. Blast amino acid alignment of *M. ruber* Mrub_0628 and *E.coli* b2039. Mrub_0628 is the query sequence and b2039 is the subject sequence. This analysis was performed using NCBI BLAST bioinformatics tool at http://blast.ncbi.nlm.nih.gov.

Out of the total 236 amino acids 81 of them are conserved (34%). The bit score of 122 is somewhat high suggesting the two sequences are related. This alignment has an E-value of 5e-37 which is very small suggesting that the sequence is not conserved due to chance. Since only 34% of the sequence was conserved more evidence had to be collected to confirm or deny that Mrub_0628 and b2039 are orthologous.

The next bioinformatics tool listed was the CDD. Recall, if the *E. coli* gene and the *M. ruber* gene of interest both belong to the same COG group it suggests that they most likely are orthologous to one another. As can be seen in the table both genes belong to the COG family COG1209. Both *M. ruber* and *E. coli* have very small E-values in relation to the family suggesting that they aren't related due to chance. This was another piece of evidence suggesting the two genes are orthologous to one another.

The next column on the table represents the predicted cellular location of the proteins coded by the *M. ruber* and *E. coli* gene. This prediction was made by using bioinformatics tools such as TMHMM, SignalP, LipoP, PSORT-B, and Phobius. Figure 12 is a visual representation of the TMHMM results for both Mrub_0628 and b2039.

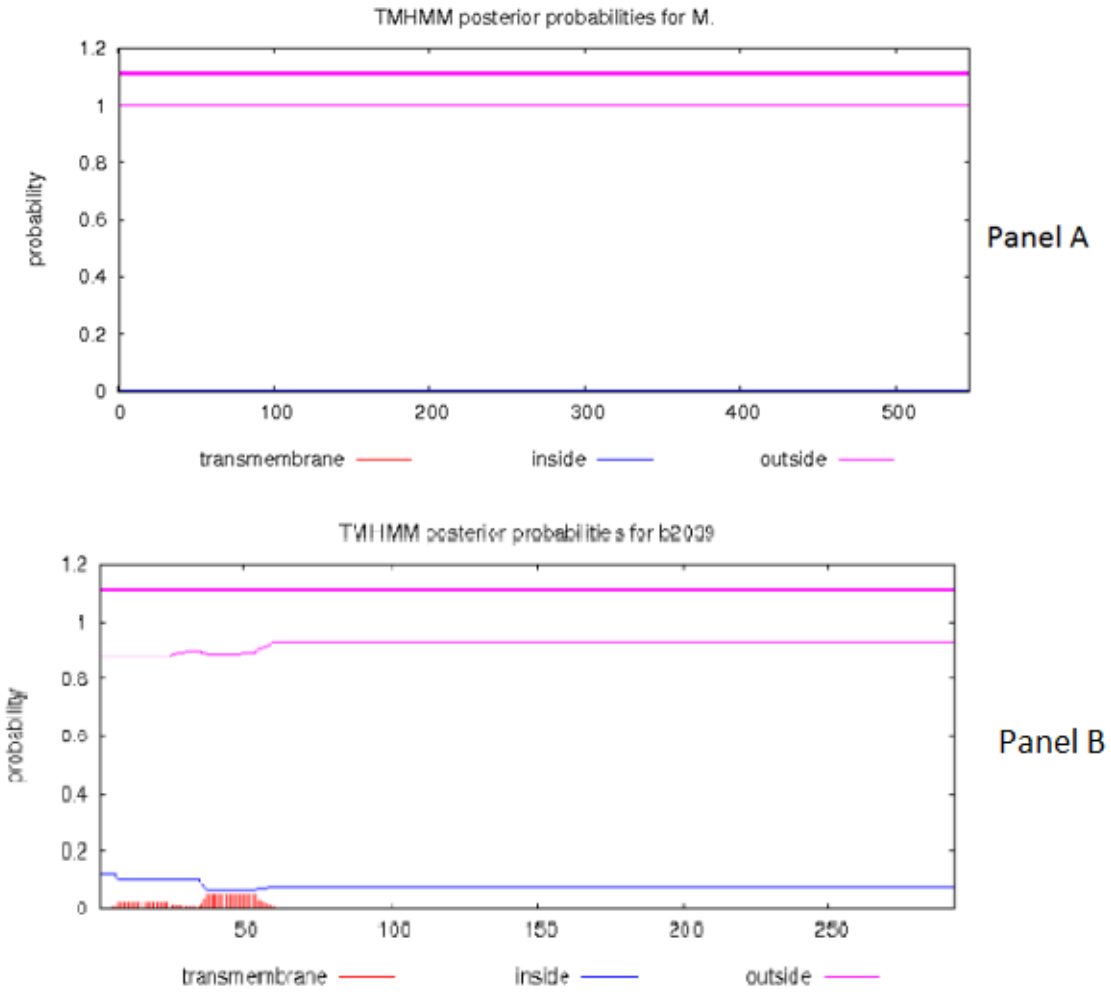


Figure 12. TMHMM transmembrane helices graph comparison of *M. ruber* Mrub_0628 and *E. coli* b2039 suggesting there are no transmembrane helices for either protein and that the protein is most likely located outside of the cell (cytoplasm). Panel A is the TMHMM transmembrane helices graph for *M. ruber* Mrub_0628 and panel B is the TMHMM transmembrane helices graph for *E. coli* b2039. TMHMM Server v. 2.0 found at http://www.cbs.dtu.dk/services/TMHMM was used to create these graphs.

In Figure 12 it can be seen that there are no red peaks in panel A (panel A representing Mrub_0628 and panel B representing b2039). In panel B there are some visible red peaks that typically would suggest that part of the protein has a transmembrane helix, but the probability is so low it is not likely. This suggests that there are no transmembrane helices for either protein these genes code for. What these graph do tell us though is that there is a high probability that the proteins that these genes code for are located outside of the membrane, which would be the cytoplasm. This was one of the pieces of information suggesting that both proteins are located in the cytoplasm and that the genes that code for them are orthologous. The next tool used was SignalP. Recall SignalP is used to determine whether or not the predicted transmembrane helices are actually that or if they are signal peptides. A visual representation of this data can be seen in Figure 13.

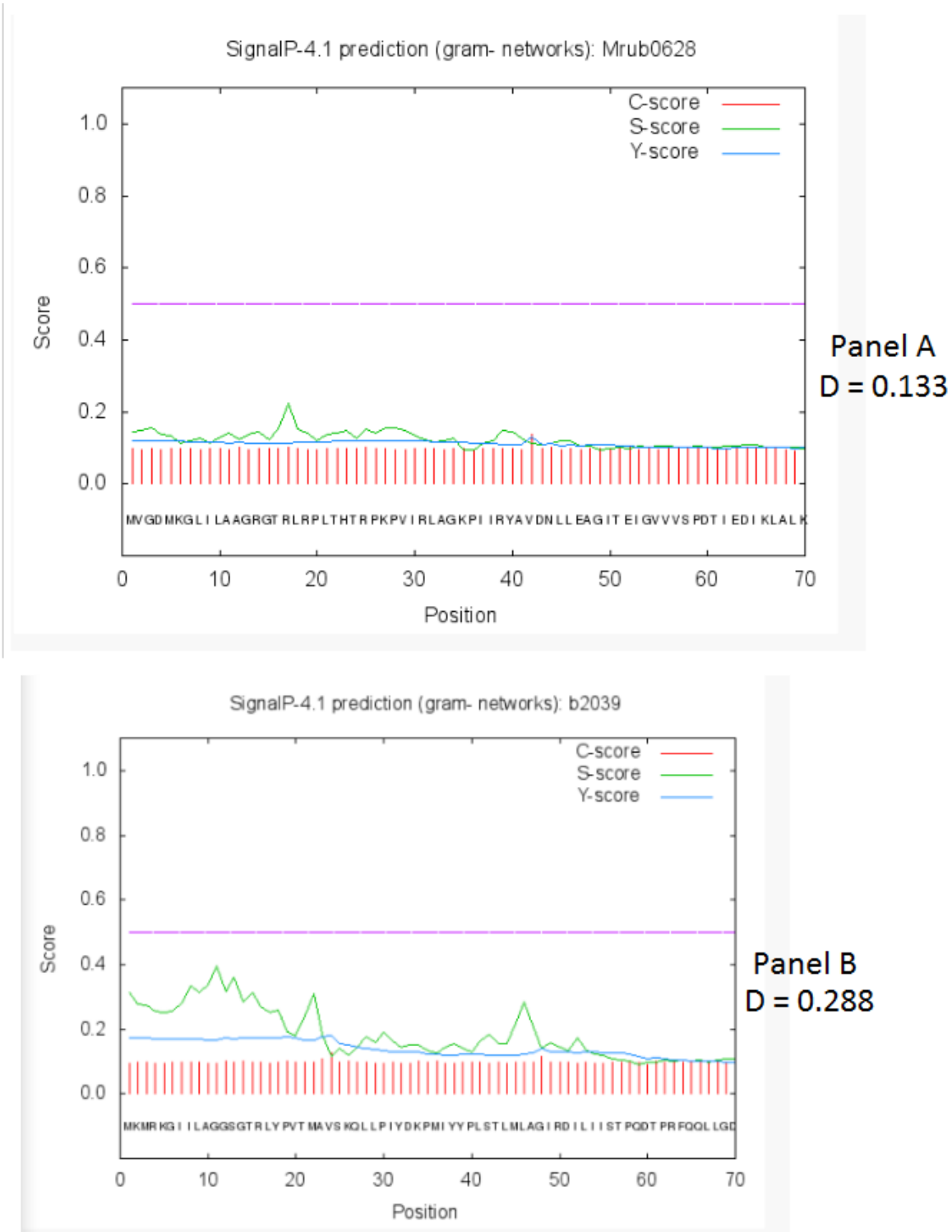


Figure 13. SignalP graphical representation of *M. ruber* Mrub_0628 and *E. coli* b2039 suggesting there are no transmembrane helices or signal peptides. Panel A represents *M. ruber* Mrub_0628 and panel B represents *E. coli* b2039. D values are located under the panel names on the graph. SignalP server v. 4.1 http://www.cbs.dtu.dk/services/SignalP was used to create these plots.

In Figure 13, there are four notable pieces of information that can be used to determine signal peptides. Those include the C-score, which distinguishes signal peptide cleavage sites from everything else, the S-score, which distinguishes the signal peptide position, the Y-score, which is a combined score of the C and S score, and the D value which it the probability that there is a signal peptide. On both panels A and B (panel A representing Mrub_0628 and panel B representing b2039) the C, S, and Y scores are very low. This makes sense because, as was suggested in Figure 12, there are no transmembrane helices that might be a signal peptide. The D value for both panel A and B is very low also confirming that there are no signal peptides. The fact that both Mrub_0628 and b2039 do not have transmembrane helices or signal peptides suggest that they may be orthologous.

After the bioinformatics tools TMHMM and SignalP were used LipoP, PSORT-B, and Phobius were also used. LipoP predicted that the proteins coded by Mrub_0628 and b2039 are found in the cytoplasm. PSORT-B confirmed this by attributing its highest score to the cytoplasm. Since Phobius is a visual representation of the results of TMHMM and SignalP and there were no results for either of these tools, the phobius was not useful. These tools suggested that both the protein coded for by Mrub_0628 and b2039 are found in the cytoplasm of the cell. This was another piece of evidence suggesting these genes were orthologous.

The next column displays what TIGRFAM protein family (domain) they are most closely related to. Notice that both *M. ruber* and *E. coli* are related to the same family with very low E-values. This family suggests that the identity of both proteins that Mrub_0628 and b2039 code for is most likely glucose-1-phosphate thymidylyltransferase. This again suggests they are orthologous to one another. The next column is the Pfam protein family. On the table it can be seen that both Mrub_0628 and b2039 are related to the same protein family with low E-values. A visual representation of the comparison the top Pfam protein family to the gene sequences can be seen in Figure 14.

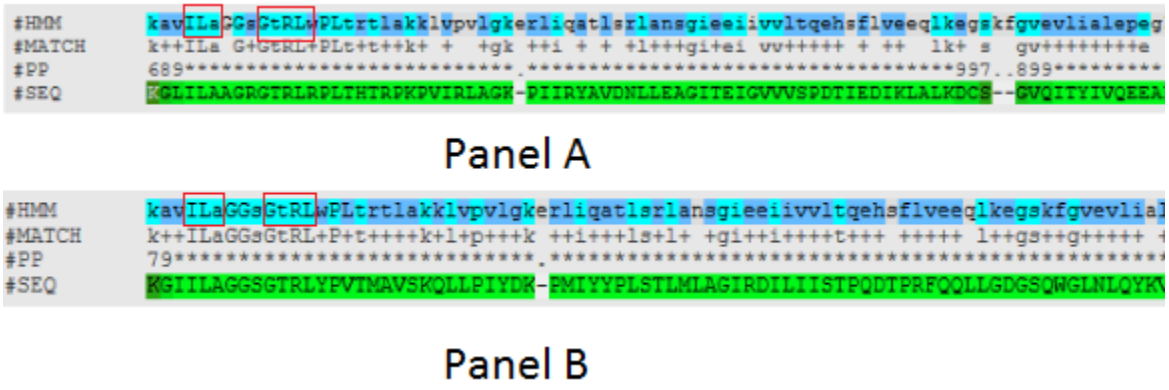


Figure 14. Pfam protein family PF00483 comparison to *M. ruber* Mrub_0628 and *E. coli* b2039 displaying conserved amino acids. Panel A represents *M. ruber* Mrub_0628 and panel B represents *E. coli* b2039. The red boxes display the similar conserved amino acids to PF00483. This pairwise alignment was created using the Pfam website http://pfam.sanger.ac.uk/search.

Figure 14 shows the comparison of both Mrub_0628 and b2039 to the protein family amino acid sequence. As can be noted by the red boxes and the light blue color in the Figure there are a

number of similarities in conserved amino acids. This further concludes that they are highly related and also suggests that the protein these genes code for is glucose-1-phosphate thymidylyltransferase. The next column is the PDB. Despite the fact that Mrub_0628 and b2039 were paired with different PDB database hits both genes were suggested to be glucose-1-phosphate thymidylyltransferase. The E.C. number also confirmed this identity prediction.

The second to last column on the table is what pathway the proteins the genes code for are involved in. Both are involved in streptomycin biosynthesis. A graphic of this process can be seen in Figure 15.

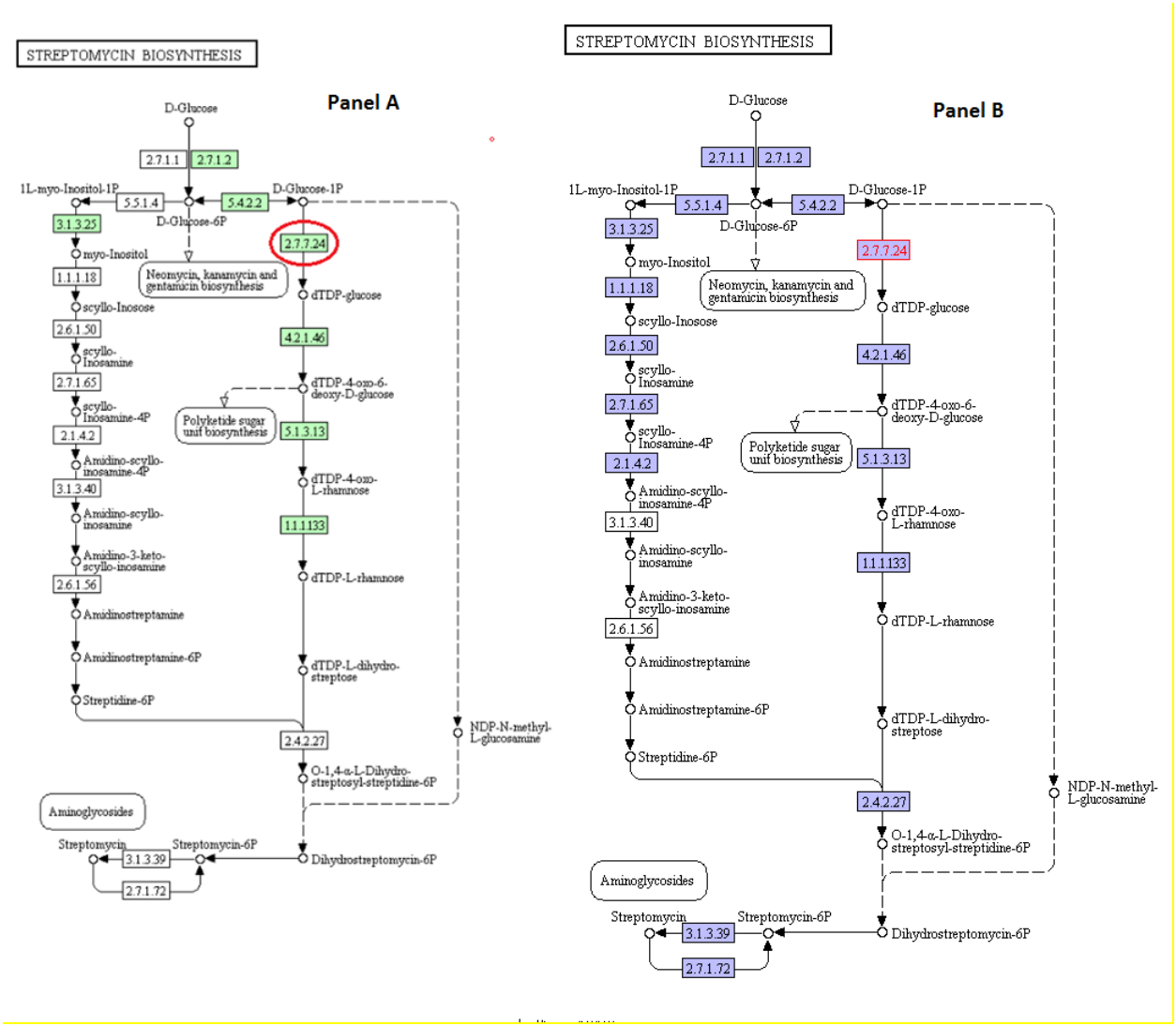


Figure 15. Enzyme E.C. number 2.7.7.24 presence in the streptomycin biosynthesis pathway. The protein coded by Mrub_0628 and b2039 is represented by E.C. number 2.7.7.24. The Kyto Encyclodpedia of Genes and Genomes (KEGG) data base was used to create this map at http://www.genome.jp /kegg/pathway.html

Figure 15 displays the entirety of the streptomycin biosynthesis pathway. The enzymes (E.C numbers) colored green can be found in *E. coli* and *M. ruber*. The red circle highlights E.C. number 2.7.7.24 which represents Mrub_0628 and b2039. As can be seen in the Figure this enzyme is suggested to convert converts D-Glucose-1P to dTDP-glucose (also known as dTDP-α-D-glucose). The KEGG pathway map was also used to determine if there are any known paralogs for the enzyme in both *M. ruber* and *E. coli*. Surprising enough Mrub_0628 is paralogs with Mrub_2034 and b2039 is paralogs with b3789. This is particularly surprising because Mrub_2034 is believed to be orthologous with b3789.

Some additional information that was collected that is not included in table 2 is a comparison of the Mrub_0628 and b2039 gene neighborhood maps. In Figure 16 this comparison can be seen.

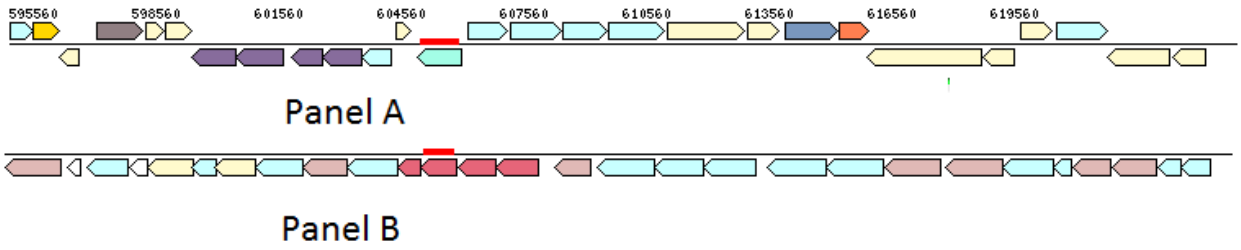
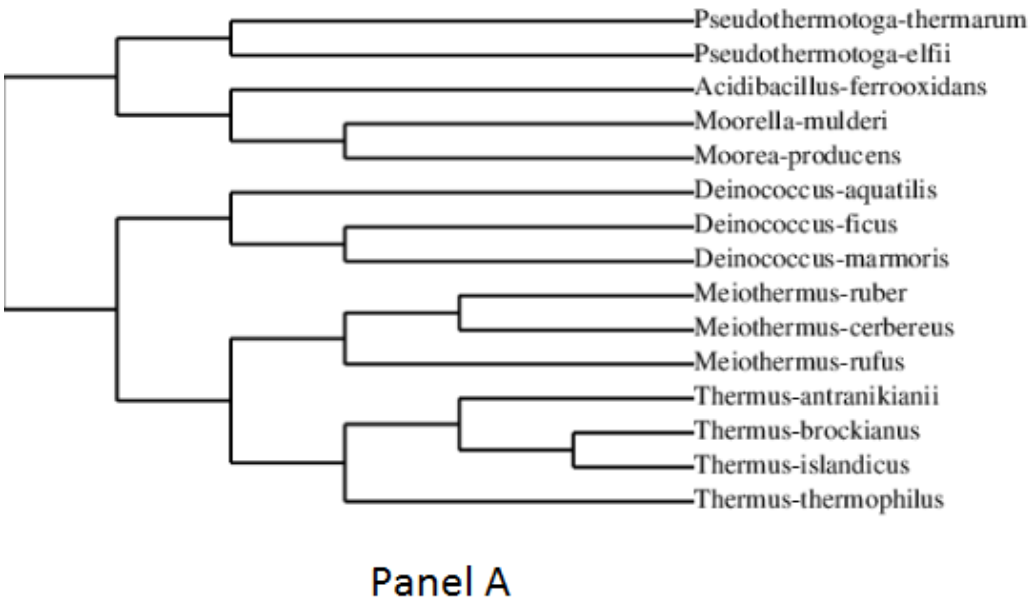


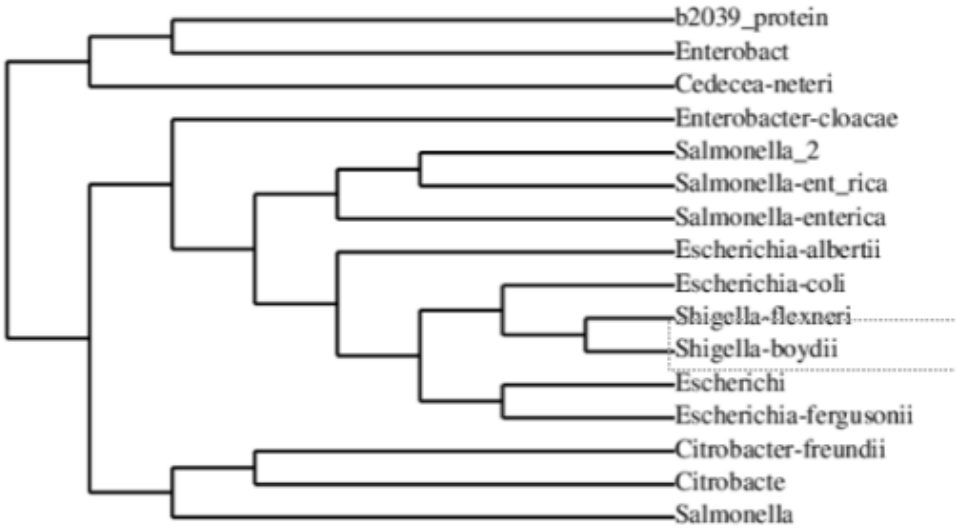
Figure 16. Comparison of *M. ruber* Mrub_0628 and *E. coli* b2039 gene neighborhood maps showing b2039 is part of an operon. Panel A represents *M. ruber* Mrub_0628 and panel B represents *E. coli* b2039. Both are underlined by a red line. Images were taken from http://img.jgi.doe.gov/.

Each gene in a gene neighborhood map is represented by an arrow. The function of that gene is represented by the color of the arrow. If there are numerous genes that are the same color pointing in the same direction it means that they are part of an operon. The gene in *M. ruber* (underlined by a red line) doesn't appear to be a part of an operon. On the other hand the gene in *E.coli* does appear to be in an operon. Strangely enough the two genes are not the same color suggesting that they do not have the same function. This is particularly odd considering all the other information collected up to this point has suggested that these genes are orthologous. This is the first piece of evidence to suggest that they are not.

Another piece of additional information that was collected that is not included in table 2 is a comparison of the Mrub_0628 and b2039 phylogenetic trees. In Figure 17 this comparison can be seen.



Panel A



Panel B

Figure 17. Comparison of *M. ruber* Mrub_0628 and *E. coli* b2039 phylogenic trees suggest horizontal gene transfer did not occur. Panel A represents M. ruber Mrub_0628 and panel B represents E. coli b2039. Images were created using http://www.phylogeny.fr

The images in Figure 17 show the phylogenetic trees of *M. ruber* and *E. coli*. Panel A represents *M. ruber* and panel B represents *E. coli.* All the organisms in panel A are part of the Deinococcus-Thermus phylum. All the organisms in panel B are part of the proteobacteria

phylum. Since the organisms close to *M ruber* and *E. coli* in each panel belong to the same phylum there is no evidence that horizontal gene transfer occurred.

Many of the bioinformatics tools suggested that Mrub_0628 and b2039 are orthologous, but gene neighborhood and PDB suggest otherwise. The identity of these genes has been suggested to be glucose-1-phosphate thymidylyltransferase. Due to the overwhelming amount of evidence suggesting they are it has been predicted that Mrub_0628 and b2039 are orthologous.

The next two genes to be compared to determine whether or not they are orthologous were Mrub_2034 and b3789. Recall that it was previously determined that Mrub_2034 and Mrub_0628 are paralogs as well as b3789 and b2039. Table 3 is a summary of the results of a variety of different bioinformatics tools that predict that the *M. ruber* gene Mrub_2034 and *E. coli* gene b3789 might be orthologous and report the final prediction of the identity of the *M. ruber* gene.

Table 3: Mrub_2034 and b3789 might be orthologous to one another

Bioinformatics tools used	***M. ruber* Mrub_2034**	***E. coli* b3789**
BLAST *E.coli* against *M. ruber*	Score: 160 bits E-value: 2e-51	
CDD Data (COG category)	COG number: COG1209 glucose-1-phosphate thymidylyltransferase	
	E-value: 1.19e-44	E-value: 3.31e-31
Cellular Localization	Cytoplasm of the cell	
TIGRFAM – protein family	TIGRFAM number: TIGR01208 (glucose-1-phosphate thymidylyltransferase)	
	Score: 643.4 E-value: 2.8e-190	Score: 647.6 E-value: 1.5e-191
Pfam – protein family	Pfam number: PF00483 (NTP_transferase)	
	E- values: 1.1e-56	E- values: 1.5e-72
PDB – protein database	5IDS Glucose-1-phosphate Thymidylyltransferase from Burkholderia vietnamiensis	
	E-value: 1.64572e-44	E-value: 1.54322e-114
Enzyme commission number – E.C. number	E.C. 2.7.7.24 Glucose-1-phosphate thymidylyltransferase	
KEGG pathway map	Streptomycin Biosynthesis	
Identity	Glucose-1-phosphate thymidylyltransferase	

The first bioinformatics tool listed in the table is the protein BLAST of *E.coli* against *M. ruber*. A bit score and E-value of the alignment is listed. Recall, a large bit score and small E-value suggests that the two sequences are very similar. As can be seen in table 3, the BLAST alignment of Mrub_2034 against b3789 shows a high bit score and a low E-value. This means

the two genes have a very similar amino acid sequence and that the alignment isn't due to chance. A more detailed description of the BLAST results can be seen in Figure 18.

Score	Expect	Method	Identities	Positives	Gaps
160 bits(406)	2e-51	Compositional matrix adjust.	94/244(39%)	134/244(54%)	2/244(0%)

```
Query  3    LKGLILSGGKGTRLRPLTYTRAKQLIPIAGKPNLFYALEDLLEAGIRDIGVVLSPETGDE  62
            +KG+IL+GG GTRL P+T   +KQL+PI  KP ++Y L  L+ AGIR+I ++ +PE
Sbjct  1    MKGIILAGGSGTRLHPITRGVSKQLLPIYDKPMIYYPLSVLMLAGIREILIITTPEDKGY  60

Query  63   VRAALGDGSRWGVRLTYIVQEAPLGIAHAVKTAQGFLSDSPFVLYLGDNLLSG-GIKHLV  121
             +  LGDGS +G++L Y  Q +P G+A A    + FL+  P  L LGDN+  G G    +
Sbjct  61   FQRLLGDGSEFGIQLEYAEQPSPDGLAQAFIIGETFLNGEPSCLVLGDNIFFGQGFSPKL  120

Query  122  EEYRQTRPEAIVLLTPVEDPRAFGVVVLDGAGKVVRLLEKPKDPPSNLALVGVYLFSPAI  181
                  A V    V DP  FGVV  D   + + L EKPK P SN A+ G+Y +   +
Sbjct  121  RHVAARTEGATVFGYQVMDPERFGVVEFDDNFRAISLEEKPKQPKSNWAVTGLYFYDSKV  180

Query  182  HSIINRLKPSGRGEYEITEAIQGLVDEGKRVVAHQVRGW-WKDTGKPEDLLDANRLALSS  240
                ++KPS RGE EIT   Q  ++ G   V    RG+ W DTG  + L++A+    +
Sbjct  181  VEYAKQVKPSERGELEITSINQMYLEAGNLTVELLGRGFAWLDTGTHDSLIEASTFVQTV  240

Query  241  LTRR  244
              R+
Sbjct  241  EKRQ  244
```

Figure 18. Blast amino acid alignment of *M. ruber* Mrub_2034 and *E.coli* b3789. Mrub_2034 is the query sequence and b3789 is the subject sequence. This analysis was performed using NCBI BLAST bioinformatics tool at http://blast.ncbi.nlm.nih.gov.

Out of the total 244 amino acids 94 of them are conserved (39%). The bit score of 160 is somewhat high suggesting the two sequences are related. This alignment has an E-value of 2e-51 which is very small suggesting that the sequence is not conserved due to chance. Since only 39% of the sequence was conserved more evidence had to be collected to confirm or deny that Mrub_2034 and b3789 are orthologous.

The next bioinformatics tool listed was the CDD. Recall, if the *E. coli* gene and the *M. ruber* gene of interest both belong to the same COG group it suggests that they most likely are orthologous to one another. As can be seen in the table both genes belong to the COG family COG1209. Both *M. ruber* and *E. coli* have very small E-values in relation to the family suggesting that they aren't related due to chance. This was another piece of evidence suggesting the two genes are orthologous to one another.

The next column on the table represents the predicted cellular location of the proteins coded by the *M. ruber* and *E. coli* gene. This prediction was made by using bioinformatics tools such as TMHMM, SignalP, LipoP, PSORT-B, and Phobius. Figure 19 is a visual representation of the TMHMM results for both Mrub_2034 and b3789.

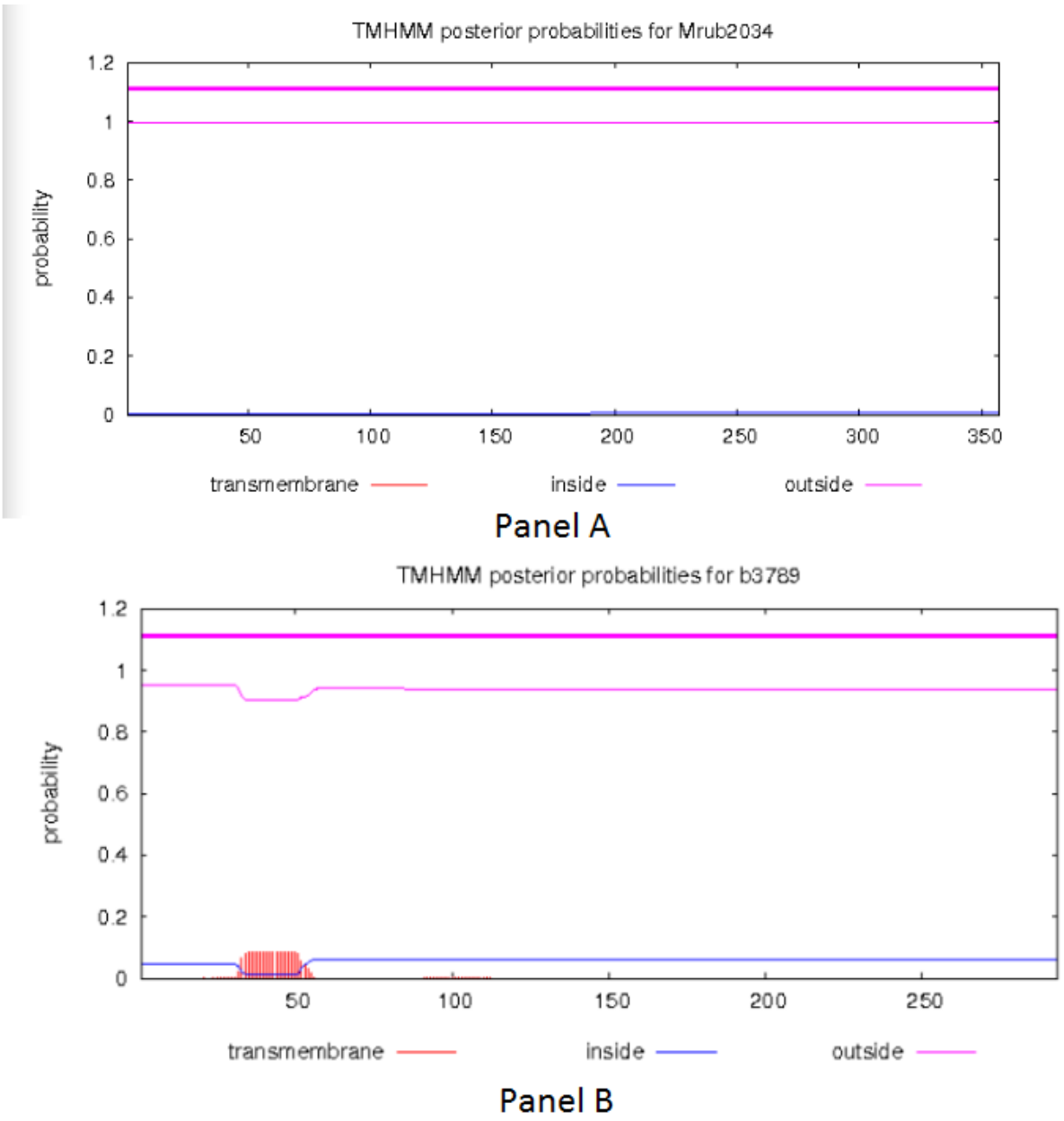


Figure 19. TMHMM transmembrane helices graph comparison of *M. ruber* Mrub_2034 and *E. coli* b3789 suggesting there are no transmembrane helices for either protein and that the protein is most likely located outside of the cell (cytoplasm). Panel A is the TMHMM transmembrane helices graph for *M. ruber* Mrub_2034 and panel B is the TMHMM transmembrane helices graph for *E. coli* b3789. TMHMM Server v. 2.0 found at http://www.cbs.dtu.dk/services/TMHMM was used to create these graphs.

In Figure 19 it can be seen that there are no red peaks in panel A (panel A representing Mrub_2034 and panel B representing b3789). In panel B there are some visible red peaks that typically would suggest that part of the protein has a transmembrane helix, but the probability is

so low it is not likely. This suggests that there are no transmembrane helices for either protein these genes code for. What these graph do tell us though is that there is a high probability that the proteins that these genes code for are located outside of the membrane, which would be the cytoplasm. This was one of the pieces of information suggesting that both proteins are located in the cytoplasm and that the genes that code for them are orthologous. The next tool used was SignalP. Recall SignalP is used to determine whether or not the predicted transmembrane helices are actually that or if they are signal peptides. A visual representation of this data can be seen in Figure 20.

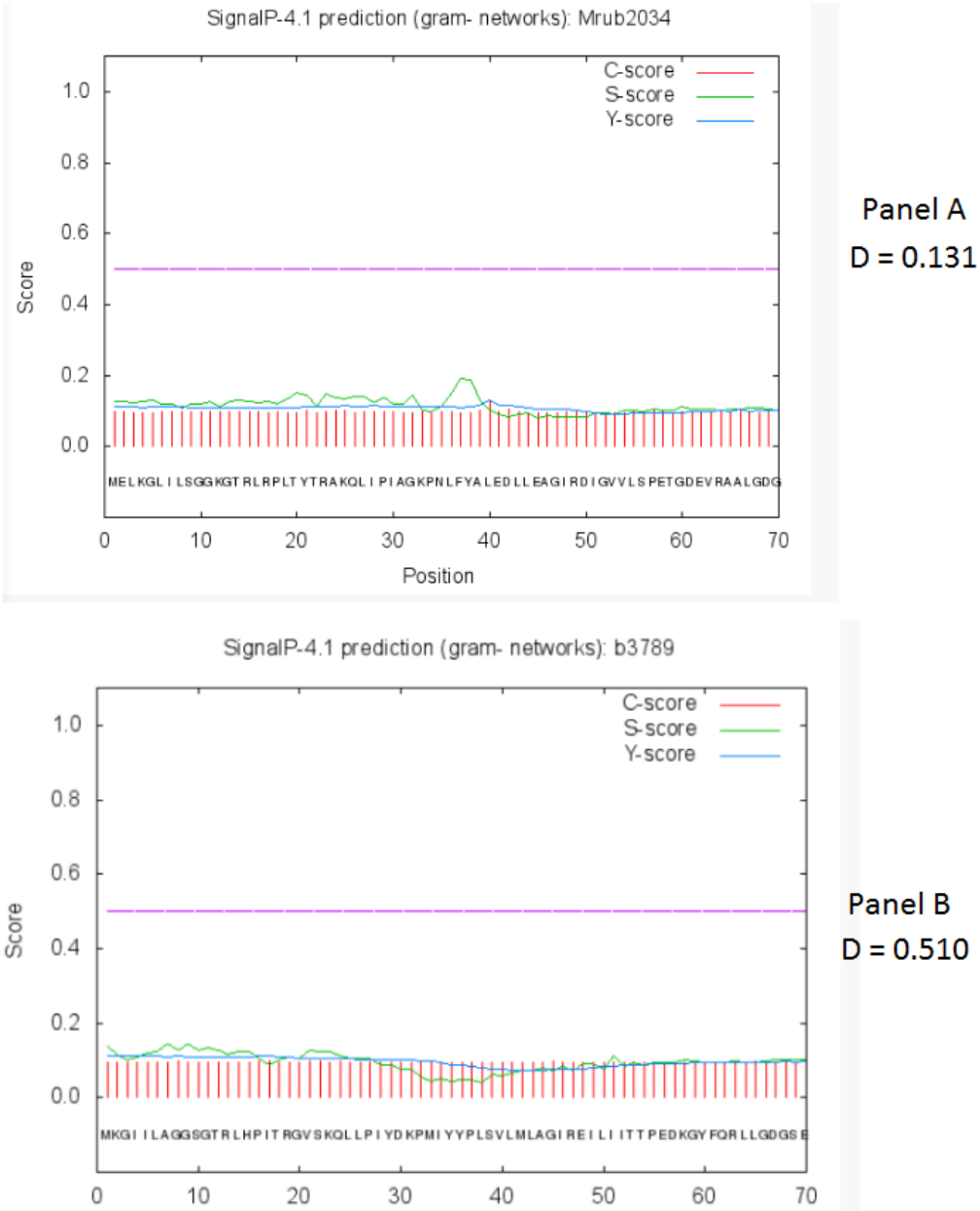


Figure 20. SignalP graphical representation of *M. ruber* Mrub_2034 and *E. coli* b3789 suggesting there are no transmembrane helices or signal peptides. Panel A represents *M. ruber* Mrub_2034 and panel B represents *E. coli* b3789. D values are located under the panel names on the graph. SignalP server v. 4.1 http://www.cbs.dtu.dk/services/SignalP was used to create these plots.

In Figure 20, there are four notable pieces of information that can be used to determine signal peptides. Those include the C-score, which distinguishes signal peptide cleavage sites from everything else, the S-score, which distinguishes the signal peptide position, the Y-score, which is a combined score of the C and S score, and the D value which it the probability that there is a signal peptide. On both panels A and B (panel A representing Mrub_2034 and panel B representing b3789) the C, S, and Y scores are very low. This makes sense because, as was suggested in Figure 19, there are no transmembrane helices that might be a signal peptide. The D value for both panel A and B is very low also confirming that there are no signal peptides. The fact that both Mrub_2034 and b3789 do not have transmembrane helices or signal peptides suggest that they may be orthologous.

After the bioinformatics tools TMHMM and SignalP were used LipoP, PSORT-B, and Phobius were also used. LipoP predicted that the proteins coded by Mrub_2034 and b3789 are found in the cytoplasm. PSORT-B confirmed this by attributing its highest score to the cytoplasm. Since Phobius is a visual representation of the results of TMHMM and SignalP and there were no results for either of these tools, the phobius was not useful. These tools suggested that both the protein coded for by Mrub_2034 and b3789 are found in the cytoplasm of the cell. This was another piece of evidence suggesting these genes were orthologous.

The next column displays what TIGRFAM protein family (domain) they are most closely related to. Notice that both *M. ruber* and *E. coli* are related to the same family with very low E-values. This family suggests that the identity of both proteins that Mrub_2034 and b3789 code for is most likely glucose-1-phosphate thymidylyltransferase. This again suggests they are orthologous to one another. The next column is the Pfam protein family. On the table it can be seen that both Mrub_2034 and b3789 are related to the same protein family with low E-values. A visual representation of the comparison the top Pfam protein family to the gene sequences can be seen in Figure 21.

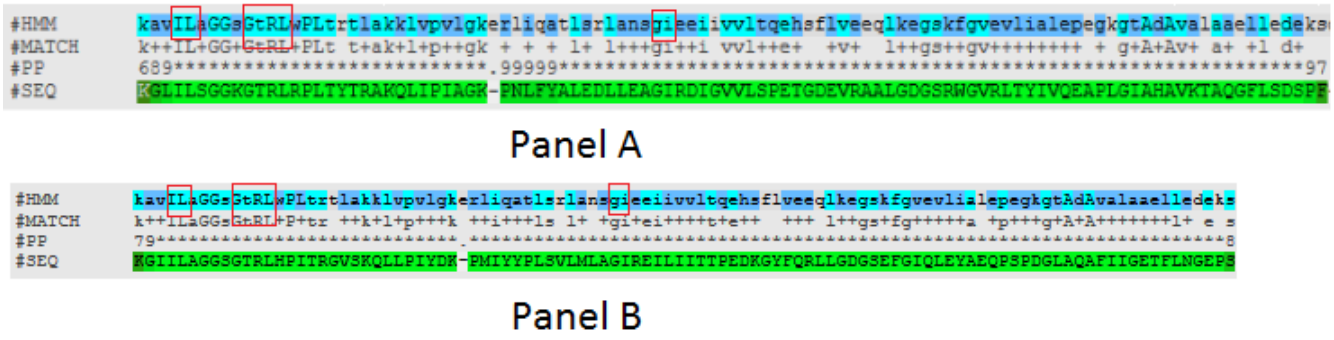


Figure 21. Pfam protein family PF00483 comparison to *M. ruber* Mrub_2034 and *E. coli* b3789 displaying conserved amino acids. Panel A represents *M. ruber* Mrub_2034 and panel B represents *E. coli* b3789. The red boxes display the similar conserved amino acids to PF00483. This pairwise alignment was created using the Pfam website http://pfam.sanger.ac.uk/search.

Figure 21 shows the comparison of both Mrub_2034 and b3789 to the protein family amino acid sequence. As can be noted by the red boxes and the light blue color in the Figure there are a number of similarities in conserved amino acids. This further concludes that they are highly

related and also suggests that the protein these genes code for is glucose-1-phosphate thymidylyltransferase. The next column is the PDB. Mrub_2034 and b3789 were paired with the same PDB database hit. That hit was for glucose-1-phosphate thymidylyltransferase. The E.C. number also confirmed this identity prediction.

The second to last column on the table is what pathway the proteins the genes code for are involved in. Both are involved in streptomycin biosynthesis. A graphic of this process can be seen in Figure 22.

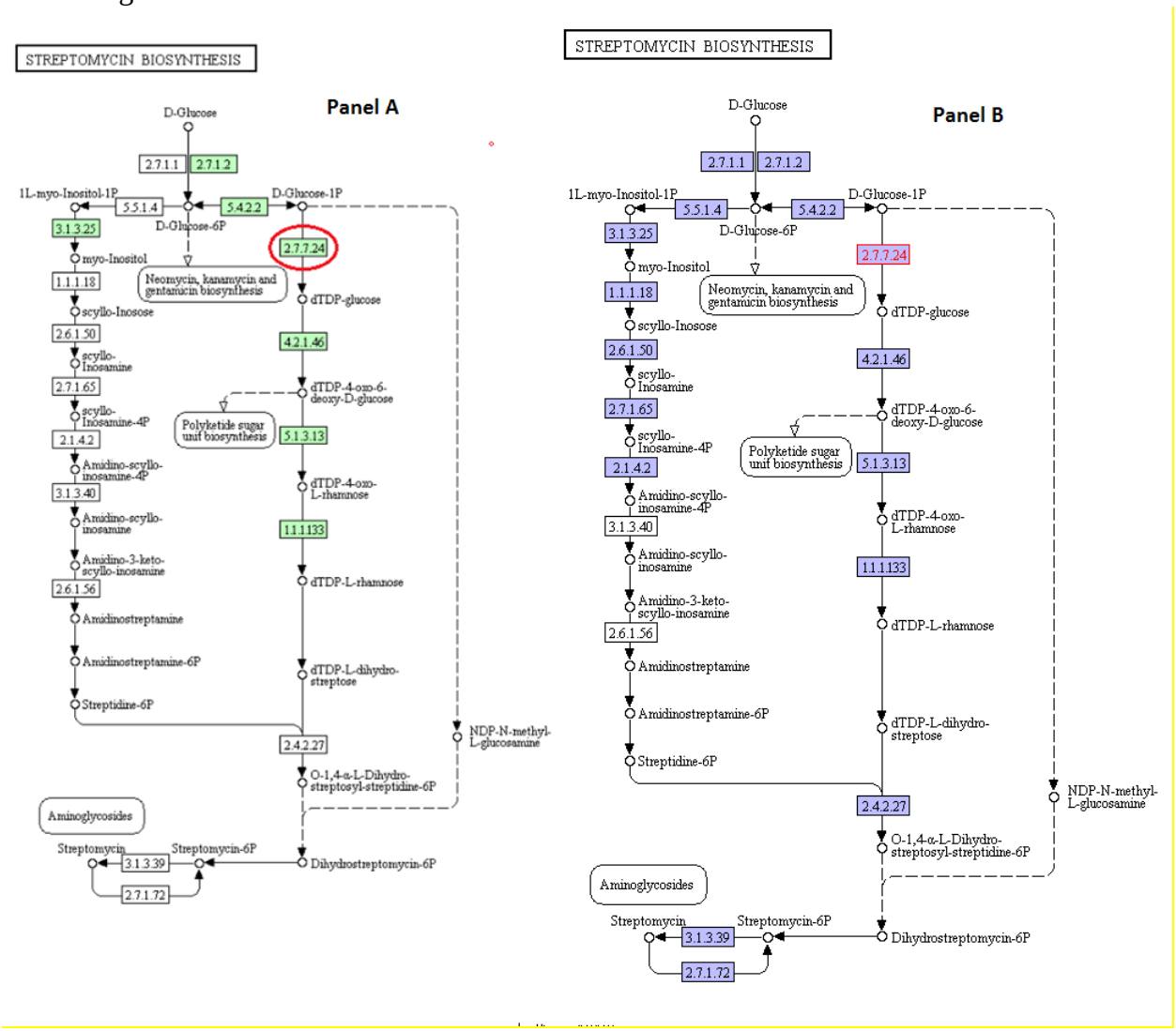


Figure 22. Enzyme E.C. number 2.7.7.24 presence in the streptomycin biosynthesis pathway. The protein coded by Mrub_2034 and b3789 is represented by E.C. number 2.7.7.24. The Kyto Encyclodpedia of Genes and Genomes (KEGG) data base was used to create this map at http://www.genome.jp /kegg/pathway.html

Figure 22 displays the entirety of the streptomycin biosynthesis pathway. The enzymes (E.C numbers) colored green can be found in *E. coli* and *M. ruber*. The red circle highlights E.C. number 2.7.7.24 which represents Mrub_2034 and b3789. As can be seen in the Figure this enzyme is suggested to convert converts D-Glucose-1P to dTDP-glucose (also known as dTDP-α-D-glucose). The KEGG pathway map was also used to determine if there are any known paralogs for the enzyme in both *M. ruber* and *E. coli*. Surprising enough Mrub_2034 is paralogs with Mrub_0628 and b3789 is paralogs with b2039. This is particularly surprising because Mrub_0628 might be orthologous with b2039.

Some additional information that was collected that is not included in table 3 is a comparison of the Mrub_2034 and b3789 gene neighborhood maps. In Figure 23 this comparison can be seen.

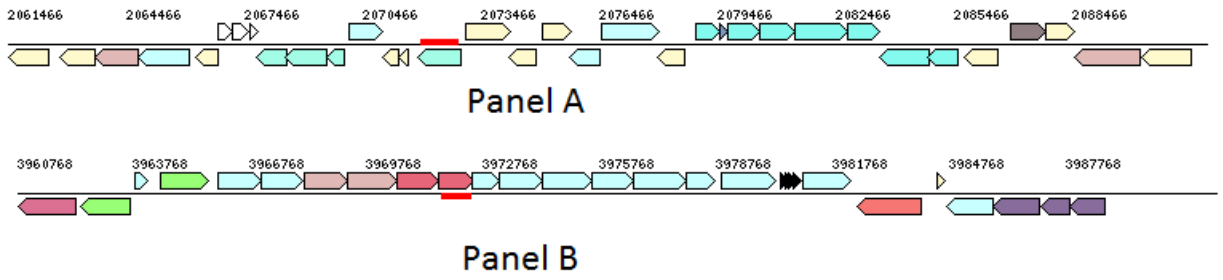


Figure 23. Comparison of *M. ruber* Mrub_2034 and *E. coli* b3789 gene neighborhood maps. Panel A represents *M. ruber* Mrub_2034 and panel B represents *E. coli* b3789. Both are underlined by a red line. Images were taken from http://img.jgi.doe.gov/.

Each gene in a gene neighborhood map is represented by an arrow. The function of that gene is represented by the color of the arrow. If there are numerous genes that are the same color pointing in the same direction it means that they are part of an operon. Mrub_2034 (underlined by a red line) doesn't appear to be a part of an operon, which is confirmed in Figure 24 (a comparison of flanking genes in related species). On the other hand, the gene in *E. coli* does appear to be in an operon, which is confirmed on the Ecocyc page for this gene (Keseler, 2013). While being a component of an operon that shares multiple genes is strong evidence of a shared evolutionary history, the lack of an operon for both genes is not refuting evidence for our hypothesis of their orthologous relationship. *E. coli* and *M. ruber* are in different phyla and not closely related. Chromosomal rearrangements are common even between closely related species, as can be seen in the chromosome maps in Figure 24.

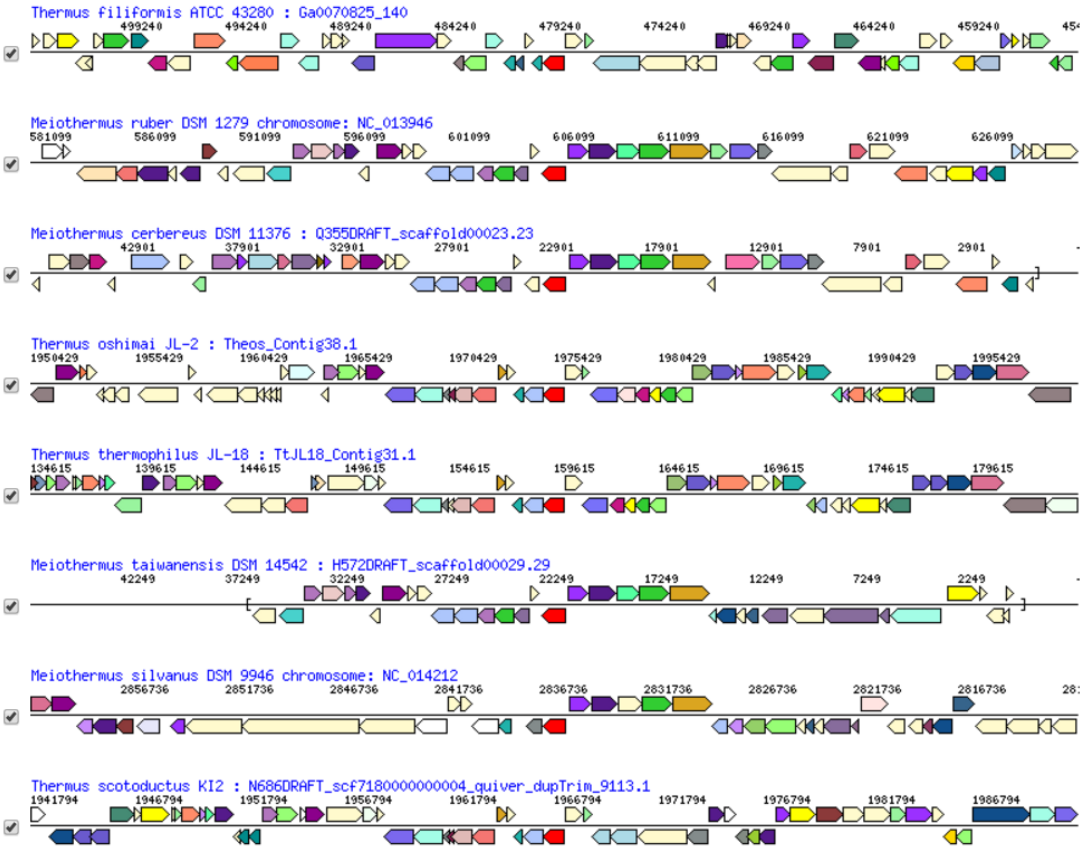


Figure 24. A comparison of flanking genes in related species including *Meiothermus ruber*. Mrub_2034 and related genes are represented by the red arrow. Numerous similar species are compared.

Another piece of additional information that was collected that is not included in table 3 is a comparison of the Mrub_2034 and b3789 phylogenetic trees. In Figure 25 this comparison can be seen.

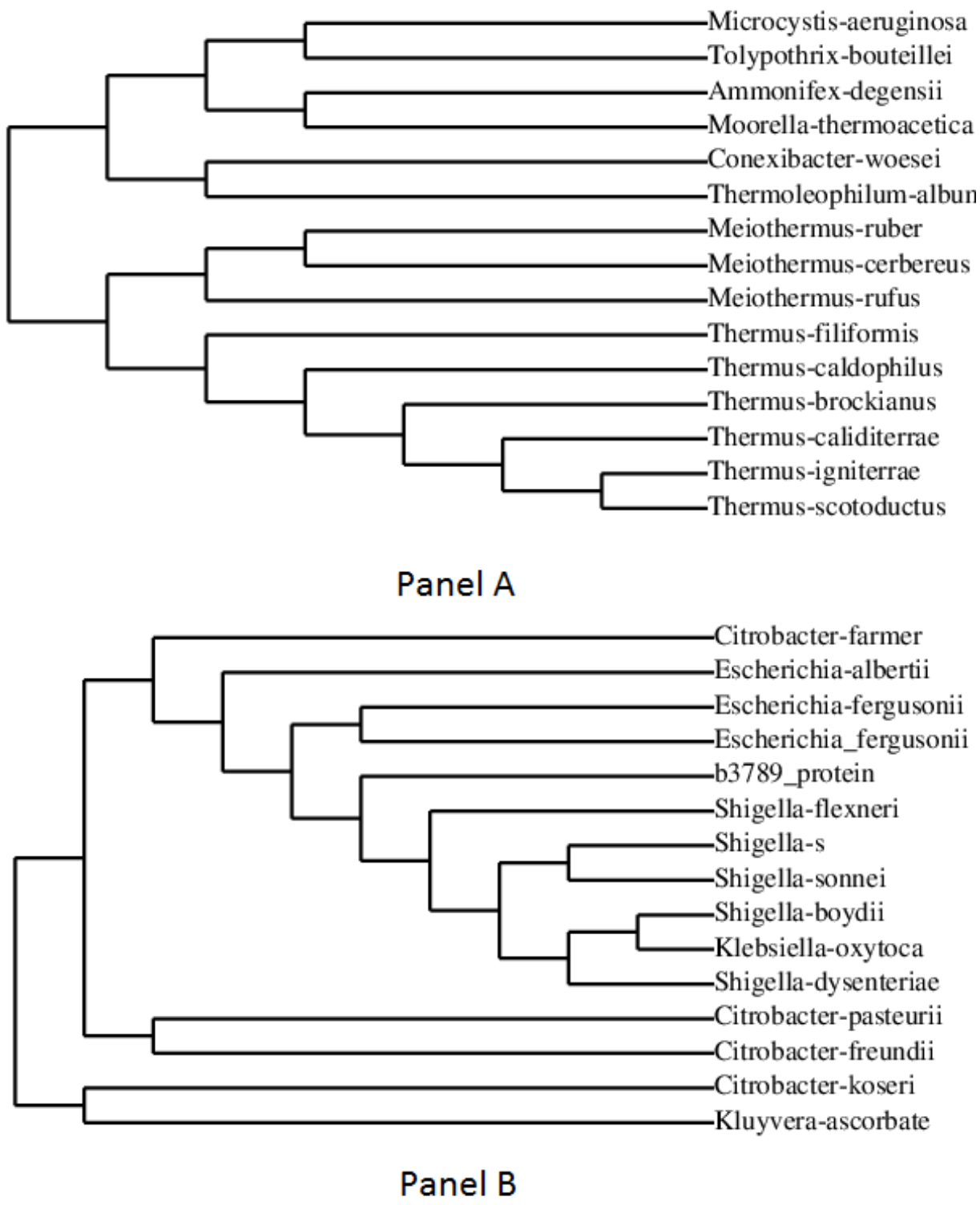


Figure 25. Comparison of *M. ruber* Mrub_2034 and *E. coli* b3789 phylogenic trees suggest horizontal gene transfer did not occur. Panel A represents M. ruber Mrub_2034 and panel B represents E. coli b3789. Images were created using http://www.phylogeny.fr

The images in Figure 25 show the phylogenetic trees of *M. ruber* and *E. coli*. Panel A represents *M. ruber* and panel B represents *E. coli.* All the organisms in panel A are part of the Deinococcus-Thermus phylum. All the organisms in panel B are part of the proteobacteria phylum. Since the organisms close to *M ruber* and *E. coli* in each panel belong to the same phylum there is no evidence that horizontal gene transfer occurred.

Considering that only one of the bioinformatics tools (gene neighborhood) suggested that Mrub_2034 and b3789 were not orthologous and the rest suggest that it is, it can be strongly predicted that they are in fact orthologous. The identity of these genes has been suggested to be glucose-1-phosphate thymidylyltransferase

Conclusion

Based on the results of these many bioinformatics tools it can be concluded that *M. ruber* Mrub_2052 and *E. coli* b0688 are orthologous to one another. Every bioinformatics tool utilized in the comparison unanimously suggested that both of these genes coded for the protein phosphoglucomutase and are orthologous to one another. One strong piece of evidence that confirmed this hypothesis was the protein BLAST of Mrub_2052's amino acid against b0688's amino acid sequence. This confirmed that 63% of the amino acids were identical at the same positions. Proteins with similar amino acid sequences tend to have similar function as well as suggest that they have evolved from a common ancestral gene. These are the qualities of orthologous genes (Gabaldón, 2013). These results can be seen in Figure 4. Another piece of strong evidence suggesting that *M. ruber* Mrub_2052 and *E. coli* b0688 are orthologous to one another was the results of TIGRFAM and Pfam. Both of these bioinformatics tools matched these genes to the same protein families (TIGRFAM: TIGR01132; Pfam: PF02878 & PF02880) with very small E-values (if an E- value is smaller than 10^{-100}, it is sometimes given as 0.0). This suggests that they were not paired with this protein family simply by chance. Both these tools also suggested that the identity of Mrub_2052 and b0688 was most likely phosphoglucomutase. These similarities suggest that the genes are orthologous. These results can be seen in Figure 7. The gene neighborhood map also helped confirm that the two genes are orthologous. On the map it can be seen that both Mrub_2052 and b0688 are the same color. This suggests that they have similar functions. This is yet another piece of evidence suggesting they both code for phosphoglucomutase and are orthologous.

The results for Mrub_0628 and b2039 as well as Mrub_2034 and b3789 were slightly different. During the analysis process it was determined that Mrub_0628 and Mrub_2034 were paralogs (using the KEGG pathway map). The same could be said about b2039 and b3789. Many of the bioinformatics tools suggested that the gene pairs were orthologous and that the identity of the protein they coded for was glucose-1-phosphate thymidylyltransferase. One example of this is the Pfam results. All four matched with PF00483. This makes sense because that protein family is for nucleotidyl transferases (NTP_transferases), enzymes which transfer nucleotides onto phosphosugars. Glucose-1-phosphate thymidylyltransferase is a nucleotidyl transferase, suggesting that this is the correct identity of Mrub_0628 and b2039 as well as Mrub_2034 and b3789, and that the pairs are orthologous. In addition to this they all belonged to the same COG family (COG1209) again suggesting the identity is glucose-1-phosphate thymidylyltransferase.

On the other hand, there is some refuting evidence suggesting that the Mrub_0628 and b2039 as well as Mrub_2034 and b3789 pairs are not orthologous. The first piece of evidence suggesting this is the protein BLAST. For Mrub_0628 and b2039 there were only 34% identical amino acids and for Mrub_2034 and b3789 there were only 39%. Despite the fact that the similarity is not as large as the similarity between Mrub_2052 and E. coli b0688 (63%) it does not mean that they don't code for orthologous proteins with similar function. Another piece of refuting evidence was the gene neighborhoods. For both the Mrub_0628 and b2039 as well as Mrub_2034 and b3789 pairs the *M. ruber* genes and *E. coli* genes were different colors. This suggests that they code for proteins with different function. When the functions the gene neighborhood was suggesting were observed, it could be seen that all 4 genes it could be seen there was no consistency and that the functions suggested were very random amongst all 4. This suggests that there possibly was an error in the gene neighborhood maps. All though there are these hiccups in the analysis, due to the large amount of information suggesting the Mrub_0628 and b2039 as well as Mrub_2034 and b3789 pairs are orthologous, these pieces of refuting evidence can be discounted. More research needs to be conducted to understand why there were these hiccups.

The phylogenetic trees for all three *M. ruber* genes analyzed suggests that horizontal gene transfer did not occur because all of the organisms in the tree were part of the deinococcus-thermus phylum. This suggests they came from a common ancestor rather than a gene not part of the tree.

If one of these genes were to be chosen to be studied through site-directed mutagenesis, Mrub_2052 would be the best candidate due to how strong the results were. Based on the WebLogo which shows the conserved amino acids across a wide variety of organisms linked to Mrub_2052, mutagenesis to H60 would most likely be a loss of function mutation. A visual description of this can be seen in Figure 26.

Figure 26. WebLogo of Mrub_2052 showing the amino acid H is highly conserved at position 60. These images were obtained from http://weblogo.berkeley.edu and http://pfam.sanger.ac.uk/search.

As can be seen on Figure 26, the amino acid histidine (H) is highly conserved at position 60. This means that across a variety of different species this amino acid is conserved. Histidine is typically very important to protein function because it is ideal residue for protein functional centres (Betts, 2003). It H60 were disrupted there is a large possibility that the phosphoglucomutase would not function properly. A Figure of what primer could be used to perform this site directed mutagenesis can be seen in Figure 27.

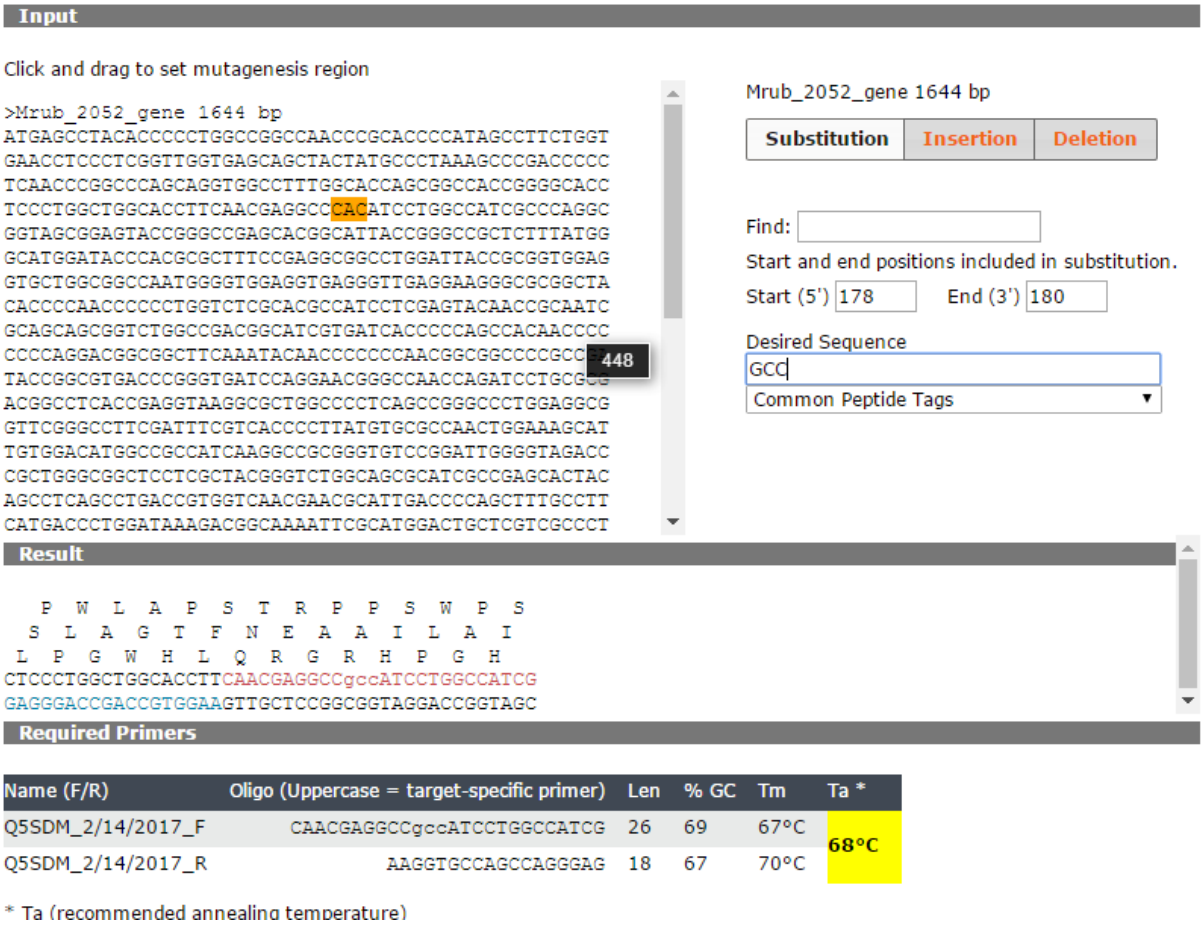


Figure 27. Primers designed to perform site directed mutagenesis on histidine 60 in Mrub_2052. The primers designed would replace histidine at position 60 with an alanine. Images obtained from http://nebasechanger.neb.com/.

The histidine at position 60 in the amino acid sequence is represented by the codon CAC in the nucleotide sequence ranging from 178 to 180. As mentioned before histidine is typically important to the function of a protein. The primers designed in Figure 27 would perform a site directed mutagenesis replacing the codon CAC with GCC, which is a codon for alanine. Alanine is not essential for protein function, thus by performing this replacement there will most likely be a loss of function in Mrub_2052. This in turn would hinder the streptomycin biosynthesis pathway because phosphoglucomutase would no longer be able to convert D-Glucose- 6P (also known as D-glucopyranose 6-phosphate) to D-Glucose-1P (also known as α-D-glucopyranose 1-phosphate).

Literature Cited

A DOE Office of Science User Facility of Lawrence Berkeley National Laboratory. DOE Joint Genome Institute. 2017 Jan 19. http://jgi.doe.gov/

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., Russell, R.B. Amino acid properties and consequences of substitutions. Bioinformatics for Geneticists. 2003. Availible at: http://www.russelllab.org/aas/His.html

Blattner, F. R.. The Complete Genome Sequence of Escherichia Coli K-12. Science 277.5331 (1997): 1453-462.

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

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

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.

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

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

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

Keseler, I.M., Mackie, A., Peralta-Gil, M., Santos-Zavaleta, A., Gama-Castro, S., Bonavides-Martinez, C., Fulcher, C., Huerta, A.M., Kothari, A., Krummenacker, M., Latendresse, M., Muniz-Rascado, L., Ong, Q., Paley, S., Schroder, I., Shearer, A., Subhraveti, P., Travers, M., Weerasinghe, D., Weiss, V., Collado-Vides, J., Gunsalus, R.P., Paulsen, I., and Karp, P.D. 2013. EcoCyc: fusing model organism databases with systems biology Nucleic Acids Research 41:D605-612.

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

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

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

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

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

Podell S, Gaasterland T. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. Genome Biology . 2007;8(2).

Schatz A, Bugle E, Waksman SA. Streptomycin, a Substance Exhibiting Antibiotic Activity Against Gram-Positive and Gram-Negative Bacteria.* . Experimental Biology and Medicine. 1944;55(1):66–69.

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

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

Walker, Margaret S., Walker, James B. 1971. Streptomycin Biosynthesis. 246(22): 7034-7086.

Yu, N.Y., Wagner, J.R, Laird, M.R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S.C., Ester, M., Foster, L.J., Brinkman, F.S.L. (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes, Bioinformatics 26(13):1608-1615