Mrub_0258 gene is an ortholog of the b4226 gene \((ppa)\) found in \textit{Escherichia coli}; Mrub_1198 gene is an ortholog of the b2501 gene \((ppk)\) found in \textit{Escherichia coli};

Brandon M. Wills  
\textit{Augustana College, Rock Island Illinois}

Dr. Lori Scott  
\textit{Augustana College, Rock Island Illinois}

Follow this and additional works at: http://digitalcommons.augustana.edu/biolmruber

Part of the\ Bioinformatics Commons, Biology Commons, Genomics Commons, Molecular Biology Commons, and the Molecular Genetics Commons

Recommended Citation
Wills, Brandon M. and Scott, Dr. Lori. "Mrub_0258 gene is an ortholog of the b4226 gene \((ppa)\) found in \textit{Escherichia coli}; Mrub_1198 gene is an ortholog of the b2501 gene \((ppk)\) found in \textit{Escherichia coli};" (2016). \textit{Meiothermus ruber Genome Analysis Project}.  
http://digitalcommons.augustana.edu/biolmruber/6

This Student Paper is brought to you for free and open access by the Biology at Augustana Digital Commons. It has been accepted for inclusion in Meiothermus ruber Genome Analysis Project by an authorized administrator of Augustana Digital Commons. For more information, please contact digitalcommons@augustana.edu.
Mrub_0258 gene is an ortholog of the b0258 gene (ppa) found in Escherichia coli; Mrub_1198 gene is an ortholog of the b2501 gene (ppk) found in Escherichia coli

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

INTRODUCTION

Why Study Meiothermus Ruber?
Meiothermus ruber (M. ruber) is a unique, red-pigmented, thermophilic bacterium that preferentially grows in high-temperature environments ranging from 35-70°C (Tindall et al., 2010). In fact, this bacterium was first isolated in the Russian city of Kamchatka, where the city’s hot springs provide optimal temperatures in which these bacteria thrive (Loginova et al., 1975). Unlike E. coli and Salmonella, both of which have over 30,000 publications available on PubMed, M. ruber has only had 28 such publications (Scott 2013). Due to the lack of studies performed on this organism, there is quite a bit of information missing in regards to the genes found within this organism’s genome and the function that these genes play. With so much information already available on E. coli and Salmonella, one might ask why even bother studying these less well-known organisms in the first place. To answer that question, we can look to the Joint Genome Institute, which has started a project called the Genomic Encyclopedia of Bacteria and Achaea (GEBA) that focuses specifically on the study of lesser-known bacteria. The study of organisms like M. ruber may reveal genes or variants of certain processes that are lacking in the other bacteria that are well studied (Wu et al., 2009). Therefore, it is important that we do not overlook bacterial species just because they are not particularly well-studied. Using E. coli as a control, we attempt to fill in some of the gaps that currently exist in the understudied M. ruber bacterial genome.

Inorganic Pyrophosphatase
The b4226 locus tag corresponds to a gene found in the E. coli bacteria called ppa that codes for inorganic pyrophosphatase. The gene is located at coordinates 4449122..4449652 in the E. coli genome and does not make up part of an operon. Inorganic pyrophosphatase is a small protein residing in the cytoplasm that consists of 175 amino acids and functions as a single polypeptide (Lahti et al., 1998). As shown in Figure 1, this enzyme is indirectly involved with oxidative phosphorylation and catalyzes an important reaction involving energy metabolism. During catalysis, a polyphosphate ion is cleaved into two separate phosphate ions (Yang et al, 2009). The resulting phosphates can be used by ATP synthase to produce molecules of ATP. Because of the versatile nature of phosphate ions, the products of this reaction can also be utilized during the synthesis of many important biological macromolecules (Yang et al., 2009).
Phosphate ions also play a critical role in glycolysis, as many of the enzymes involved with the production of pyruvate depend on the presence of phosphate and diphosphate ions (Terkeltaub 2001).

Other studies on this unique enzyme have also revealed that it plays an important role in maintaining proper cellular growth for E. coli bacteria (Chen et al., 1990). In addition to creating phosphate ions, the reaction catalyzed by inorganic pyrophosphatase is exergonic, meaning that energy is released during the reaction (Sun et al., 2014). This is due to the fact that the breaking of the phosphate-phosphate bond results in two molecules that are more stabilized by resonance, experience less electrostatic interaction, and can be solvated more easily. As we learned in biochemistry, the energy that is released during exergonic reactions may be used to power other reactions that would normally not proceed on their own. Therefore, not only is this enzyme critical to the production of many essential cellular molecules, but it can also be used as a driving force to propel energetically unfavorable reactions forward.

A large amount of research has been focused on cloning this gene and studying its active site (Yang et al, 2009; Lahti et al., 1988). Crystal structures of this enzyme have be obtained for a number of different organisms and there is a great deal of information available on the E. coli form of this protein. Also, the structure of this crystallized enzyme has also been determined for Thermus thermophilus, which is a bacteria closely related to Meiothermus ruber that is used for this study (Teplyakov et al., 1994). The fact that so much research has been conducted on the active site of this enzyme is indicative that this gene does code for inorganic pyrophosphatase. According to Yang et al. (2009), the solved structure for the E. coli version of inorganic pyrophosphatase shows that this short protein contains 4 metal ions in the center surrounded by a number of positively charged amino acids. The metal ions at the center of the protein form the active site of the molecule and play a critical role in maintaining proper active site orientation. Because the substrate binding to this site is a negatively charged pyrophosphate ion, the positive charges help to both promote substrate binding and stabilize the transition state as the high energy phosphate bond is cleaved between two phosphate molecules (Yang et al., 2009). Not only are the metal ions in the active site important, but the amino acids near the active site also playing a role in the protein’s catalytic ability. Studies have shown that of the amino acids that surround the active site, the aspartic acid residue at the 67th position is of most importance. In the protonated state, this amino acid is involved with the reverse reaction of converting individual phosphate ions back into diphosphate.

Figure 1. Reaction catalyzed by the inorganic pyrophosphatase enzyme. Image taken from http://metacyc.org/META/NEW-IMAGE?type=REACTION&object=INORGPYROPHOSPHAT-RXN
ions (Yang et al., 2009). Therefore we can see that the structure of this enzyme allows the enzyme to efficiently carry out this catalysis.

**Polyphosphate Kinase**

The locus tag *ppk* corresponds to a different gene found in the *E. coli* bacteria that codes for the enzyme polyphosphate kinase. As shown in figure 2 below, this enzyme is responsible for the lengthening (forward reaction) or shortening (reverse reaction) of polyphosphate chains within the cell. The lengthening of polyphosphate chains requires the use of energy and with each reaction a phosphate is added onto an already existing polyphosphate chain. The opposite can be said for the reverse reaction, in which energy is released as the bond between two phosphate molecules is cleaved.

The diphosphates created from this enzyme can be catalyzed by inorganic pyrophosphatase, which can then be utilized as described in the above paragraphs. The gene is located at coordinates 2623044..2625110 in the *E. coli* genome and, just like inorganic pyrophosphatase, this gene also does not make up part of an operon. This protein is not found in the cytoplasm, but is rather a peripheral membrane protein, meaning that it can attach and detach itself from the outer membrane depending on where it is needed for functioning (Akiyama et al., 1992).

Similar to inorganic pyrophosphatase, the *E. coli* version of polyphosphate kinase has been cloned and purified and its active site has also been studied extensively (Alkiyama et al., 1992; Zhu et al., 2005). Unlike inorganic pyrophosphatase, polyphosphate kinase is a very large protein consisting of 687 amino acid residues (Alkiyama et al., 1992). According to Zhu et al. (2005), the protein is made up of 4 distinct domains an N-terminus domain, a head (middle) domain, and two C-terminus domains called C1 and C2. Both of the C domains are the most important for providing the enzyme with its catalytic function. All four of the domains come together in the middle of the protein to form a tunnel where both the ATP molecules and the polyphosphate needed for catalysis can bind (Zhu et al., 2005). From here, catalysis begins and the polyphosphate can either be shortened or lengthened depending on what is needed.

Polyphosphate kinase is a very versatile enzyme that plays a number of important roles within the cell. Studies on the biological functions of this enzyme have found that polyphosphate
kinase makes virulence factors and aids in cellular movement (Zhu et al., 2005). Furthermore, other studies have shown that this enzyme is important for the production of biofilms in *Porphyromonas gingivalis*, a bacteria that is responsible for causing gum inflammation in humans (Chen et al., 2002). Furthermore, other research on this protein indicates that polyphosphate kinase is also part of RNA degradosome, and therefore also plays a part in degrading molecules of RNA (Blum et al., 1997). Therefore, this enzyme can function both on its own and as a complex of proteins.

In this project, we utilize a variety of bioinformatics tools to determine if the *ppa* and the *ppk* genes are orthologous to the genes Mrub_0258 and Mrub_1198, respectively. By using the programs on the GENI-ACT website, we can determine differences and similarities that exist between the gene pairs selected for this study. Doing so requires knowledge of E-values, a value generated by the bioinformatics tool that helps to determine the significance of the results. High E-values indicate that the sequences entered into the bioinformatics program likely line up due to chance alone. On the other hand, low E-values indicate that the sequence alignment was significant, and that they almost certainly did not line up due to chance. Before starting the project we initially performed BLAST searches of *ppa* versus the Mrub_0258 gene and the *ppk* gene versus the Mrub_1198 gene. Based on the low E-value outputs given by the BLAST program for both of these gene pairs, we hypothesize that the gene pairs selected for this study are orthologous to one another.

**METHODS**

To gather data on the *E. coli* and *M. ruber* genes, the GENI-ACT gene annotation website instructions were followed (http://www.geni-act.org/education/main/) with a few deviations. We started off by performing a BLAST of the *E. coli* genes versus their respective *M. ruber* genes to determine similarities between the two sequences. Once we had established that the sequence pairs were similar to one another, we continued on by filling out the various modules on the GENI-ACT site by using their associated bioinformatics programs. One deviation we had from the GENI-ACT instructions is that instead of using the recommended top 10 BLAST hits for the T-coffee analysis, we instead used the top twenty hits. Another deviation is that we used the map colored by KEGG for the Gene Context of the Horizontal Gene Transfer module. Also for the Gene Context module, I did not find the Chromosome Viewer for multiple bacterial species because the Chromosome viewer for the assigned genes was enough to prove that they were not in an operon. Phobius modules were omitted for all genes because the TMHMM data showed no transmembrane helices. We also did not fill out the Chromosome Viewer GC Heat Map. All other modules were completed following the instructions provided on the website.
RESULTS/DISCUSSION

Table 1. *E. coli ppa* and Mrub_0258 are orthologs

<table>
<thead>
<tr>
<th>Bioinformatics tool used</th>
<th><em>E. coli ppa</em> gene</th>
<th>Mrub_0258 gene</th>
</tr>
</thead>
</table>
| BLAST *E. coli* against *M. ruber* | Score: 148 bits  
E-value: 2e-49 | |
| CDD Data (COG category) | COG Number: COG0221  
Inorganic pyrophosphatase | E-value: 3.64e-85  
E-value: 1.27e-68 |
| Cellular Localization | | Cytoplasm of the cell |
| TIGRfam – protein family | | N/A |
| Pfam – protein family | Pf00719 (inorganic pyrophosphatase) |  
E-value: 1.1e-55  
E-value: 4.6e-52 |
| Protein Database | 4UM4 and 2PRD  
*Structure of inorganic pyrophosphatase from Escherichia coli in complex with sulfate & Crystal structure of inorganic pyrophosphatase from Thermus thermophilus.*  
E-value: 2.27548e-100  
E-value: 3.01152e-71 |
| Enzyme commission number | E.C. 3.6.1.1 – Inorganic pyrophosphatase |
| KEGG pathway map | Oxidative phosphorylation  
KEGG number: 00190 |

Table 1 summarizes the results from a variety of bioinformatics tools that were used to compare the *E. coli ppa* gene to the Mrub_0258 gene. The first row of data shows the results of the initial BLAST analysis as discussed in the introduction. Because these two proteins vary in
length, the bit score is not particularly relevant in this situation. Of more importance is the E-value output of the BLAST results which is 2e-49. Since this number is close to zero, we can be certain that these two sequences do not align due to random chance alone and that they share many of the same amino acids. This provides evidence that these two organisms are evolutionarily related and may have this gene in common. The CDD pulled the same COG number (COG0221) and name (inorganic pyrophosphatase) from the database, both with very small E-values indicating significance. Thus, the genes likely code for the same enzyme that plays a role in energy metabolism. All of the bioinformatics tools used to analyze cellular location (TMH, SignalP, LipoP, and PSORT-B) suggested that both proteins are located in the cytoplasm of the cell and that both also lack cleavage sites. Therefore, the cellular location of the enzyme coded by *ppa* and Mrub_0258 are the same, which further suggests that the two genes are orthologs. Interestingly this enzyme does not have a TIGRFam number, a result that was confirmed via the TIGRFam website as well as the BLAST website. Both sites did not list a TIGRFam number when the protein sequences were entered into the search bar. However, both genes were pulled from the Pfam database. Analysis of the genes with Pfam confirmed a hit of the same one-domain protein with a Pfam number of PF00719 (inorganic pyrophosphatase). The protein database pulled different numbers for each gene. The number pulled for *ppa* was the *E. coli* crystalized protein while the number pulled for Mrub_0258 was the *Thermus thermophilus* crystalized protein. Either way, both of the PDB analyses pulled up the same enzyme. Also enzyme commission number of E.C.3.6.1.1 was the same for both genes. Lastly, both genes were predicted to be involved in the same KEGG pathway for energy metabolism in oxidative phosphorylation.

Figure 3. The b4226 and Mrub_0258 genes have a significant number of similar amino acids. E-value for this BLAST analysis was 2e-49, which means that it is very unlikely that these sequences aligned due to random chance. NCBI Basic Local Alignment Search Tool (BLAST) was used to perform the analysis [http://www.ncbi.nlm.nih.gov/blast](http://www.ncbi.nlm.nih.gov/blast).

Figure 3 shows the results of a BLAST search of *E. coli ppa* vs. Mrub_0258 before I started filling in any of the GENI-ACT modules. As we can see from the data, 46% of the amino acids were exactly the same between both sequences, and 113 of the amino acids were similar in character. Of most importance is the E-value calculated for the BLAST, which is 2e-49. This E-value is very close to being zero and therefore there is an extremely small probability that these sequences aligned due to random chance. Therefore, we can see that *E. coli ppa* and Mrub_0258 appear to
share some primary structural similarities. This is our first piece of evidence that suggests that the two genes might be evolutionarily related to one another.

Figure 4. Mrub_0258 and E. coli b4226 do not contain TMH regions; a cytoplasmic location is predicted for both proteins. Panel A shows the TMHMM for b4226; Panel B shows the TMHMM data for Mrub_0258. TMHMM Server v 2.0 http://www.cbs.dtu.dk/services/TMHMM was used to create these two hydropathy charts.

The images in figure 4 above display TMH hydropathy plots for E. coli ppa and Mrub_0258. Red peaks that appear on the plots indicate the presence of transmembrane helices. As we can see, red peaks did not appear on either of the plots after the analysis. Therefore, the TMHMM hydropathy plots of both the ppa gene and Mrub_0258 are consistent with one another, each demonstrating that the protein coded by these genes are present in the cytoplasm as opposed to the membrane.
Figure 5. Mrub_0258 and *E. coli* b4226 do not contain cleavage sites; D values for both plots were below the cutoff value in both cases. Panel A shows the plot for b4226; Panel B shows the plot for Mrub_0258. Signal P server v. 4.1 [http://www.cbs.dtu.dk/services/SignalP](http://www.cbs.dtu.dk/services/SignalP) created these plots.

The above plots in Figure 5 are SignalP graphs generated for both *E. coli* *ppa* and Mrub_0258. This bioinformatics tool is used to predict protein cleavage sites by assigning each protein a D-value, which is calculated using the S-score and Y-score, and a cutoff value, which is indicated by the
purple line. For *E. coli* ppa (Panel A), the D-value (0.103) is lower than the cutoff value (0.450), which means that this protein does not contain any cleavage sites. The same can be said for Mrub_0258 (Panel B), as the D-value of 0.103 is again below the cutoff of 0.450. Therefore, these data are consistent for both genes, suggesting that neither have cleavage sites.

**Figure 6.** Mrub_0258 and *E. coli* b4226 are present in the same biochemical pathway. It is predicted that both of these genes are involved in energy metabolism for oxidative phosphorylation, both coding for the enzyme inorganic pyrophosphatase. Panel A shows the KEGG pathway after selecting for *Escherichia coli*. Panel B shows the KEGG pathway after selecting for *M. ruber*. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database at http://www.genome.jp/kegg/ was used to locate the KEGG pathway these genes are involved in.

Figure 6 shows the biochemical pathway that both of these genes are involved in. Enzymes that are colored in green are thought to be present in the organism. From the pathways, we see that *E. coli* ppa and Mrub_0258 are both involved with energy metabolism, both coding for the enzyme inorganic pyrophosphatase. This provides even more evidence to suggest that these organisms are evolutionarily related and that these two genes are orthologous.

**Panel A**

**Panel B**

**Figure 7.** Mrub_0258 and *E. coli* b4226 have the same highly conserved amino acids and code for the same protein, inorganic pyrophosphatase. As we can see from this pairwise alignment, the residues at the E22, K29, Y54, G55, T60, G64, D65, D67, P68 and D70 positions seem to be the most conserved of the amino acids in the full sequence. Panel A shows the pairwise alignment for *Escherichia coli*. Panel B shows the pairwise alignment for *M. ruber*. The pairwise alignment was created using the Pfam website http://pfam.sanger.ac.uk/search.

As we can see from the pairwise alignments in Figure 7, both *E. coli* ppa and Mrub_0258 contain the same highly conserved residues in the middle of the sequence. Rather than the two sequences being compared to one another as in the BLAST search, this pairwise alignment
compares the provided sequence to a consensus sequence created from hundreds of other proteins. The fact that both the *E. coli ppa* and Mrub_0258 protein sequences both pulled out the same exact consensus sequence gives us even more information to support that the two genes might be orthologous to one another.

Panel A

Panel B

Figure 8. *E. coli b4226* and Mrub_0258 genes are not part of an operon. Both genes are also predicted to be involved with energy metabolism, which is consistent with the rest of the data. Chromosome Viewer maps were colored by KEGG. Panel A: *E. coli b4226* Chromosome Viewer; Panel B: Mrub_0258 Chromosome Viewer. Genes are marked with a red line. Images were taken from http://img.jgi.doe.gov/

The above images in Figure 8 give us even further information that these two genes are orthologous to one another. As noted by the differences in color from the other genes near them, neither of these genes are part of an operon. The *ppa* gene is brown and the Mrub_0258 gene is pink, which is indicative of genes that are involved with energy metabolism. Therefore, the Chromosome Viewer map adds even more evidence that these two genes might be orthologs.

Table 2. *E. coli ppk* and Mrub_1198 are orthologs

<table>
<thead>
<tr>
<th>Bioinformatics tool used</th>
<th><em>E. coli ppk</em> gene</th>
<th>Mrub_1198 gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLAST <em>E. coli</em> against <em>M. ruber</em></td>
<td>Score: 262 bits E-value: 2e-81</td>
<td></td>
</tr>
<tr>
<td>CDD Data (COG category)</td>
<td>COG Number: COG0855 Polyphosphate kinase</td>
<td></td>
</tr>
<tr>
<td></td>
<td>E-value: 0.00</td>
<td>E-value: 3.97e-152</td>
</tr>
<tr>
<td>Cellular Localization</td>
<td>Cytoplasmic membrane</td>
<td></td>
</tr>
</tbody>
</table>
Table 2 summarizes the results from a variety of bioinformatics tools that were used to compare the *E. coli* *ppk* gene to Mrub_1198. The first row of data shows the results of the initial BLAST analysis as discussed in the introduction. Because these two proteins vary in length, the bit score is not particularly relevant in this situation. Of more importance is the E-value output of the BLAST results which is 2e-81. Since this number is close to zero, we can be certain that these two sequences do not align due to random chance alone and that they share many of the same amino acids. This provides evidence that these two organisms are evolutionarily related and may have this gene in common. The CDD pulled the same COG number (COG0855) and name (polyphosphate kinase) from the database, both with very small E-values indicating significance. Thus, the genes both code for the same enzyme involved with energy metabolism. Furthermore, the TMHMM and SignalP were consistent between the two genes, indicating a

<table>
<thead>
<tr>
<th>TIGRfam – protein family</th>
<th>TIGR03705 (polyphosphate kinase)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>E-value: 0.00</td>
</tr>
<tr>
<td>Pfam – protein family</td>
<td>PF13089 (N-terminus Domain), PF02503 (Middle Domain) PF13090 (C-terminus Domain)</td>
</tr>
<tr>
<td></td>
<td>Polyphosphate Kinase</td>
</tr>
<tr>
<td></td>
<td>E-value</td>
</tr>
<tr>
<td></td>
<td>5.8e-143</td>
</tr>
<tr>
<td></td>
<td>4.3e-88</td>
</tr>
<tr>
<td>Protein Database</td>
<td>1XDO</td>
</tr>
<tr>
<td></td>
<td>Crystal structure of Escherichia coli polyphosphate kinase</td>
</tr>
<tr>
<td></td>
<td>E-value: 0.00</td>
</tr>
<tr>
<td>Enzyme commission number</td>
<td>E.C 2.7.4.1 – polyphosphate kinase</td>
</tr>
<tr>
<td>KEGG pathway map</td>
<td>Oxidative phosphorylation</td>
</tr>
<tr>
<td></td>
<td>KEGG number: 00190</td>
</tr>
</tbody>
</table>
lack of transmembrane helices and signal peptides, respectively. According to the bioinformatics tool LipoP we might believe that the protein is located in the cytoplasm of the cell. However, the data from PSORT-B suggests otherwise, indicating that both of the proteins are located in the cytoplasmic membrane. Upon research of the enzyme, it was found that the *E. coli* version of polyphosphate kinase is a peripheral outer membrane protein, which means that the cytoplasmic prediction is only partially correct and PSORT-B is completely incorrect (Akiyama et al., 1992). This would also explain why no transmembrane helices were predicted, because although it polyphosphate kinase is a membrane-associated protein, it does not go through the membrane. Out of all the bioinformatics programs used for this study, the cellular location modules were the only ones that showed deviations between the two genes. Further evidence of these genes being orthologous can be found in the TIGRfam results. The TIGRfam number pulled from the database for both protein sequences was the same, TIGR03705, which also came up as polyphosphate kinase. Furthermore, Pfam confirmed that both proteins contain the same three domains, a catalytic C-terminus domain (PF13090), a middle domain (PF02503), and an N-terminus domain (PF13089). According to research on the structure of this enzyme, the C domain pulled from the Pfam database is actually two separate domains that are called C1 and C2 (Zhu et al., 2005). The protein database (PDB) also pulled the same number and name for both sequences and the enzyme commission number of E.C.2.7.4.1 was the same for both genes. Lastly, both genes were predicted to be involved in the same KEGG pathway involved with energy metabolism for oxidative phosphorylation. All of these similarities of these genes found using bioinformatics suggests that these genes could also be orthologous to one another.
Figure 9 shows the results of a BLAST search of *E. coli* *ppk* vs. Mrub_1198 before I started filling in any of the GENI-ACT modules. As we can see from the data, 30% of the amino acids were exactly the same between both sequences, and 322 of the amino acids were similar in character. Of most importance is the E-value calculated for the BLAST, which is $2e^{-81}$. This E-value is very close to being zero and therefore there is an extremely small probability that these sequences aligned due to random chance. Therefore, we can see that *E. coli* *ppk* and Mrub_1198 appear to share some primary structural similarities. This is our first piece of evidence that suggests that the two genes might be evolutionarily related to one another.
The images in figure 10 above display TMH hydropathy plots for E. coli ppk and Mrub_1198. Red peaks that appear on the plots indicate the presence of transmembrane helices. As we can see, red peaks did not appear on either of the plots after the analysis. Therefore, the THM hydropathy plots of both the ppk gene and Mrub_1198 are consistent with one another, each demonstrating that the protein coded by these genes is not part of a transmembrane helix.

Figure 10. Mrub_1198 and E. coli ppk do not contain TMH regions; Panel A shows the TMHMM for ppk; Panel B shows the TMHMM data for Mrub_1198. TMHMM Server v 2.0 http://www.cbs.dtu.dk/services/TMHMM was used to create these two hydropathy charts.
Figure 11. Mrub_1198 and *E. coli* ppk do not contain cleavage sites; D values for both plots were below the cutoff value in both cases. Panel A shows the plot for *ppk*; Panel B shows the plot for Mrub_1198. Signal P server v. 4.1 [http://www.cbs.dtu.dk/services/SignalP](http://www.cbs.dtu.dk/services/SignalP) created these plots.

The above plots in Figure 11 are SignalP graphs generated for both *E. coli* ppk and Mrub_1198. For *E. coli* ppk (Panel A), the D-value (0.105) is lower than the cutoff value (0.450), which means that this protein does not contain any cleavage sites. The same can be said for Mrub_1198 (Panel B), as the D-value of 0.114 is again below the cutoff of 0.450. Therefore, these data are consist for both genes, suggesting that neither have cleavage sites.
Figure 12. Mrub_1198 and E. coli ppk are present in the same biochemical pathway. It is predicted that both of these genes are involved with energy metabolism for oxidative phosphorylation, both coding for the enzyme polyphosphate kinase. Panel A shows the KEGG pathway after selecting for Escherichia coli. Panel B shows the KEGG pathway after selecting for M. ruber. The Kyoto Encyclopedia of Genes and Genomes (KEGG) database at http://www.genome.jp/kegg/ http://www.genome.jp/kegg/pathway.html was used to locate oxidative phosphorylation KEGG pathway maps.

Figure 12 shows the biochemical pathway that both of these genes are involved in. Enzymes that are colored in green are thought to be present in the organism. From the pathways, we see that E. coli ppk and Mrub_1198 are both involved with energy metabolism, both coding for the enzyme inorganic pyrophosphatase. This provides even more evidence to suggest that these organisms are evolutionarily related and that these two genes are orthologous.

Panel A
Figure 13. Mrub_1198 and *E. coli* ppk have the same highly conserved amino acids and code for the same protein, polyphosphate kinase. The alignment shown is for the C-terminus domains of the four-domain protein, which are catalytic. As we can see from this pairwise alignment, the residues at the positions P346, Y347, A360, Y374, and P375 seem to be the most conserved of the amino acids in the full sequence. Panel A shows the pairwise alignment for *Escherichia coli*. Panel B shows the pairwise alignment for *M. ruber*. The pairwise alignment was created using the Pfam website [http://pfam.sanger.ac.uk/search](http://pfam.sanger.ac.uk/search).

As we can see from the pairwise alignments in Figure 13, both *E. coli* ppk and Mrub_1198 contain the same highly conserved residues in the C-terminus domains. Rather than the two sequences being compared to one another as in the BLAST search, this pairwise alignment compares the provided sequence to a consensus sequence created from hundreds of other proteins. The fact that both the *E. coli* ppk and Mrub_1198 protein sequences both pulled out the same exact consensus sequence gives us even more information to support that the two genes might be orthologous to one another.

Figure 14. *E. coli* ppk and Mrub_1198 genes are not part of an operon. Chromosome Viewer maps were colored by KEGG. Panel A: *E. coli* ppk Chromosome Viewer; Panel B: Mrub_1198 Chromosome Viewer. Genes are underlined in red. Images were taken from [http://img.jgi.doe.gov/](http://img.jgi.doe.gov/)
The above images in Figure 14 give us even further information that these two genes are orthologous to one another. As noted by the differences in color from the other genes near them, neither of these genes are part of an operon. Interestingly, both genes were not given the same color for *E. coli* and *M. ruber*. According to the gene details page of the analysis, polyphosphate kinase plays two different roles, energy metabolism and RNA degradation (Panel C). That is why in Panel A the gene is colored purple (energy metabolism) while in Panel B it is colored green (degradation). The fact that both genes are not part of an operon and both have this dual function further suggests that these two genes are orthologous to one another.

**CONCLUSION**

The results obtained from this study revealed that *E. coli* b2446 and Mrub_0258 are orthologous genes, which means that these organisms are related through a common ancestry. The same conclusion can be drawn for the genes *ppk* and Mrub_1198, as the bioinformatics tools also suggest that these two genes are orthologous. Evidence for these linkages was first suggested by the BLAST analyses comparing the protein sequences of the *E. coli* genes to that of *M. ruber* genes. Further confirming this result were cellular location bioinformatics tools such as TMH, SignalP, LipoP, and PSORT-B, which all indicated a cellular location for genes b2446 and Mrub_0258. Research publications suggest that the genes *ppk* and Mrub_1198 are located in the outer membrane of the cell. While TIGRfam data could not be obtained for the b2446 or Mrub_0258 genes, Pfam data were consistent with one another, pulling up the single domain protein inorganic pyrophosphate. Additionally, TIGRfam and Pfam consistently matched the protein sequences of *ppk* and Mrub_1198 to that of polyphosphate kinase and the domains of which the protein is composed. There were also additional bioinformatics programs utilized for this project that also presented the same results for the pairs of genes being studied. In fact, the only deviation between gene pairs were the cellular location data for *ppk* and Mrub_1198. A literature search helped to confirm which of the bioinformatics programs was correct. Based on the frequency with which the bioinformatics data matched up between the gene pairs, we can be pretty certain that *E. coli ppa* and Mrub_0258 are orthologous to one another. The same can be said for the *ppk* gene and the Mrub_1198 gene as well.


Scott LR. Meiothermus ruber Genome Analysis Project. [Internet]. GENI-ACT; [2015 Dec 16]. Available from: http://geni-science.org/secure/projects/view/


