2018

Examination of Orthologous Genes (Mrub_2518 and b3728, Mrub_2519 and b3727, Mrub_2520 and b3726, Mrub_2521 and b3725) Responsible for ABC Phosphate Transporters in Two Species *M. ruber* and *E. coli*

Margaret Meyer
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Examination of Orthologous Genes (Mrub_2518 and b3728, Mrub_2519 and b3727, Mrub_2520 and b3726, Mrub_2521 and b3725) Responsible for ABC Phosphate Transporters in Two Species *M. ruber* and *E. coli*

Margaret Meyer
Dr. Lori Scott
Augustana College

Introduction

ABC Transporters

Arguably the one of the most important functions of a cell is transport. Cells transport molecules to communicate, maintain gradients, gain nutrients and remove waste. While this can be accomplished in a number of ways, active transport allows for molecules to be moved against their gradient, which is vital for normal cell operations. ABC transporters are protein complexes that use ATP to actively transport molecules across membranes. These transporters are so important to the function of a cell that between 1 and 3% of the bacterial and archaeal genome is dedicated solely to ABC transport protein subunits (Wilkens, 2015).

Phosphate Transporters

In this article we will be focusing on the phosphate ABC transporter. In *Escherichia coli* (*E. coli*), as well as other gram negative bacteria, it is known that there are two cell membranes separated by a periplasmic space (Tomii, 1998). The outer membrane is fairly rigid, yet quite porous, allowing many larger molecules, such as ions or sugars, to pass through unaided. The inner membrane, which separates the cytoplasm of the cell from the periplasmic space, is quite different. It does not allow for large molecules to cross, unless through a carrier protein, and therefore maintains many chemical gradients. In *E. coli,* phosphate ABC transporters are active carrier proteins that use ATP to move a single inorganic phosphate ion from the periplasmic space into the cytoplasm of the cell (Rao, 1990).

For the annotations presented in this article the *E. coli* system is used as a comparison system to the model organism. As the phosphate ABC transport system in *E. coli* is well known, this allows us to make predictions about the model organism and possible common ancestors between the two. The phosphate ABC transporter system in *E. coli* is made up of four subunits, which are commonly labeled PstA, PstB, PstC and PstS. These protein subunits are encoded by the genes *pstA, pstB, pstC,* and *pstS* (locus tags: b3726, b3725, b3727 and b3728 respectively), and the entire system is considered the *pst* operon (Rao, 1990).

The protein complex is made up of two membrane embedded proteins, one free floating binding protein (within the periplasm), and one cytoplasmic protein. The free floating binding protein, PstS, moves throughout the periplasmic space and binds with inorganic phosphates. Once bound, PstS will return and bind to the two membrane embedded proteins, PstA and PstC. These two proteins create the channel through the membrane, and the phosphate is released from PstS and allowed to move through PstA and PstC. PstB is the cytoplasmic unit, and is permanently attached to PstA and PstC. This unit is responsible for binding to ATP, and gathering the energy needed to transport the phosphate. Using this energy from the ATP, PstA and PstC move the phosphate into the cytoplasm of the cell (Rao, 1990, Rees, 2009).

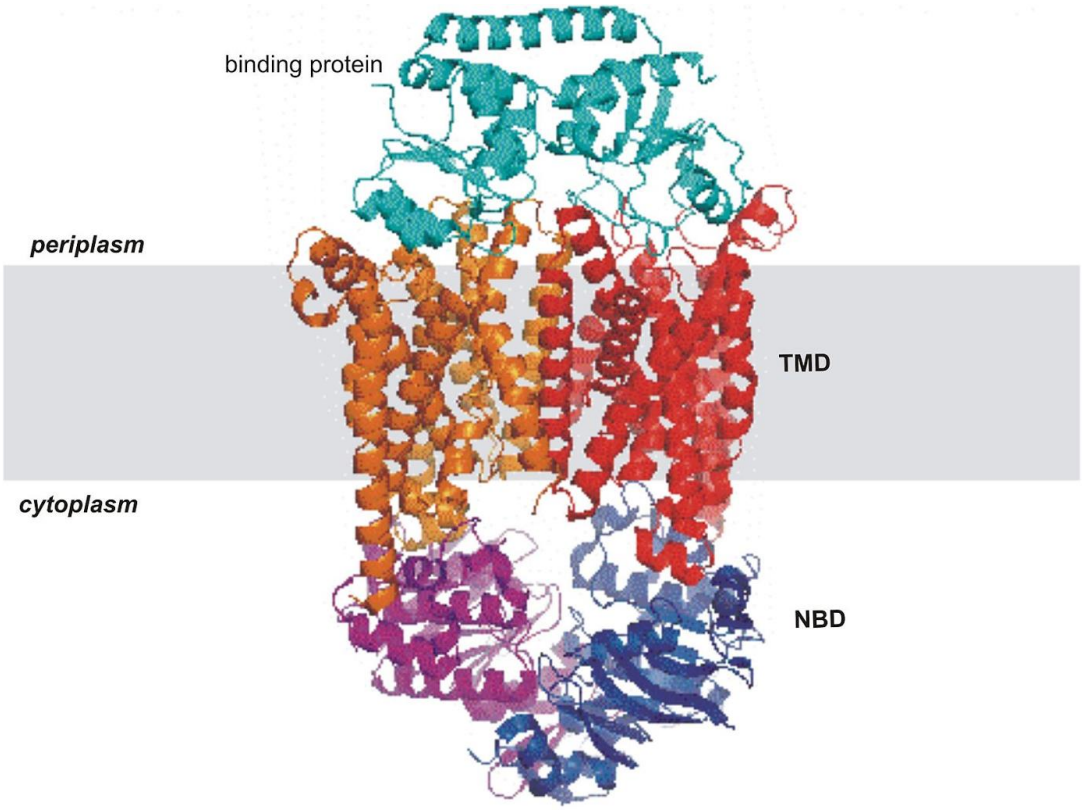


Figure 1: The structure of a common ABC transporter is shown. The light blue portion within the periplasm represents the binding protein (PstS), the orange and red portions are the transmembrane domains (PstA and PstC), and the dark blue and purple portions are the nucleotide binding domains (PstB), or where the ATP binds within the cytoplasm of the cell. While some ATP transporters have two subunits in the NBD as shown, the phosphate ABC transporter that is focused on within this article only has a single NBD subunit within the cytoplasm (Berman, 2000).

Model Organism: *Meiothermus ruber*

The organism studied in this article is *Meiothermus ruber* (*M. ruber*). *M. ruber* is a gram negative thermophilic bacteria. *M. ruber* is rod shaped, non-motile and prefer temperatures around 60 degrees Celsius (Tindall, 2010). This bacterium has a genome with 3,097,457 bp, 5.5% of which is devoted to inorganic ion transport and metabolism (Tindall, 2010). *M. ruber* was chosen for this project as it is has been studied little in the past, although it's genome has been sequenced. This allows for a wide range of studies to be done on many different genes of the bacteria. Four *M. ruber* genes will be discussed here, Mrub_2518, Mrub_2519, Mrub_2520 and Mrub_2521. We predict that these genes are an operon for the phosphate ABC transporter in *M. ruber*. We also predict that the four genes are orthologous to the *E. coli* phosphate ABC transporter genes (Mrub_2518 and b3728, Mrub_2519 and b3727, Mrub_2520 and b3726, Mrub_2521 and b3725).

Materials and Methods

Finding Potential Orthologs

As one of our main goals was to determine if the two species had orthologous genes we started by searching the online bioinformatic tools KEGG (Kanehisa *et al.,* 2016) and Blast (Madden, 2002). Within KEGG we searched for the ABC transporter pathway, which shows every known gene pathway for different ABC transporters in a large number of species. Once the phosphate pathway was found, each gene for each species was verified to be within the pathway by searching for species specific (*M. ruber* or *E.coli*) pathways. Genes could be confirmed by clicking on the appropriate pathway.

Blast was used for each *E. coli* gene to check for orthologs in the opposing species. For each gene, the amino acid sequence (found from KEGG information on individual genes) was pasted into an input box and was "Protein Blasted" against the entire genetic code of *M. ruber*, omitting any species except for the target. The Blast program returned any similar sequence found in the genome, and any matching sequence was then examined to see if it was a viable option as an ortholog. KEGG and Blast also are good tools for background on the gene of interest. If more information about the *E. coli* genes is needed the cite ecocyc provides a number of pathways and other background information (Keseler *et al*., 2013).

Confirming Start Codons

Many of the bioinformatics tools used in this article require the correct nucleotide or amino acid sequence of the gene to produce an output. As such it is important that the sequences obtained from KEGG were confirmed to be correct. Start codons are often points within the sequence that are mislabeled by online tools. By first confirming the start codon of any gene being worked with, you can then proceed with any other tests on the gene knowing that the sequence is correct. To confirm the start codon the program IMG (Markowitz *et al.*, 2012) was used. Each genes locus tag was searched under the Gene Search section of the program. The results page for each gene includes an enormous amount of information about the gene, but for this purpose a link entitled "Sequence Viewer For Alternate ORF Search" was used. This link allows the user to view any desired number of nucleotides upstream or downstream of the target gene. By viewing the nucleotides upstream of the gene, the user can find potential shine-dalgarno sequences and view them in relation to potential start codons.

If there are a number of possible start codons or the shine-dalgarno sequence is absent it may be helpful to confirm the start codon in other ways using the tools TCoffee (Notredame *et al.,* 2000) and WebLogo (Crooks *et al.,* 2004). Using Blast find a variety of similar sequences to the gene of interest and download their protein sequences. These sequences can be put into the input section of TCoffee, which will align the sequences. This input can then be used for WebLogo. The TCoffee results should show the user the exact alignment of the sequences, and the WebLogo output will show the user a graphic depicting the most commonly occurring amino acids at each position. If the sequences are well aligned and the expected start codon very common, then it is likely that this codon is correct. While TCoffee and WebLogo can be helpful tools, they were not employed for these genes, as all sequences studied were found to have a single start codon within range of the shine-dalgarno sequence. The genes for *E. coli* were not examined for start codons, as these genes have already been extensively studied are unlikely to be incorrect.

Cell Features

Once a gene's sequence has been confirmed, this sequence can be utilized with several other tools to gain a better understanding of the gene's function. We first set out to better understand the predicted location of the protein within the cell. For the *E. coli* genes, this process was simply a conformation of an already known location as *E. coli* has already been widely studied. However, for the *M. ruber* genes this process was more vital as the location and function of these genes was only suspected and not confirmed. To start we utilized the program TMHMM (Krogh *et al.*, 2016). By inputting the amino acid sequence of the gene of interest, this program will predict the number of transmembrane helices within your protein. The next program, SignalP, will predict the appearance of signal peptides (Petersen *et al.*, 2011), while LipoP will predict a general location of the protein (Juncker *et al*., 2003). The PSORT-B program gives a more precise location prediction (Yu *et al*., 2010), and Phobius will output a graphic showing location probability (Kall *et al*., 2004).

To predict structure (and therefore function) of each gene, we returned to the program Blast. On the results page of a blast output, there is a graphic depicting the gene. If this graphic is clicked on it will take the user to a page with more gene information called the CDD (Marchler-Bauer *et al.*). Here the user should look for genes labeled COG (cluster of orthologous groups), as each COG label represents genes of a similar function to the gene of interest. After looking at the COG's, inputting the amino acid sequence into the programs TIGRFAM (Haft *et al.*, 2001) and PFAM (Finn *et al*., 2016) will give outputs showing the domain, clan, and/or family names of the genes. Knowing the categories that the genes fall into can help predict their function. Lastly the program PDB will give the user a list of similar functioning genes to the gene of interest, when using the amino acid sequence to search its database (Berman, 2000).

Finding Potential Operons

The last goal of this project was to confirm that the genes of interest were indeed part of an operon. One way to do this is to return to the IMG cite. On this page directly under the "Sequence Viewer for Alternate ORF Search" link is a drop down menu. On the menu select KEGG, and this link will bring you to a graphic depicting all of the genes surrounding your gene of interest. This allows you to confirm that the genes being studied are in the same area or even in line with each other. The genes in an operon should have locus tags that are in order, however, double checking this with IMG allows for a visual of the entire operon and other surrounding genes.

Results

Potential Ortholog Results

The KEGG program provided two pathways, one for each species. Each pathway showed the genes involved with phosphate ABC transporters. The first, the *E. coli* pathway, shows the genes b3728, b3727, b3726 and b3725 (labeled as PstS, PstC, PstA, and Pst B respectively). The second, the *M. ruber* pathway, shows the genes Mrub_2518, Mrub_2519, Mrub_2520 and Mrub_2521 (labeled as PstS, PstC, PstA and PstB respectively). This preliminary step shows that the following genes are most likely orthologs: b3728 and Mrub_2518, b3727 and Mrub_2519, b3726 and Mrub_2520, and b3725 and Mrub_2521.

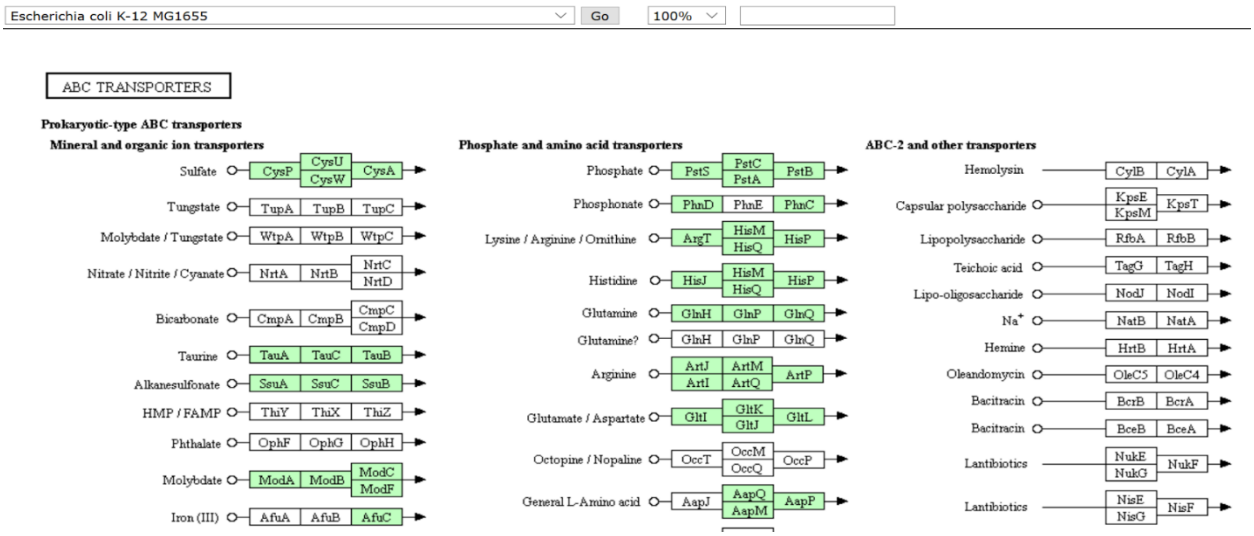


Figure 2: ABC transporter pathways for *E. coli*. Phosphate transporter pathway can be seen at the top of the middle column highlighted in green.

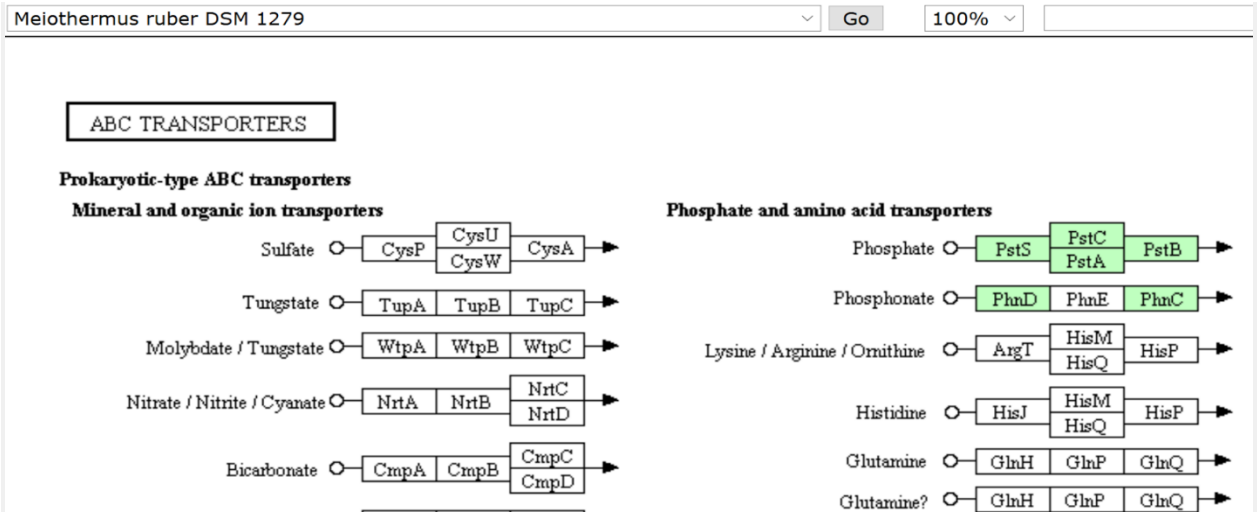
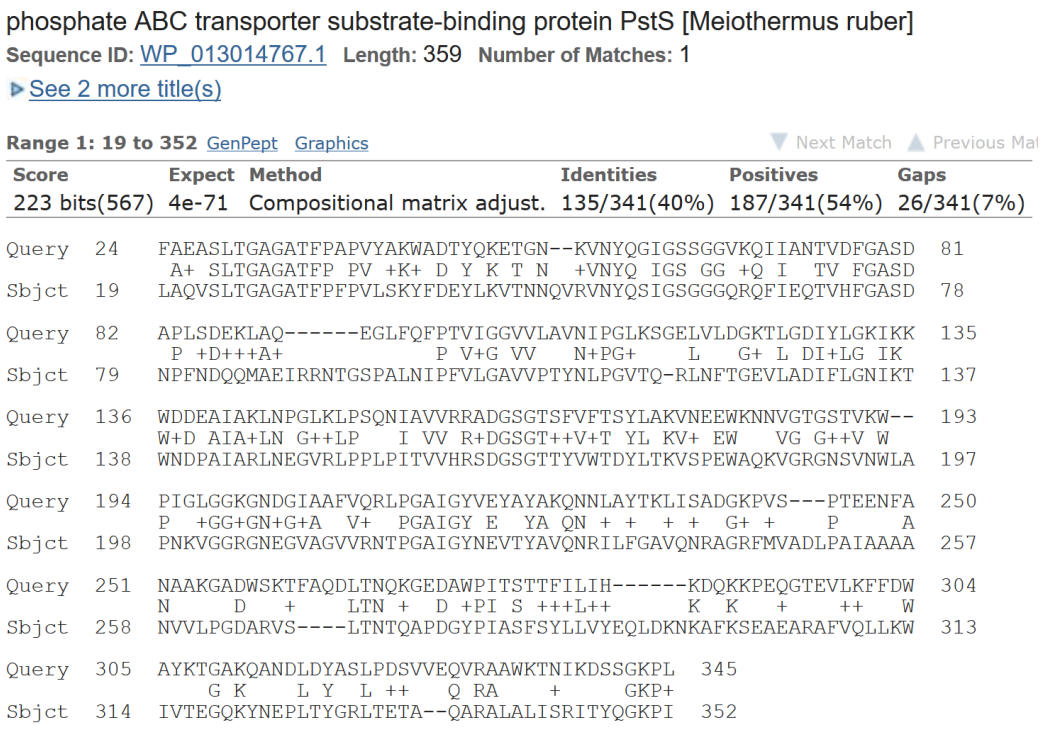


Figure 3: ABC transporter pathways for *M. ruber*. Phosphate pathway can be seen at the top of the second column highlighted in green.

Confirmation of these orthologous genes came from the Blast results, shown in the figures below. The *E. coli* genes all returned matches to the suspected orthologous *M. ruber* genes.



phosphate ABC transporter substrate-binding protein PstS [Meiothermus ruber]

Sequence ID: WP_013014767.1 **Length:** 359 **Number of Matches:** 1

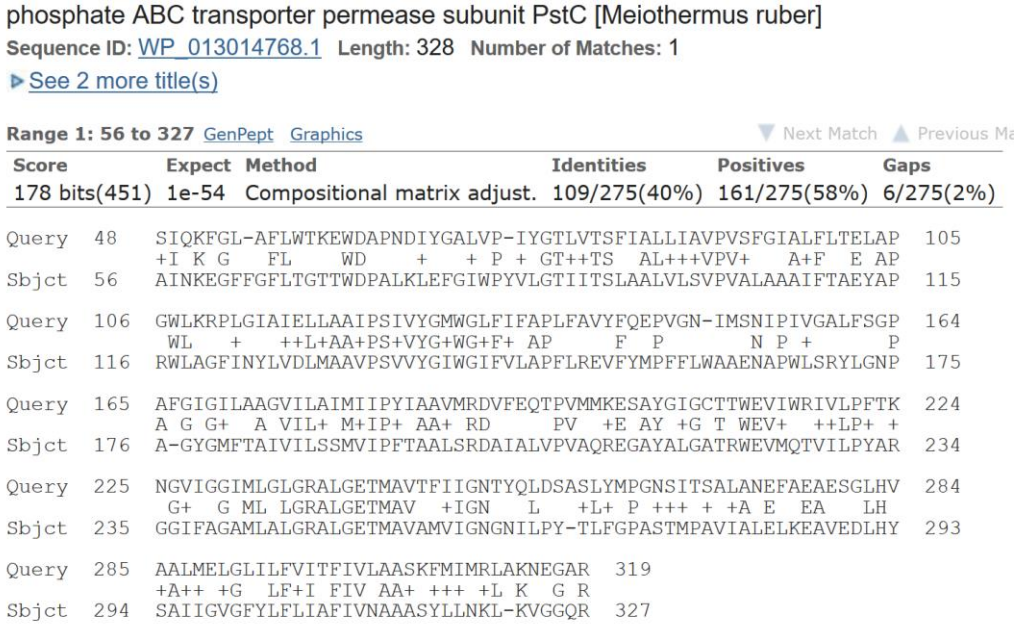
See 2 more title(s)

Range 1: 19 to 352 GenPept Graphics Next Match Previous Mat

Score	Expect	Method	Identities	Positives	Gaps
223 bits(567)	4e-71	Compositional matrix adjust.	135/341(40%)	187/341(54%)	26/341(7%)

```
Query  24   FAEASLTGAGATFPAPVYAKWADTYQKETGN--KVNYQGIGSSGGVKQIIANTVDFGASD  81
            A+ SLTGAGATFP PV +K+ D Y K T N  +VNYQ IGS GG +Q I  TV FGASD
Sbjct  19   LAQVSLTGAGATFPFPVLSKYFDEYLKVTNNQVRVNYQSIGSGGGQRQFIEQTVHFGASD  78

Query  82   APLSDEKLAQ------EGLFQFPTVIGGVVLAVNIPGLKSGELVLDGKTLGDIYLGKIKK  135
             P +D+++A+            P V+G VV   N+PG+    L   G+ L DI+LG IK
Sbjct  79   NPFNDQQMAEIRRNTGSPALNIPFVLGAVVPTYNLPGVTQ-RLNFTGEVLADIFLGNIKT  137

Query  136  WDDEAIAKLNPGLKLPSQNIAVVRRADGSGTSFVFTSYLAKVNEEWKNNVGTGSTVKW--  193
            W+D AIA+LN G++LP   I VV R+DGSGT++V+T YL KV+ EW   VG G++V W
Sbjct  138  WNDPAIARLNEGVRLPPLPITVVHRSDGSGTTYVWTDYLTKVSPEWAQKVGRGNSVNWLA  197

Query  194  PIGLGGKGNDGIAAFVQRLPGAIGYVEYAYAKQNNLAYTKLISADGKPVS---PTEENFA  250
            P  +GG+GN+G+A  V+  PGAIGY E  YA QN + +  + +  G+ +    P     A
Sbjct  198  PNKVGGRGNEGVAGVVRNTPGAIGYNEVTYAVQNRILFGAVQNRAGRFMVADLPAIAAAA  257

Query  251  NAAKGADWSKTFAQDLTNQKGEDAWPITSTTFILIH------KDQKKPEQGTEVLKFFDW  304
            N     D   +    LTN +  D +PI S +++L++      K  K   +    ++   W
Sbjct  258  NVVLPGDARVS----LTNTQAPDGYPIASFSYLLVYEQLDKNKAFKSEAEARAFVQLLKW  313

Query  305  AYKTGAKQANDLDYASLPDSVVEQVRAAWKTNIKDSSGKPL  345
                G K    L Y  L ++   Q RA    +     GKP+
Sbjct  314  IVTEGQKYNEPLTYGRLTETA--QARALALISRITYQGKPI  352
```

Figure 4: Blast of b3728 against the *M. ruber* genome resulted in a hit with the *M. ruber pstS* gene.



phosphate ABC transporter permease subunit PstC [Meiothermus ruber]

Sequence ID: WP_013014768.1 **Length:** 328 **Number of Matches:** 1

See 2 more title(s)

Range 1: 56 to 327 GenPept Graphics Next Match Previous Ma

Score	Expect	Method	Identities	Positives	Gaps
178 bits(451)	1e-54	Compositional matrix adjust.	109/275(40%)	161/275(58%)	6/275(2%)

```
Query  48   SIQKFGL-AFLWTKEWDAPNDIYGALVP-IYGTLVTSFIALLIAVPVSFGIALFLTELAP  105
            +I K G   FL    WD    +   + P + GT++TS  AL+++VPV+   A+F  E AP
Sbjct  56   AINKEGFFGFLTGTTWDPALKLEFGIWPYVLGTIITSLAALVLSVPVALAAAIFTAEYAP  115

Query  106  GWLKRPLGIAIELLAAIPSIVYGMWGLFIFAPLFAVYFQEPVGN-IMSNIPIVGALFSGP  164
             WL   +   ++L+AA+PS+VYG+WG+F+ AP     F  P       N P +      P
Sbjct  116  RWLAGFINYLVDLMAAVPSVVYGIWGIFVLAPFLREVFYMPFFLWAAENAPWLSRYLGNP  175

Query  165  AFGIGILAAGVILAIMIIPYIAAVMRDVFEQTPVMMKESAYGIGCTTWEVIWRIVLPFTK  224
            A G G+  A VIL+ M+IP+ AA+ RD     PV  +E AY +G T WEV+  ++LP+ +
Sbjct  176  A-GYGMFTAIVILSSMVIPFTAALSRDAIALVPVAQREGAYALGATRWEVMQTVILPYAR  234

Query  225  NGVIGGIMLGLGRALGETMAVTFIIGNTYQLDSASLYMPGNSITSALANEFAEAESGLHV  284
             G+  G ML LGRALGETMAV  +IGN   L   +L+ P +++ + +A E  EA   LH
Sbjct  235  GGIFAGAMLALGRALGETMAVAMVIGNGNILPY-TLFGPASTMPAVIALELKEAVEDLHY  293

Query  285  AALMELGLILFVITFIVLAASKFMIMRLAKNEGAR  319
            +A++ +G  LF+I FIV AA+ +++ +L K  G R
Sbjct  294  SAIIGVGFYLFLIAFIVNAAASYLLNKL-KVGGQR  327
```

Figure 5: Blast of b3727 against the *M. ruber* genome resulted in a hit with the *M. ruber pstC* gene.

phosphate ABC transporter permease PtsA [Meiothermus ruber]

Sequence ID: WP_013014769.1 **Length:** 276 **Number of Matches:** 1

See 1 more title(s)

Range 1: 2 to 249 GenPept Graphics Next Match Previous

Score	Expect	Method	Identities	Positives	Gaps
160 bits(405)	1e-48	Compositional matrix adjust.	92/250(37%)	152/250(60%)	2/250(0%)

```
Query  17   RKMQARRRLKNRIALTLSMATMAFGLFWLIWILMSTITRGIDGMSLALFTEMTPPPNTEG  76
            R ++AR   ++R+ L + +         L  +L   +  G   ++    T+   PP   G
Sbjct  2    RNLRARYT-RDRLILIVVIIGTILAALPLTLVLGYALVNGFSSLNWDFLTQGPKPPGELG  60

Query  77   GGLANALAGSGLLILWATVFGTPLGIMAGIYLAEYGRKSWLAEVIRFINDILLSAPSIVV  136
            GG+A A+ G+ ++     +  TP GI AGI LAEY     L   +R ++D L   P+I+
Sbjct  61   GGMAPAIVGTLIITGAGLLMATPFGIGAGILLAEYPDNK-LNPTLRLLSDTLNGMPAILK  119

Query  137  GLFVYTIVVAQMEHFSGWAGVIALALLQVPIVIRTTENMLKLVPYSLREAAYALGTPKWK  196
            GL  Y +VV     FSG +G +A+A + +PI+ +TTE++LKLVP+++REA  ALG P+W+
Sbjct  120  GLLAYVLVVKAQGSFSGLSGALAMAFIMIPIIAKTTESVLKLVPWNIREAGLALGLPRWR  179

Query  197  MISAITLKASVSGIMTGILLAIARIAGETAPLLFTALSNQFWSTDMMQPIANLPVTIFKF  256
            +I ++ L A+ +G++TG+LLA AR AGE APL+FTA  N   + +++QP+  LP+ ++ +
Sbjct  180  VIMSLVLPAARAGVVTGLLLATARAAGEAAPLIFTAFGNSLLTYNLLQPMDALPLRLYAY  239

Query  257  AMSPFAEWQQ  266
            A+SP+ +W +
Sbjct  240  AISPYEDWHR  249
```

Figure 6: Blast of b3726 against the *M. ruber* genome resulted in a hit with the *M. ruber pstA* gene.

phosphate ABC transporter, ATPase subunit [Meiothermus ruber DSM 1279]

Sequence ID: ADD29272.1 **Length:** 273 **Number of Matches:** 1

Range 1: 16 to 273 GenPept Graphics Next Match Previous M

Score	Expect	Method	Identities	Positives	Gaps
290 bits(741)	7e-100	Compositional matrix adjust.	137/258(53%)	181/258(70%)	1/258(0%)

```
Query  1    MSMVETAPSKIQVRNLNFYYGKFHALKNINLDIAKNQVTAFIGPSGCGKSTLLRTFNKMF  60
            M++  +  ++I+ R +   YG+   +K++++ I  N+VTA IGPSGCGK+T LR  N+M
Sbjct  16   MTLDPSLKTRIEARRVTVMYGQKAGVKDVDMPIYANKVTALIGPSGCGKTTFLRALNRMH  75

Query  61   ELYPEQRAEGEILLDGDNILTNSQDIALLRAKVGMVFQKPTPFP-MSIYDNIAFGVRLFE  119
            +L P  R  GE+LLDG N+     D   +R K+GMVFQKP PFP +SIY N+  G+RL
Sbjct  76   DLTPSARVTGEVLLDGINVYAPGVDPVEVRRKIGMVFQKPNPFPTLSIYGNVVAGLRLVG  135

Query  120  KLSRADMDERVQWALTKAALWNETKDKLHQSGYSLSGGQQQRLCIARGIAIRPEVLLLDE  179
               ++ +DE V+ AL++AALW+E KD+L+    SLSGGQQQRLCIAR +A+ PEVLL+DE
Sbjct  136  IRKKSILDEAVERALSQAALWDEVKDRLNAPSMSLSGGQQQRLCIARALAVEPEVLLMDE  195

Query  180  PCSALDPISTGRIEELITELKQDYTVVIVTHNMQQAARCSDHTAFMYLGELIEFSNTDDL  239
            P SALDPIST  IE+L+ ELK   T+VIVTHNMQQA R SD T +   GE++EF  T  L
Sbjct  196  PTSALDPISTQSIEDLLNELKNHVTIVIVTHNMQQAGRVSDFTGYFLNGEMVEFGPTSLL  255

Query  240  FTKPAKKQTEDYITGRYG  257
            FT P  K+TE YITGR+G
Sbjct  256  FTTPKDKRTEAYITGRFG  273
```

Figure 7: Blast of b3725 against the *M. ruber* genome resulted in a hit with the *M. ruber pstB* gene.

Confirming the Start Codon

IMG results are shown from the four *M. ruber* genes in the figures below. All were found to have the correct start codon. The red ATG was the predicted start codon, and as it was always in the correct range following the shine-dalgarno sequence, it remained the proper start codon.

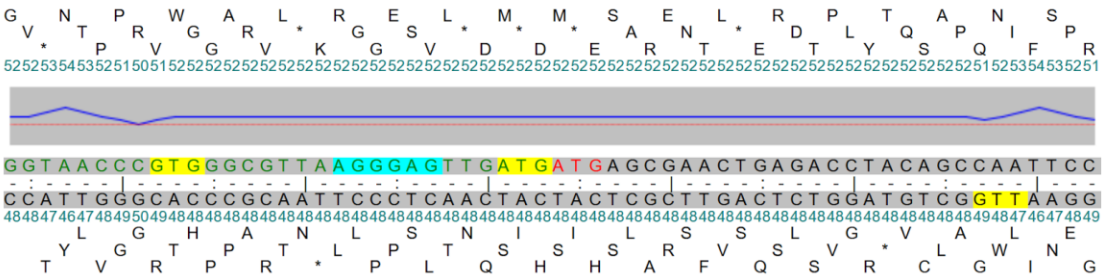


Figure 8: IMG results from Mrub_2521. Blue highlights the shine-dalgarno sequence, red is the predicted start codon, yellow is all other possible start codons.

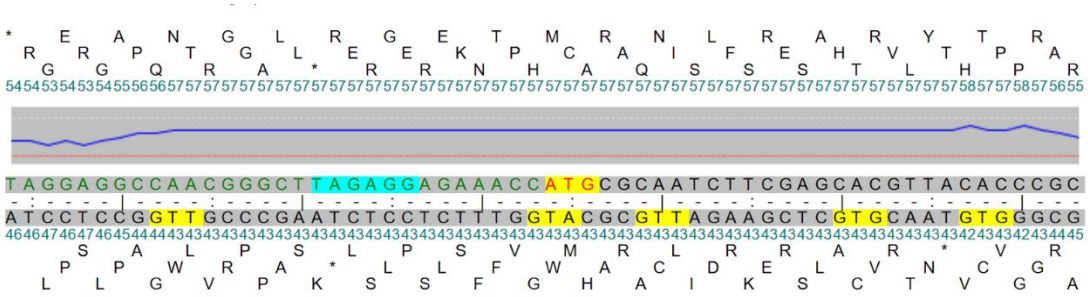


Figure 9: IMG results from Mrub_2520. Blue highlights the shine-dalgarno sequence, red is the predicted start codon, yellow is all other possible start codons.

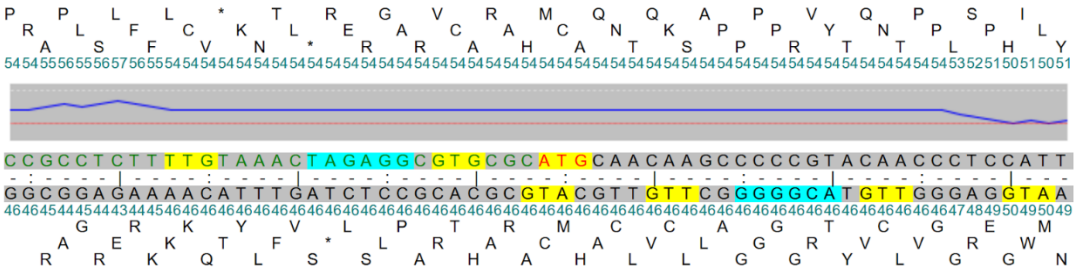


Figure 10: IMG results from Mrub_2519. Blue highlights the shine-dalgarno sequence, red is the predicted start codon, yellow is all other possible start codons.

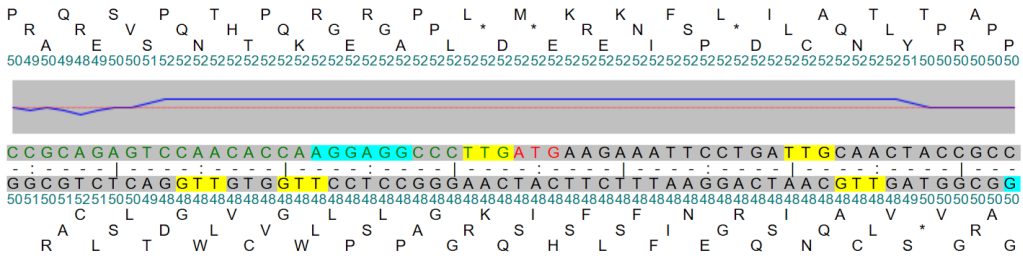


Figure 11: IMG results from Mrub_2518. Blue highlights the shine-dalgarno sequence, red is the predicted start codon, yellow is all other possible start codons.

Protein Location and Function

Orthologs b3728 and Mrub_2518 are both responsible for PstS, the phosphate binding protein within the periplasm. This is evident from the multiple sources below, which each stated the protein's location to be within the periplasm, or stated that the protein was a part of the binding protein family.

Table 1: Data from b3728 and Mrub_2518 genes.

Bioinformatics Tool Used	*E. coli* b3728 gene	Mrub_ 2518 gene
BLAST E-value	4e-71	---
Blast Score	223 bits	---
CDD Data (COG category)	PstS (COG0226)	PstS (COG0226)
Cellular Localization (PSORT-B)	Periplasm	Periplasm
TIGRfam-protein family	Phosphate ABC Transporter (TIGR00975)	Phosphate ABC Transporter (TIGR00975)
Pfam-protein family	Periplasmic Binding Protein (Pf12849.6)	Periplasmic Binding Protein (Pf12849.6)
Protein Database	4G5D	4G5D
KEGG pathway map	Phosphate ABC Transport Pst (02010)	Phosphate ABC Transport Pst (02010)

The following TMHMM output and SignalP outputs give more evidence for these two proteins being in the periplasm as a binding protein. The TMHMM shows that the majority of the protein is outside of the inner membrane. The SignalP shows that these proteins are considered signal peptides, with possible cleavage sites, both of which indicate a periplasmic protein.

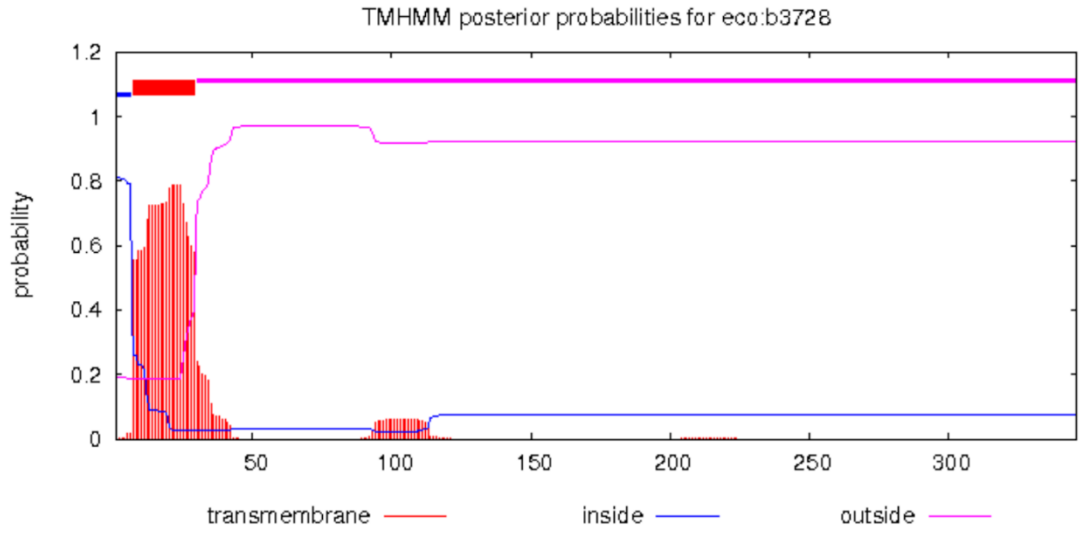


Figure 12: TMHMM output for b3728. Shows one possible transmembrane helices.

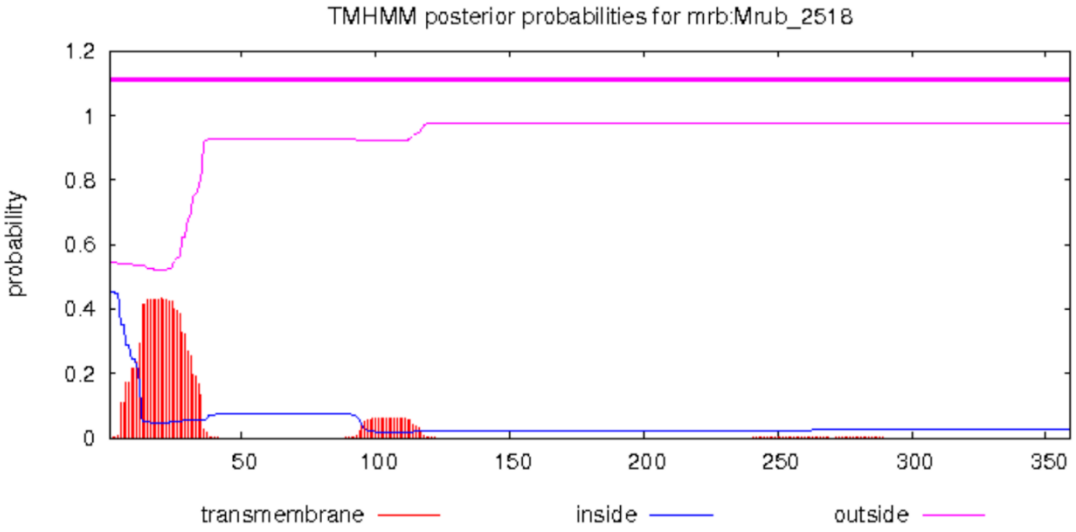


Figure 13: TMHMM output for Mrub_2518. Shows one possible transmembrane helices.

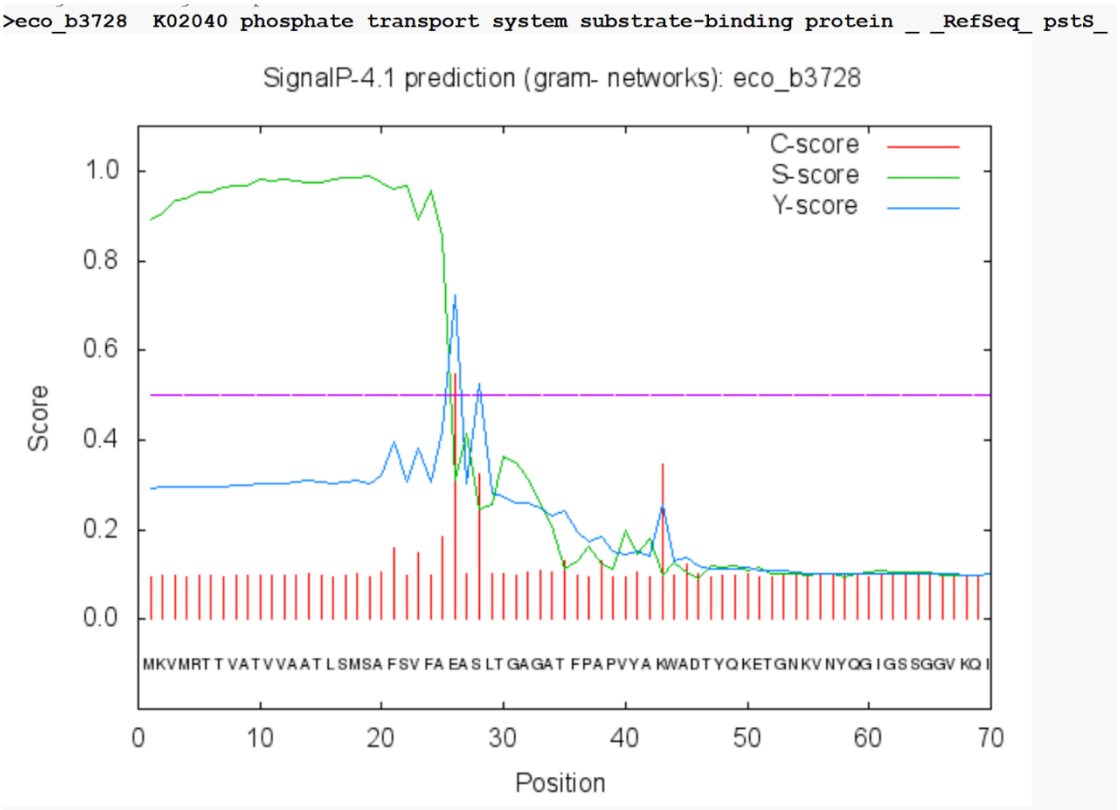


Figure 14: SignalP output for b3718. Shows a probable cleavage site between aa #25 and 26.

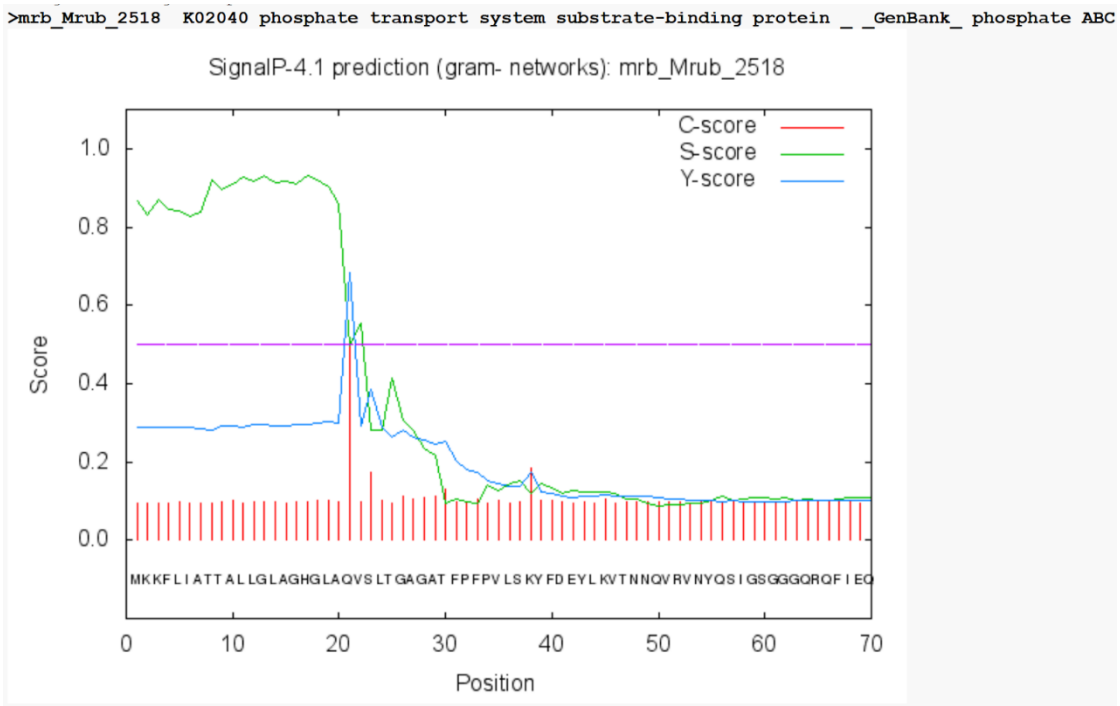


Figure 15: SignalP output for Mrub_2518. Shows a possible cleavage site between aa# 20 and 21.

Orthologs b3727 and Mrub_2519 are both responsible for PstC one of two transmembrane proteins in the phosphate ABC transporter. This protein is located within the cytoplasmic membrane, as is evident from the multiple sources below. These sources all claim these two genes as a part of the ABC phosphate transport system.

Table 2: Data from b3727 and Mrub_2519 genes.

Bioinformatics Tool Used	*E. coli* b3727 gene	Mrub_ 2519gene
BLAST E-value	1e-54	---
Blast Score	178 bits	---
CDD Data (COG category)	ABC phosphate transport sys permease component (COG0573)	ABC phosphate transport sys permease component (COG0573)
Cellular Localization (PSORT-B)	Cytoplasmic Membrane	Cytoplasmic Membrane
TIGRfam-protein family	Phostphate ABC Transporter PstC (TIGR02138)	Phostphate ABC Transporter PstC (TIGR02138)
Pfam-protein family	Binding protein dependent transport (Pf00528.21)	Binding protein dependent transport (Pf00528.21)
Protein Database	1PW4	2YZ2
KEGG pathway map	Phosphate ABC Transport Pst (02010)	Phostphate ABC Transport Pst (02010)

Further evidence for these two proteins being located within the cytoplasmic membrane can be found below in the TMHMM and SignalP outputs. Having any transmembrane helices indicates that the protein is embedded within a membrane, and because this is a transport protein, it is most likely going to be within the cytoplasmic membrane.

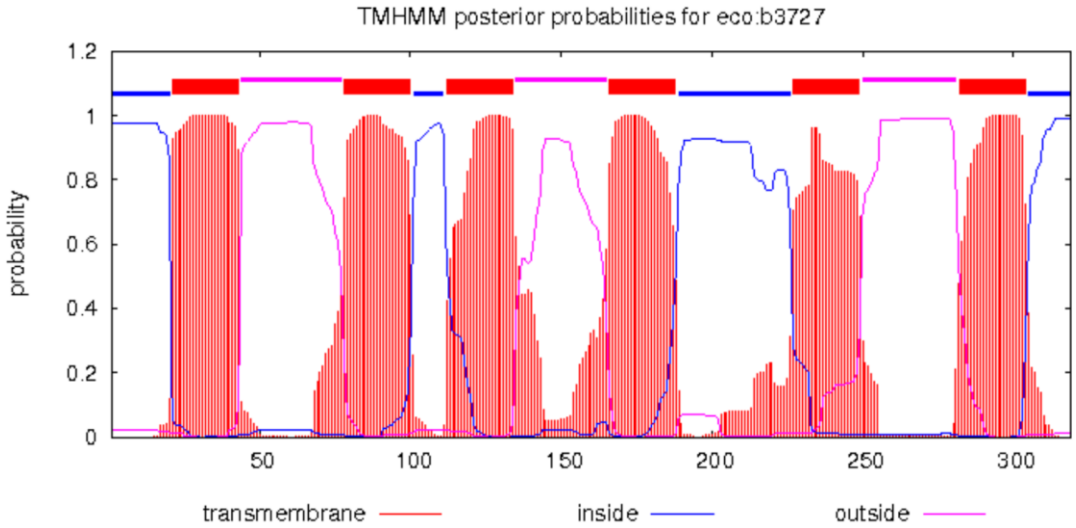


Figure 16: TMHMM output for b3727. This shows that the protein has six transmembrane helices, which indicates its location within the cytoplasmic membrane.

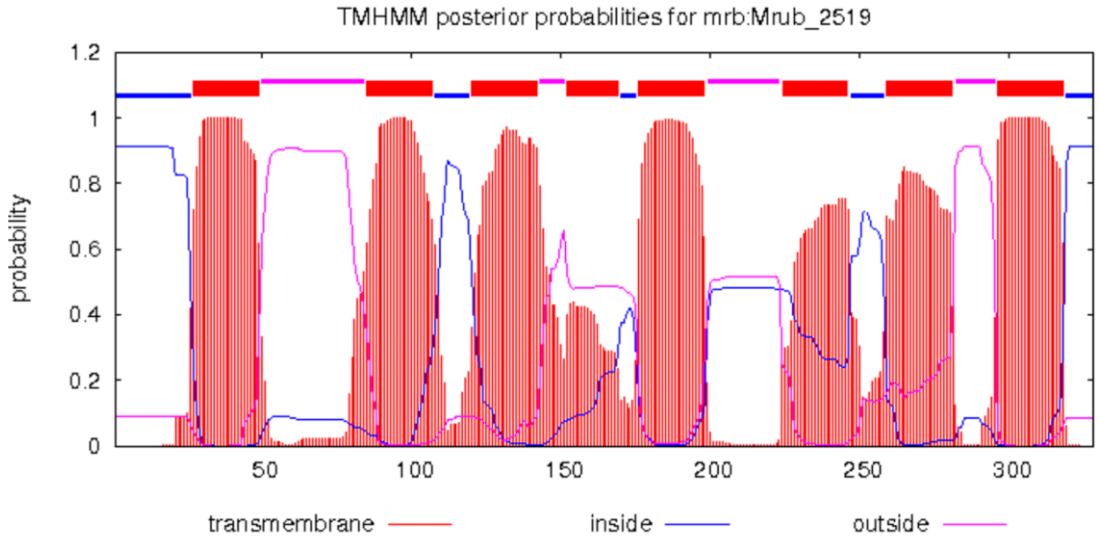


Figure 17: TMHMM output for Mrub_2519. This shows that the protein has six transmembrane helices, which indicates its location within the cytoplasmic membrane.

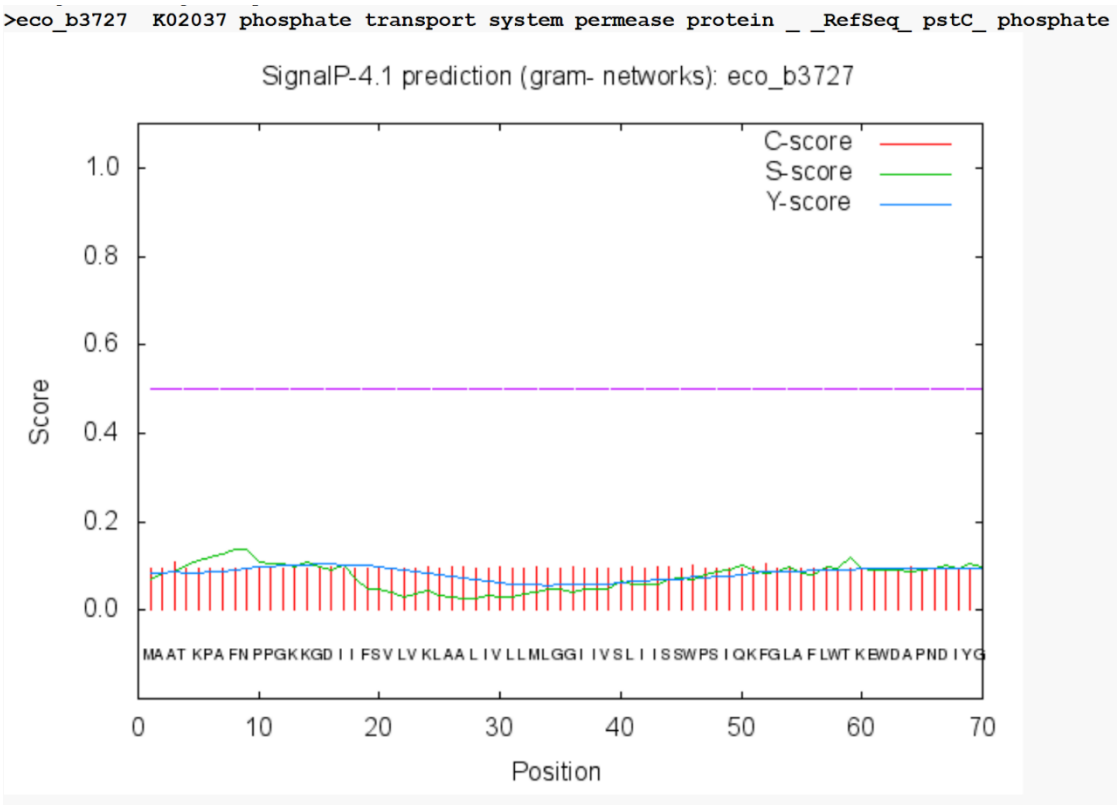


Figure 18: The SignalP output for b3727. Shows that there are no signal proteins or cleavages, indicating that the protein is not within the periplasm or outside the outer membrane.

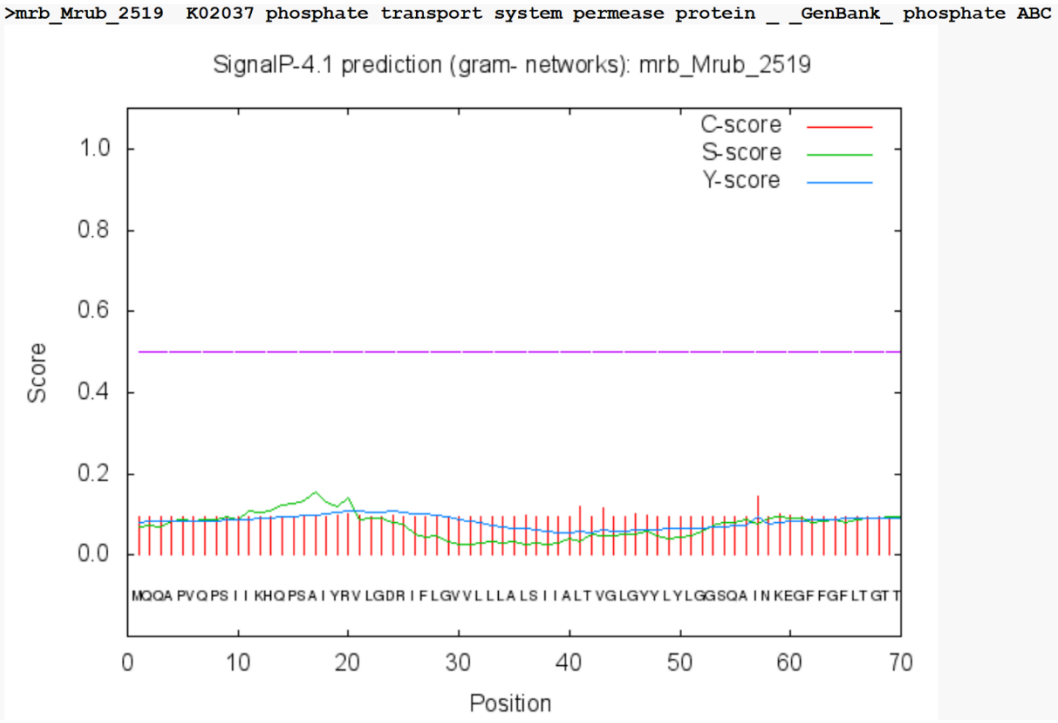


Figure 19: The SignalP output for Mrub_2519. Shows that there are no signal proteins or cleavages, indicating that the protein is not within the periplasm or outside the outer membrane.

Orthologs b3726 and Mrub_2520 are both responsible for PstA one of two transmembrane proteins in the phosphate ABC transporter. This protein is located within the cytoplasmic membrane, as is evident from the multiple sources below. These sources all claim these two genes as a part of the ABC phosphate transport system.

Table 3: Data from b3726 and Mrub_2520 genes.

Bioinformatics Tool Used	*E. coli* b3726 gene	Mrub_2520 gene
BLAST E-value	1e-48	---
Blast Score	160 bits	---
CDD Data (COG category)	ABC-type phosphate transport system, PstA (COG0581)	ABC-type phosphate transport system, PstA (COG0581)
Cellular Localization (PSORT-B)	Cytoplasmic Membrane	Cytoplasmic Membrane
TIGRfam-protein family	Phostphate ABC Transporter (TIGR00974)	Phostphate ABC Transporter (TIGR00974)
Pfam-protein family	BPD Transport (Pf00528.21)	BPD Transport (Pf00528.21)
Protein Database	1PW4	2YZ2
KEGG pathway map	Phosphate ABC Transporter Pst (02010)	Phosphate ABC Transporter Pst (02010)

Further evidence for these two genes being located within the cytoplasmic membrane can be found below in the TMHMM and SignalP outputs. Having any transmembrane helices indicates that the protein is embedded within a membrane, and because this is a transport protein, it is most likely going to be within the cytoplasmic membrane.

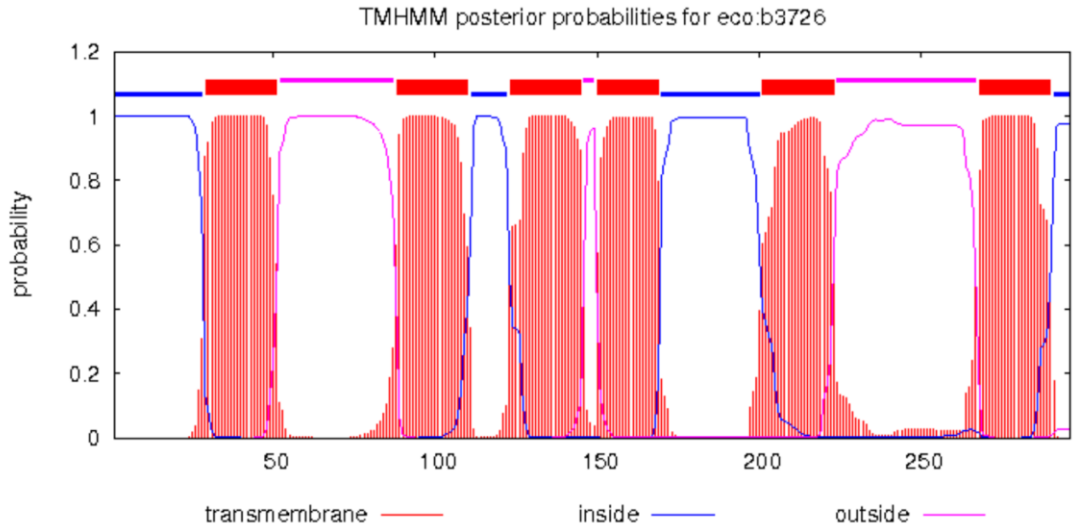


Figure 20: TMHMM output for b3726. This shows that the protein has six transmembrane helices, which indicates its location within the cytoplasmic membrane.

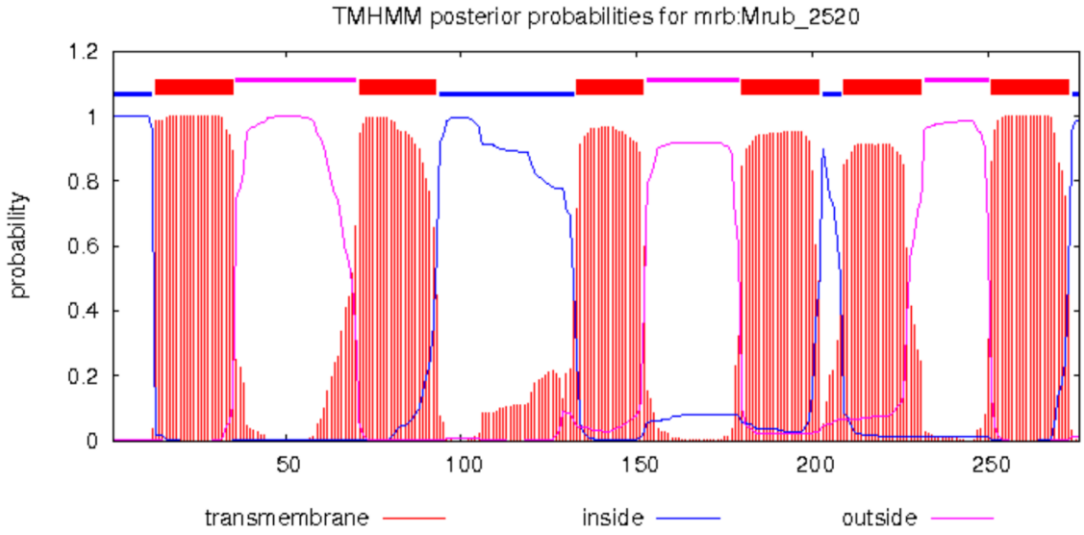


Figure 21: TMHMM output for Mrub_2520. This shows that the protein has six transmembrane helices, which indicates its location within the cytoplasmic membrane.

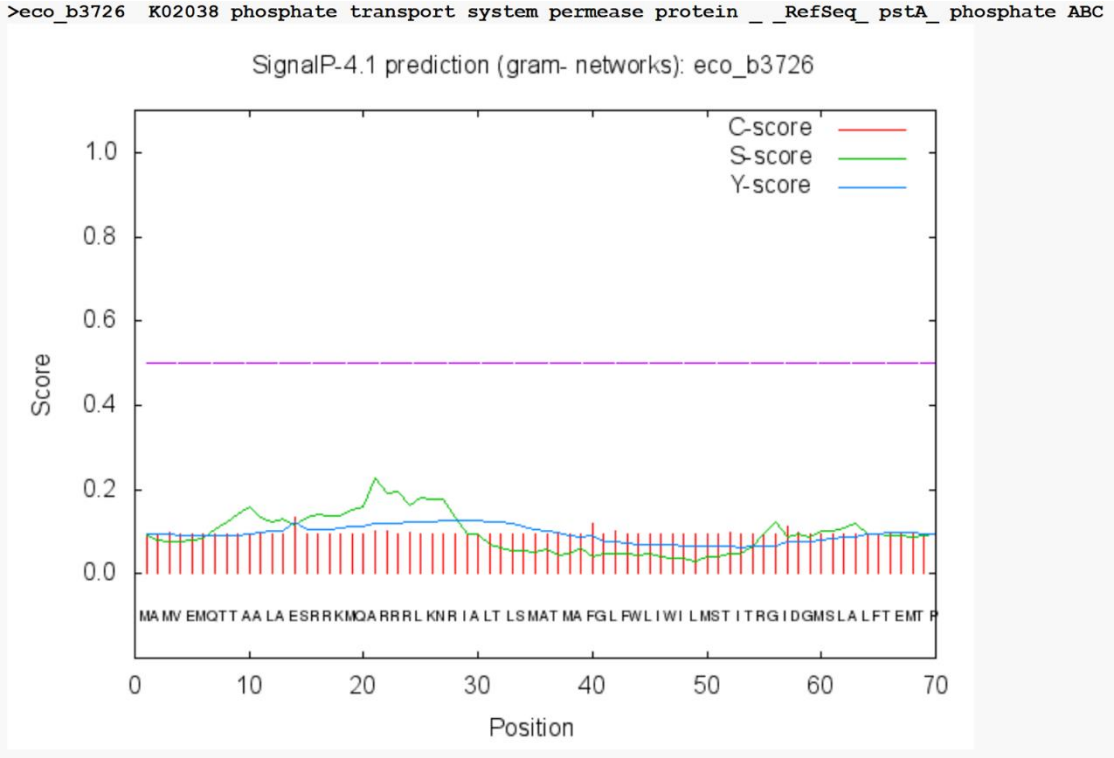


Figure 22: The SignalP output for b3726. Shows that there are no signal proteins or cleavages, indicating that the protein is not within the periplasm or outside the outer membrane.

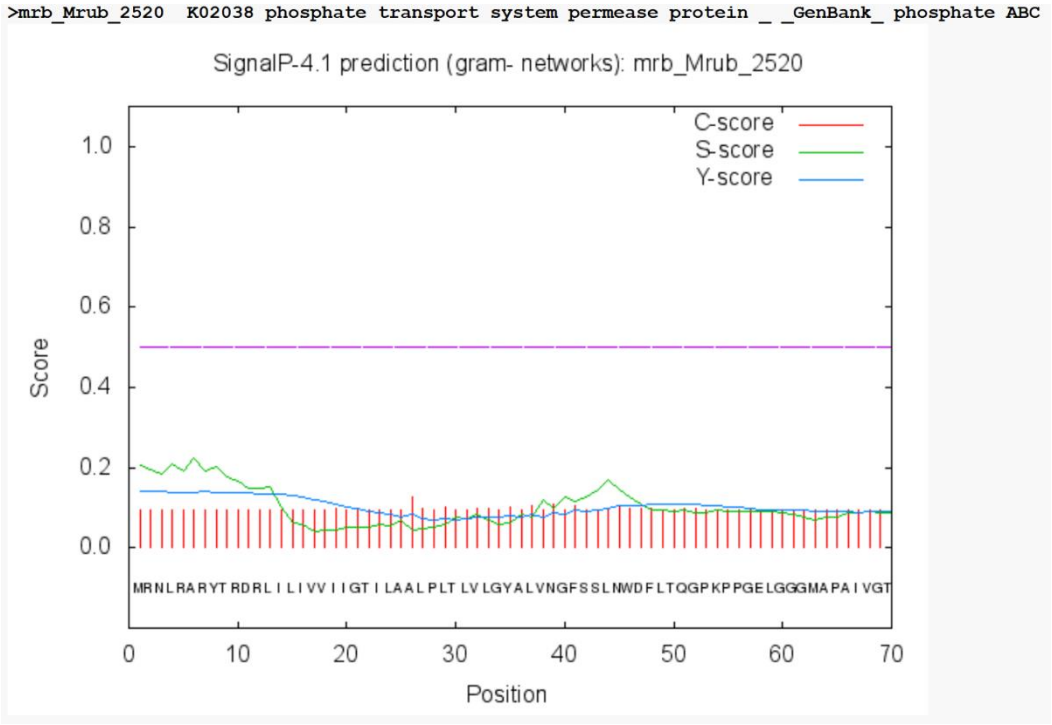


Figure 23: The SignalP output for Mrub_2520. Shows that there are no signal proteins or cleavages, indicating that the protein is not within the periplasm or outside the outer membrane.

Orthologs b3725 and Mrub_2521 are both responsible for PstB, the ATP biding site within the cytoplasm of the cell. This is evident from the multiple sources below, which site these genes as being within the ABC transporter system as the PstB protein.

Table 4: Data From b3725 and Mrub_2521

Bioinformatics Tool Used	*E. coli* b3725 gene	Mrub_2521 gene
BLAST E-value	7e-100	---
Blast Score	290 bits	---
CDD Data (COG category)	Phosphate ABC Transport PstB (COG1117)	Phosphate ABC Transport PstB (COG1117)
Cellular Localization (PSORT-B)	Cytoplasm	Cytoplasm
TIGRfam-protein family	Phostphate ABC Transporter (TIGR00972)	Phostphate ABC Transporter (TIGR00972)
Pfam-protein family	ABC Transporter (Pf00005.26)	ABC Transporter (Pf00005.26)
Protein Database	1L7V	IV43
KEGG pathway map	Phosphate ABC Transporter Pst (02010)	Phosphate ABC Transporter Pst (02010)

Further evidence for these genes being the ATP binding sites within the cytoplasm can be seem from the TMHMM and SignalP outputs. The TMHMM outputs show no transmembrane helices, and the SignalP shows no cleavage sites or signal peptides, leaving only the cytoplasm left as a potential location. The SignalP output also labels each protein as an ATP binding protein.

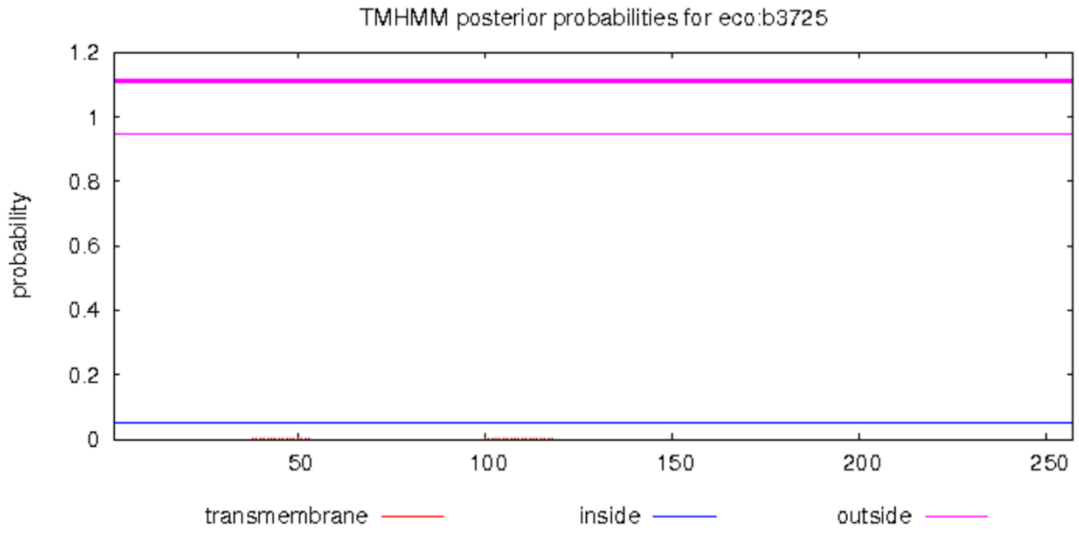


Figure 24: TMHMM output for b3725. This shows that there are no transmembrane helices within the protein.

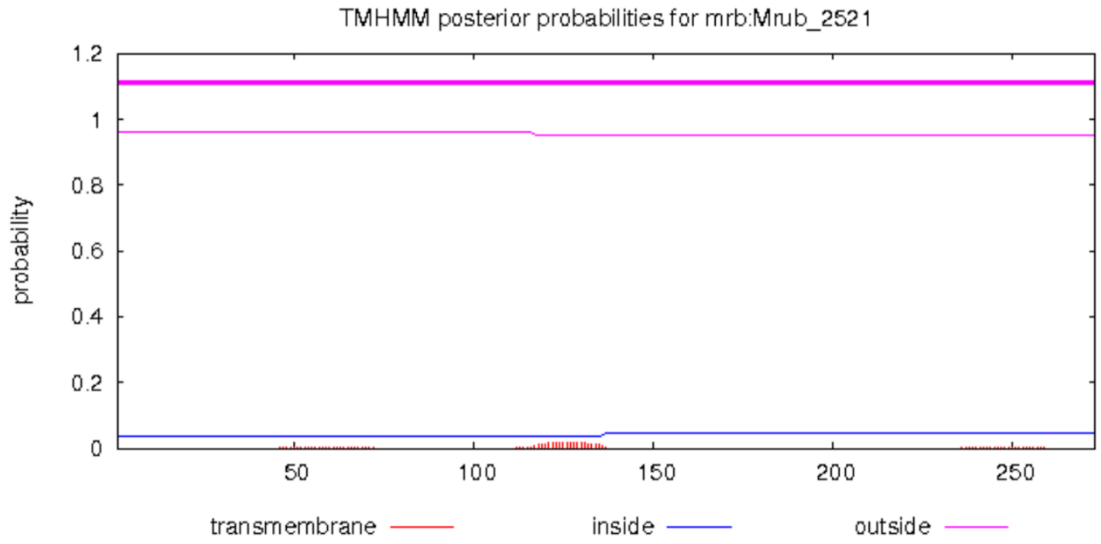


Figure 25: TMHMM output for Mrub_2521. This shows that there are no transmembrane helices within the protein.

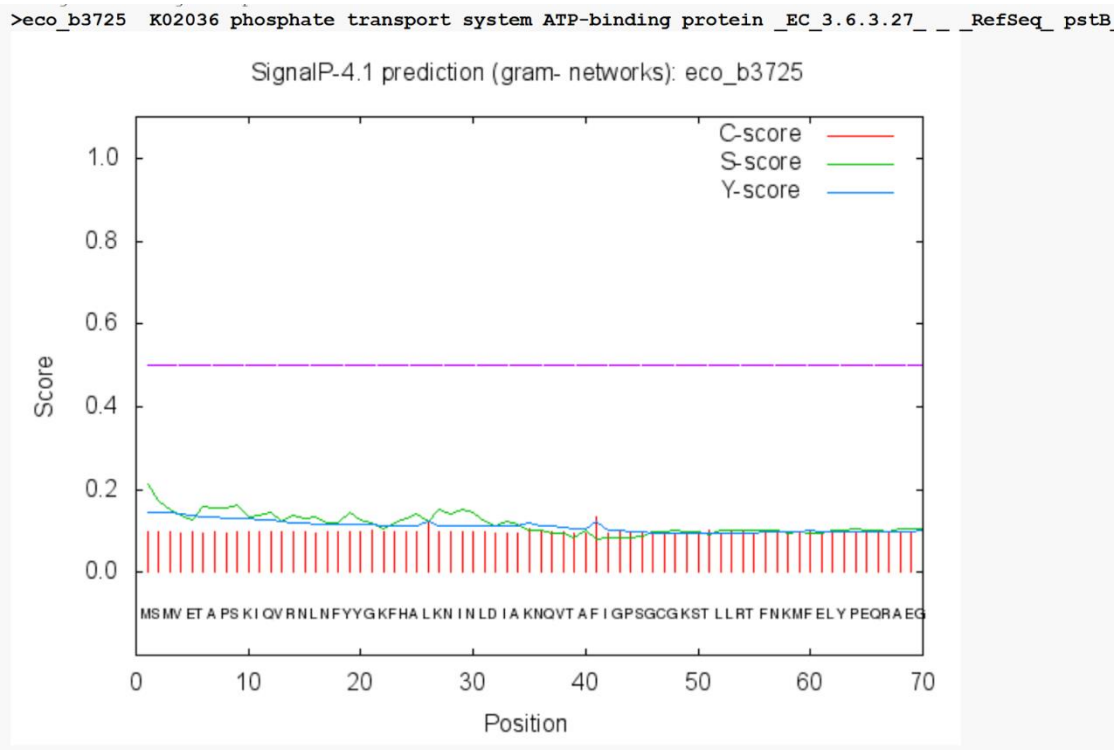


Figure 26: SignalP output for b3725. Shows that there are no cleavage sites or signal peptides.

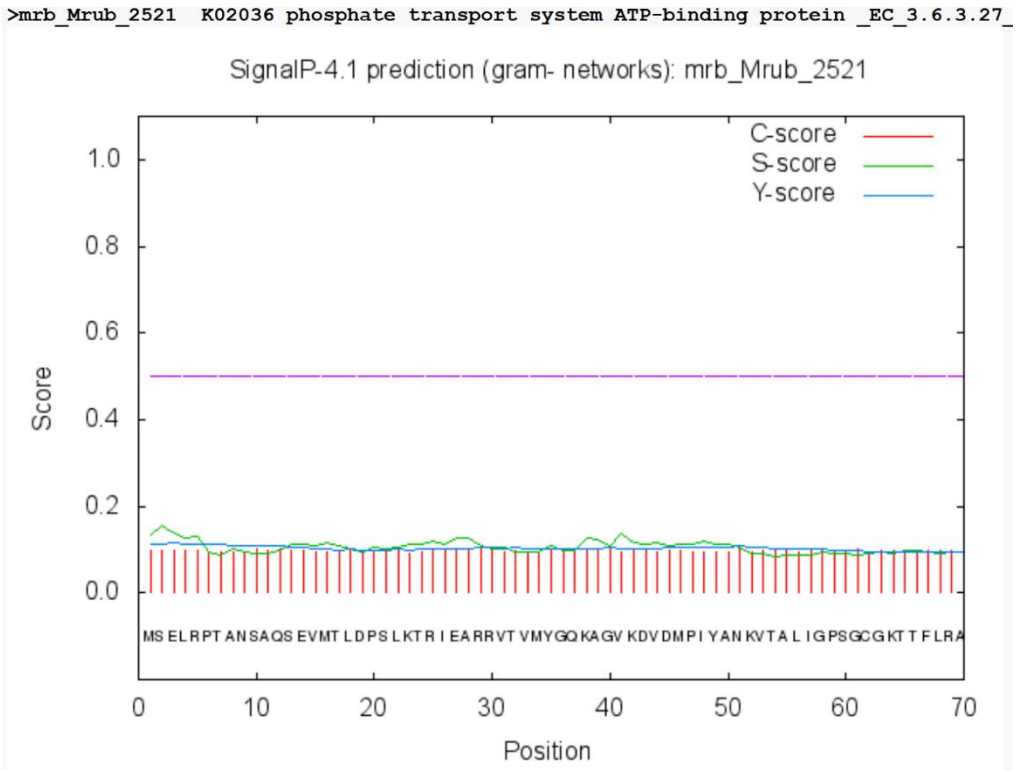


Figure 27: SignalP output for Mrub_2521. Shows that there are no cleavage sites or signal peptides.

Potential Operons

The results from IMG show that both *E. coli* and *M. ruber* have operons for the phosphate ABC transporter. The four genes in *E. Coli*, b3728, b3727, b3726 and b3725 are an operon as shown below.

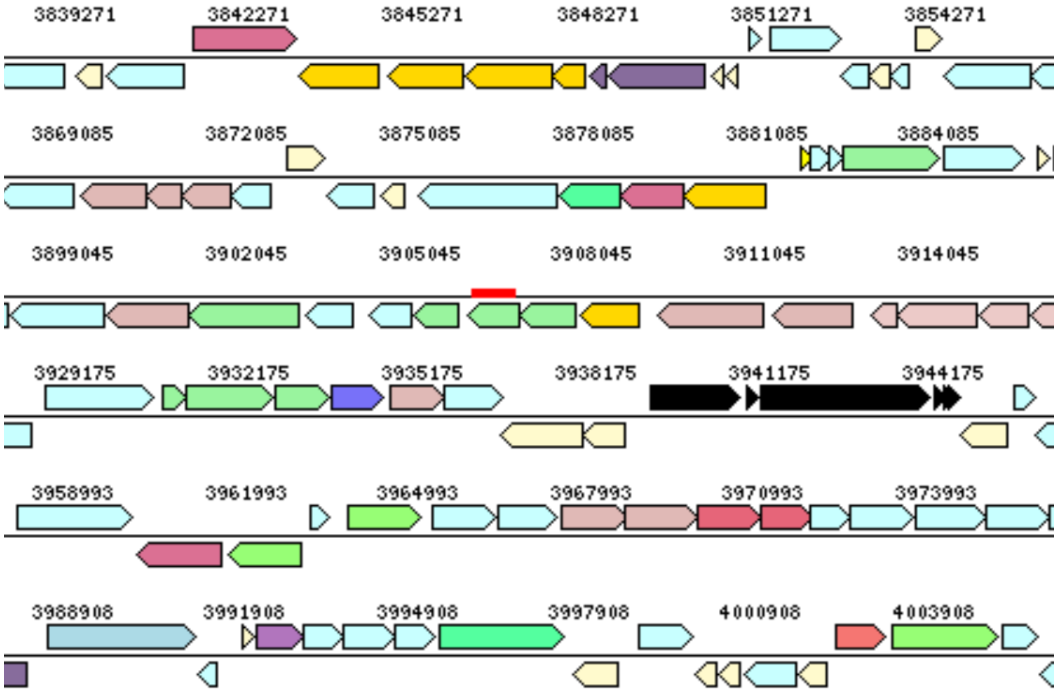


Figure 28: IMG/KEGG color output for the *E. coli pst* operon. The four genes can be seen in the middle of the figure as the three light green genes to the left of the dark yellow. One of the genes rests under a red indicator.

The four genes in *M. ruber*, Mrub_2518, Mrub_2519, Mrub_2520 and Mrub_2521 are also an operon as shown below.

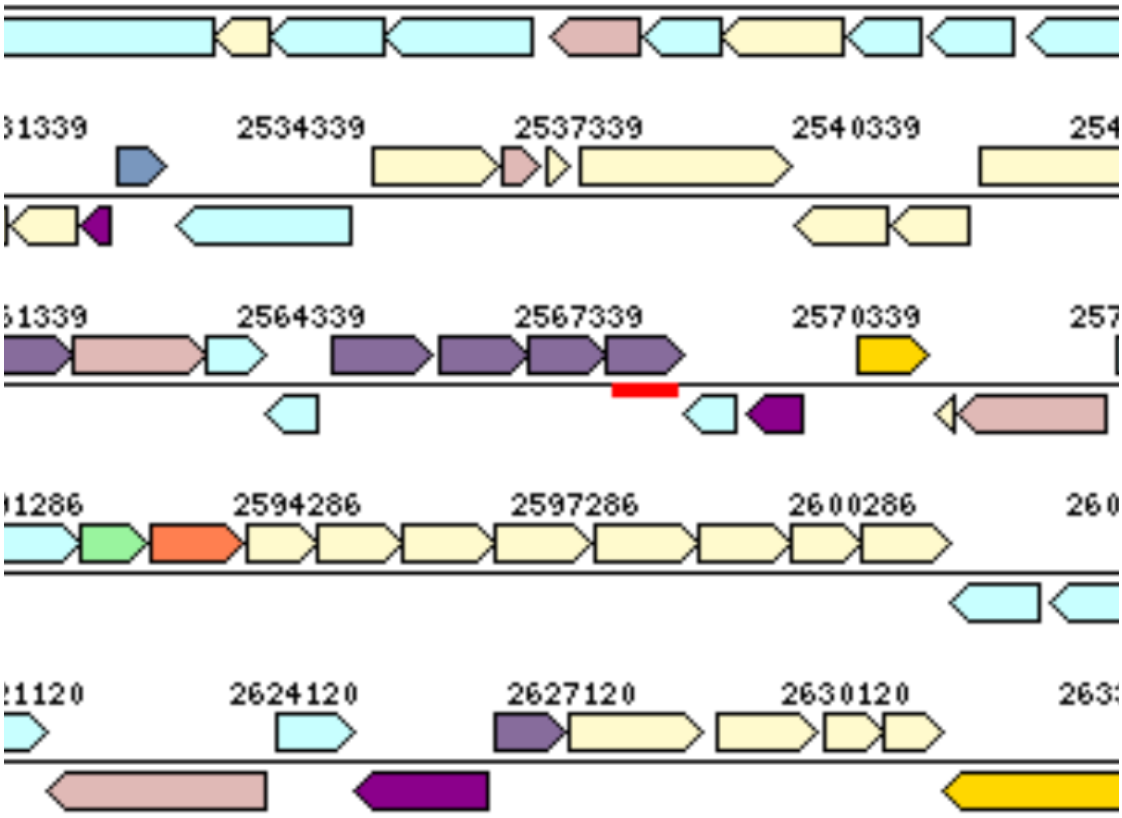


Figure 29: IMG/KEGG color output for the *M. ruber pst* operon. The four genes can be seen in the middle of the figure as the purple genes. One of the genes rests over a red indicator.

Conclusion

There were three main goals of this project. The first was to find out if our chosen genes were orthologs of one another. We confirmed that there were four sets of orthologs, b3728 and Mrub_2518, b3727 and Mrub_2519, b3726 and Mrub_2520, and b3725 and Mrub_2521. This was confirmed using the KEGG database's pathways, the Blast output and of course the locus tags. The Blast output is the most significant piece of data here, as each gene was used to search for potential matches. The *E. coli* genes, when blasted against the *M. ruber* genome, came up with exact matches to our expected genes, with low enough e-values to be significant (4e-71, 1e-54, 1e-48, and 7e-100 respective to the previously stated orthologous pairs). We can confirm that the four sets of genes are orthologs.

Our second goal of the project was to confirm the location and function of the protein. While that information was largely already available for the *E. coli* genes, the *M. ruber* genes needed to be researched. The findings that came back showed that the genes coded for the exact proteins expected of them. Mrub_2518 was found by PSORT-B to be located in the periplasm. TIGRfam and Pfam placed the gene under the family of binding proteins, and SignalP labeled it as a signal peptide 1. All of these outputs create a strong argument for the gene being the phosphate binding protein within the periplasm, or the PstS protein. This conclusion is strengthened by the fact that Mrub_2518's ortholog within *E. coli* had almost the exact same output. Mrub_2519 was found by PSORT-B to be located within the cytoplasmic membrane. TMHMM found that it had six transmembrane helices, and Pfam and CDD labeled it under the phosphate ABC transport family. These are all very good indications that the gene is, as expected, one of the two transmembrane proteins found within the phosphate system (PstC). This argument is again strengthened by the fact that this gene's *E. coli* ortholog had an almost identical output. Mrub_2520 was very similar to Mrub_2519, in that it was also found to be one of the two transmembrane proteins, with six helixes. Mrub_2520 was found to be PstA. Lastly Mrub_2521 was found to be the cytoplasmic ATP binding protein. This was supported by PSORT-B stating it to be within the cytoplasm, the SignalP outputs showing that it was not a signal peptide, and the TMHMM output showing that it was not a transmembrane protein. This gene was found to create the protein PstB.

Our last goal for this project was to confirm that the genes studied were in an operon. Using IMG we were able to visually see that the genes were in order. This supports the idea that genes with sequential locus tags of the same function are in an operon.

Overall, we can conclude that the genes Mrub_2518, Mrub_2519, Mrub_2520 and Mrub_2521 are in an operon *pst*, and that they code for the phosphate ATP transport system. We also know that the genes b3728, b3727, b3726 and b3725 are in an operon for *pst* in *E. coli*. Lastly, we can support the idea that b3728 and Mrub_2518, b3727 and Mrub_2519, b3726 and Mrub_2520, and b3725 and Mrub_2521 are all orthologs.

Literature Cited

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. [Internet]. 2000. [2018 Jan]. Available from: http://www.rcsb.org/.

Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: A sequence logo generator. [Internet]. 2004. 14:1188-1190. Genome Research; [2018 Jan]. Available from: http://weblogo.berkeley.edu/

Finn D, Coggill P, Eberhardt Y, Eddy R, Mistry J, Mitchell L, Potter C, Punta M, Qureshi M, Sangrador-Vegas A, Salazar A, Tate J, Bateman A. The Pfam protein families database: towards a more sustainable future. [Internet]. 2016. 44:D279-D285. [Jan 2018]. Available from: http://pfam.xfam.org/

Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O. TIGRFAMs: a protein family resource for the functional identification of proteins. [Internet]. 2001.Nucleic Acids Res 29(1):41-3. [Jan 2018].

Juncker A, Willenbrock H, G. von Heijne, Nielsen H, Brunak S, Krogh A. Prediction of lipoprotein signal peptides in Gram-negative bacteria. [Internet]. 2003. 12(8):1652-62. Protein Sci.; [2018 Jan]. Available from: http://www.cbs.dtu.dk/services/LipoP/

Kall L, Krough A, Sonnhammer E. Phobius: A combined transmembrane topology and signal peptide prediction method. [Internet]. 2004. 338(5):1027-36. Journal of Molecular Biology; [Jan 2018].

Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. [Internet]. 2016. 44, D457–D462. [2018 Jan]. Available from: http://www.genome.jp/kegg/

Keseler M, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martinez C, Fulcher C, Huerta 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 P, Paulsen I, Karp D. EcoCyc: fusing model organism databases with systems biology. [Internet]. 2013. 41:D605-612. Nucleic Acids Research; [Jan 2018].

Krogh A, Rapacki K. TMHMM Server, v. 2.0. Cbs.dtu.dk. [Internet]. 2016. [2018 Jan]. Available from: http://www.cbs.dtu.dk/services/TMHMM/

Madden T. The BLAST Sequence Analysis Tool. [Internet]. 2002. The NCBI Handbook; [Jan 2018]. 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. [Internet]. 28(43): D222-2. [2018 Jan]. 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. IMG: The integrated microbial genomes database and comparative analysis system. [Internet]. 2012. Nucleic Acids Research 40(D1):D115-22. [2018 Jan]. Available from: http://nar.oxfordjournals.org/content/40/D1/D115.full

Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. [Internet]. 2000. Journal of molecular biology 302 (1):205-17. [2018 Jan]. Available from: http://www.ebi.ac.uk/Tools/msa/tcoffee/

Petersen T, Brunak S, Gunnar von Heijne, Nielsen H. Discriminating signal peptides from transmembrane regions. [Internet]. 2011. 8:785-786. Nature Methods; [Jan2018].Available from: http://www.cbs.dtu.dk/services/SignalP

Rao N, Torhani A. Molecular aspects of phosphate transport in Escherichia coli. [Internet]. 1990. 4(7), 1083-1090. Molecular Microbiology; [Jan 2018]. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2958.1990.tb00682.x/epdf

Rees D, Johnson E, Lewinson O. ABC transporters: The power to change. [Internet]. 2009. 10(3): 218–227. Nat Rev Mol Cell Biol; [Jan 2018]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2830722/

Tindall et al. Complete genome sequence of Meiothermus ruber type strain (21T). [Journal]. 2010. 3:26-36. Standards in Genomic Sciences; [Jan 2018].

Tomii K, Kanehisa M. A Comparative Analysis of ABC Transporters in Complete Microbial Genomes. [Internet]. 1998. 8.10.1048. Genome Research; [Jan 2018]. Available from: http://genome.cshlp.org/content/8/10/1048.long

Wilkens S. Structure and Mechanism of ABC Transporters. [Internet]. 2015. *F1000Prime Reports* 7:14. *PMC*; [Jan 2018]. Avalible from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4338842/

Yu N, Wagner J, Laird M, Melli G, Rey S, Lo R, Dao P, Sahinalp S, Ester M, Foster L, Brinkman F. PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. [Internet]. 2010. 26(13):1608-1615. Bioinformatics; [Jan 2018].