2018

Mrub_1283, Mrub_1284 and Mrub_1285 encode for a glycine/betaine ABC transporter and are orthologs of *E. coli* proV, proW and proX

Lan Dang
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Mrub_1283, Mrub_1284 and Mrub_1285 encode for a glycine/betaine ABC transporter and are orthologs of *E. coli proV*, *proW* and *proX*

Lan Dang

Dr. Lori R. Scott Laboratory

Augustana College

639 38th Street, Rock Island, IL 61201

INTRODUCTION

ABC transporters are the proteins of interests

Cellular transporters are essential for all forms of life because they facilitate the transport of all molecules across the cellular membrane to help maintain the homeostasis of the cells. For example, in the bacteria *Escherichia coli*, about 10% of its entire genome encoding for proteins is involved in transport processes (Blattner *et al*, 1997). There are two main forms of cellular transport: passive diffusion and active transport. While passive diffusion follows the chemical gradient of molecules and does not require energy, active transport against chemical gradient requires a source of energy either by using high-energy molecule like ATP directly or the potential energy provide by a coupled reaction (Paula *et al*, 1996). Transporters using molecules like ATP are called primary transporter while transporters depending on coupled reactions are secondary transporters (Saier *et al*, 1980). In this paper, we are interested in ABC transporter - a type of primary transporters that use ATP as the energy source to drive chemicals against their gradient. ABC transporters can facilitate a wide range of substrates, from small inorganic compounds to larger organic molecules such as glucoses, amino acids, nucleosides, vitamins and metal clusters to larger organic compounds, including peptides, lipid molecules, oligonucleotides

and polysaccharides (Wilken *et al*, 2015). The catalytic/transport mechanism of ABC-transporters is of interest to all biologists in general and to bioinformatics scientist in particular. There are various types of data that can be used to understand the mechanism of ABC transporters such as structural data from crystallography, experimental data from biochemical studies and informatics data from bioinformatics algorithms. The paper targets into ABC glycine/ betaine transporters, encoded by an operon contained Mrub_1283, Mrub_1284 and Mrub_1285 in *Meiothermus ruber* genome.

Generally, ABC transporters are multi-subunit transporters that all contained essential cytoplasmic factors, which are essential to ATP hydrolysis activity (Higgins *et al*, 1992). ABC transporters are structurally characterized by two nucleotide-binding domains (NBDs) and two transmembrane domains (TMDs) (Holland *et al*, 1999). It is also common to find a phosphate-binding loop (P-loop) motif in an ABC transporter (Higgins *et al*, 1986). Since we expect Mrub_1283, Mrub_1284, and Mrub_1285 to encode the ABC transporter, we want to verify whether their protein products are similar in structure to the components of a general ABC transporter. As predicted by different bioinformatics tools and protein family data, Mrub_1283 is expected to encode for the P-loop structure, Mrub_1284 encodes for the transmembrane domains, and Mrub_1285 encodes for the substrate bind domain (NBDs) so we hypothesize that Mrub_1283, Mrub_1284, and Mrub_1285 are part of an operon that encode for the full structure of an ABC transporter that transport glycine/ betaine. Figure 1 below shows the 3D structure of some common ABC transporter with typical subunits. Since we expect that Mrub_1283, Mrub_1284, and Mrub_1285 are part of an operon that encodes for the full structure of an ABC transporter that transport glycine/ betaine, we expect to find structures that similar to figure 1 from PDB to match with protein products of these genes of interests.

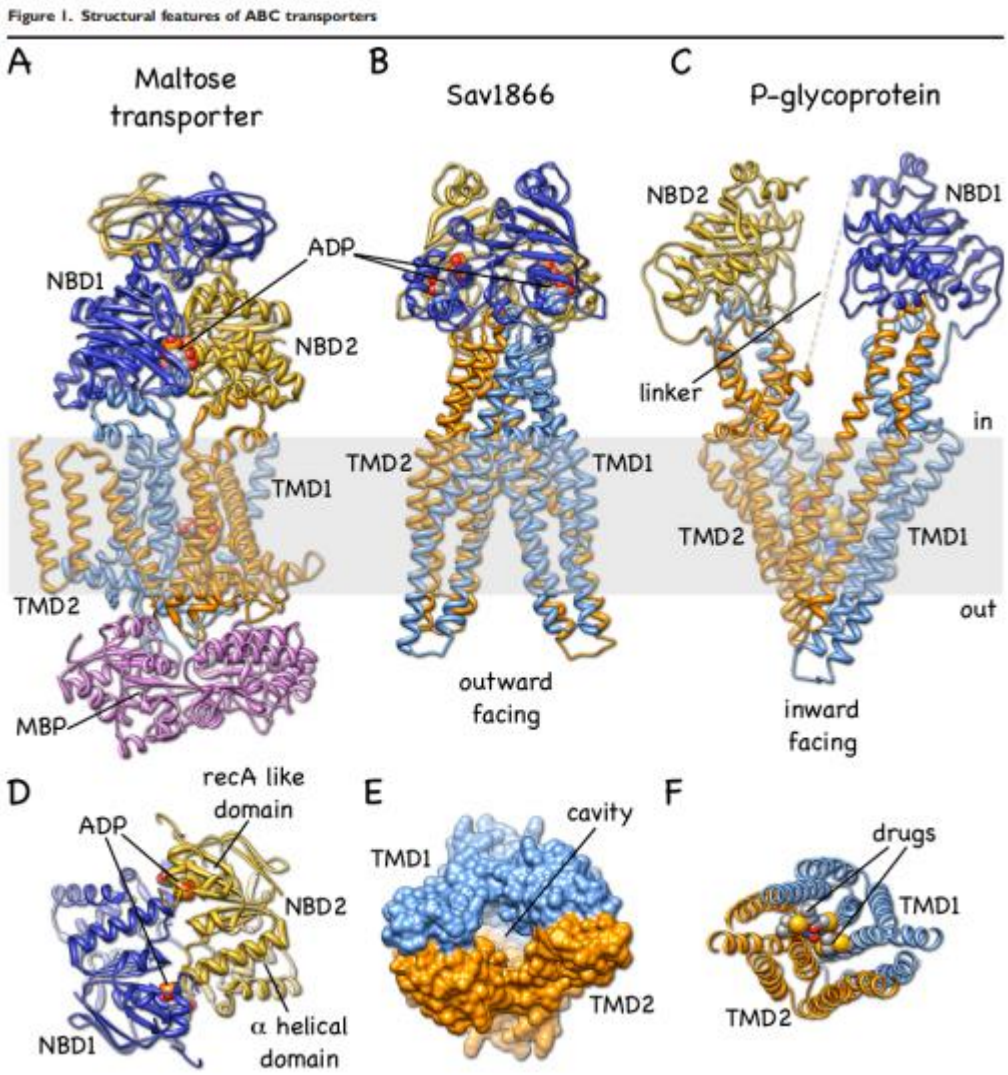


Figure 1. Figure 1 shows the structure of some common ABC transporters. (Wilken *et al*). (A) shows the outward-facing maltose transporter with ADP•VO4 in catalytic sites and maltose bound to the transmembrane domain (Oldham *et al,* 2011) (B) is the homodimeric exporter Sav1866 from *Staphylococcus aureus* in the outward-facing conformation with ADP in catalytic sites (Lewinson *et al*, 2010). (C) shows the P-glycoprotein in the inward-facing conformation with an inhibitor molecule bound at the TMDs (Li *et al,* 2014). (D) is the nucleotide-binding domain (NBD) sandwich dimer of the maltose transporter (MalK) as seen from the cytoplasmic side. (E) presents the cavity

formed by the TMDs of outward-facing Sav1866. In this structure, the cavity does not provide access to the outer leaflet of the lipid bilayer. Last but not least, (F) shows the cross-section through the TMDs of glycoprotein showing the two inhibitor molecules.

Typically, ABC transporters pump transport substrates against their chemical gradient in a single direction (either import or export).(Balakrishnan *et al*, 2004). In order to do that, the membrane domain must play the role of one or more "turnstile-like" gates and couple tightly to the catalytic cycles on the NBDs. To serve as the gates, the transmembrane domains of ABC transporters must alternate between outward and inward facing conformation (Jardetzky *et al*, 1966). We expect to see this aspect in the protein product of Mrub_1284 since it encodes the TMDs of the transporters. We also expect the protein encoded by Mrub_1285 to be a signal protein while the protein by Mrub_1283 to form a loop structure. To verify our hypothesis, a structural data module is performed and will be presented in this paper as well.

Mrub_1283, Mrub_1284 and Mrub_1285 are parts of an operon

We expect Mrub_1283, Mrub_1284, and Mrub_1285 to encode for different subunits of the glycine/betaine transporter; we expect these genes to be in the same operon. These genes are also located in close proximity and serve the same purpose, they are very likely to be in an operon. To confirm this hypothesis, several bioinformatics tools are employed and will be described later in the paper.

***Meiothermus ruber* is the study model**

Meiothermus ruber (Loginova *et al*. 1984, Nobre *et al*. 1996) is a species of the genus *Meiothermu*s. This genus has its name to indicate its presence in a hot environment (Nobre *et al*, 1996; Euzeby *et al*, 1997). The species name "ruber" was to indicate the red cell pigmentation of

the organism. (Loginova *et al*. 1984 Euzeby *et al*, 1997). Members of this genus are isolated from natural hot springs and artificial thermal environments in different Europe and Asia countries (Nobre *et al*. 1996). This genus is heterogeneous with respect to pigmentation. On the basis of the 16S rRNA gene sequence similarity, yellow species form a distinct group while the red/orange pigmented strains forming another group (Pires *et al*, 2005; Zhang *et al*, 2010). The relatively low degree of 16S rRNA gene sequence similarity makes the *Meiothermus* genus to form a separate evolutionary lineage from members of the genus *Thermus.* Among all species of *Meiothermus*, *M. ruber* can be used as a representative since it genome sequence was the first to be completely sequenced (Tindall *et al*, 2010). The reason we choose to study the function of the genome of a less common strain like *M. ruber* is that we expect to see a less common aspect that we normally cannot see in common *Eschirichia* or *Salmonella* species. The difference in optimal temperature of *Meiothermus* may also be results of some interesting metabolic pathway and protein function that we cannot expect to find out in common species.

Mrub_1283, Mrub_1284 and Mrub_1285 are orthologous to *proV, proW* and *proX* genes in *E. coli* genomes

In this study, we use *E. coli*, a very common organism as the control to study *M. ruber* genes. *E. coli*, are relatively easy to grow in the laboratory, which has allowed them to be extensively studied (Cooper 2000). A BLAST of *proV*, *proW* and *proX* genes in *E. coli* against *M. ruber* genome showed that there are orthologs of these genes in *M. ruber*. Therefore, the study of *E. coli* can help us understand better about our *M. ruber* genes of interests under the evolutionary aspects.

Bioinformatics approach

There are several different approaches to understand the function of Mrub_1283, Mrub_1284 and Mrub_1285 in *M. ruber* genome such as structural biology approach or experimental approach. Among these approaches, bioinformatics approach is a promising one since it can be efficient in time and money manner. Several bioinformatics tools are available for free and become very helpful. In this studies, the bioinformatics tools which are used include KEGG (Kanehisa M *et al*, 2016), BLAST (Madden T *et al*, 2002), EcoCyc (Keseler *et al*, 2013), T-Coffee (Notredame *et al*, 2000), WebLogo (Crooks *et al*, 2004), TMHMM (Krogh *et al*, 2016), SignalP (Petersen *et al*, 2004), LipoP (Junker *et al*, 2003), PSORT-B(Yu *et al*,2010) , Phobius (Kall *et al*, 2007), CDD search (Marchler-Bauer *et al*, 2014), IMG/M (Markowitz *et al*, 2012), TIGRFAM (Haft *et al*, 2001) , PFAM (Finn *et al*, 2016), PDB (Berman *et al*, 2000).

Purpose/ Hypothesis

Our hypothesis is that the function of Mrub_1283, Mrub_1284 and Mrub_1285 in *M. ruber* genome as encoding for subunits of glycine/betaine ABC transporters. To confirm that, we will need to confirm that Mrub_1283, Mrub_1284, and Mrub_1285 are orthologous to *proV*, *proW* and *proX* in *E. coli* genomes so that Mrub_1283, Mrub_1284 and Mrub_1285 will have the same function as *proV*, *proW* and *proX* to encode for glycine/betaine ABC transporters. We derive our hypothesis based on the low E-value of the initial BLAST of *E. coli* genes against *M. ruber* genomes.

MATERIALS AND METHODS

To collect bioinformatics data for *M. ruber* and *E. coli* genes, the GENI-ACT gene annotation was followed with some deviations. The studied was started by a BLAST of *E. coli* gene against *M. ruber* genome to looking for any potential orthologs. Once we had the similar pair of sequences, we followed up by filling out different modules on the GENI-ACT site by

using the proper bioinformatics tools. One deviation we made was that instead of using the recommended top 10 BLAST hits for the T-coffee analysis, we instead picked a list of 15-20 hits from various genus other than *Eschirichia.* We also omitted the MetaCyc path but use only KEGG ((Kanehisa M *et al*, 2016) database to obtain biological pathway. We also omitted the Open Reading Frame module for all *E. coli* genes since all *E. coli* genes were studied so well that we were certain about their reading frame. Since the protein products of our GOIs were units of transporter which were not a necessary enzyme, the enzyme function Module was omitted as well.

Follow the GENI-ACT instruction, all the bioinformatics tools which are used include KEGG (Kanehisa M *et al),* BLAST (Madden T *et al*, 2002), EcoCyc(Keseler *et al*, 2013), T-Coffee (Notredame *et al*, 2000), WebLogo (Crooks *et al*, 2004), TMHMM (Krogh *et al*, 2016), SignalP (Petersen *et al*, 2004), LipoP (Junker *et al*, 2003), PSORT-B(Yu *et al*,2010) , Phobius (Kall *et al*, 2007), CDD search (Marchler-Bauer *et al*, 2014), IMG/M (Markowitz *et al*, 2012), TIGRFAM (Haft *et al*, 2001) , PFAM (Finn *et al*, 2016), PDB (Berman *et al*, 2000). The data obtained from these bioinformatics tools are selected to determine the function of *M. ruber* GOIs.

RESULTS

Section I. KEGG (Kanehisa M *et al,* 2016) and BLASTp (Madden T *et al,* 2002) results

In this section, KEGG data and BLASTp results for each pair of genes are presented in a table of KEGG data and a figure of BLASTp result. Table 1 shows the KEGG data for Mrub_1283 and b2677. From the data, we see these two genes have the same gene name of *proV*. Their sequence lengths are also very close to each other.

Table 1. KEGG data for Mrub_1283 and *E. coli* b2677. From the data, we can see the similarity between these two genes in term of gene name and sequence length.

	b2677	Mrub_1283
Gene Name	*proV*	
KEGG map	map02010 – ABC transporters	
DNA coordinates	2804815..2806017	1307798..1308991
DNA Sequence	atggcaattaaattagaaattaaaaatctttataaaa tatttggcgagcatccacagcga gcgttcaaatatatcgaacaaggactttcaaaaga acaaattctggaaaaaactgggcta tcgcttggcgtaaaagacgccagtctggccattga agaaggcgagatatttgtcatcatg ggattatccggctcgggtaaatccacaatggtacg ccttctcaatcgcctgattgaaccc acccgcgggcaagtgctgattgatggtgtggatat tgccaaaatatccgacgccgaactc cgtgaggtgcgcagaaaaaagattgcgatggtct tccagtcctttgccttaatgccgcat atgaccgtgctggacaatactgcgttcggtatgga attggccggaattaatgccgaagaa cgccgggaaaaagcccttgatgcactgcgtcag gtcgggctggaaaattatgcccacagctacccgg atgaactctctggcgggatgcgtcaacgtgtggg attagcccgcgcgttagcgattaatccggatatatt attaatggacgaagccttctcggcgctcgatccatt aattcgcaccgagatgcaggatgagctggtaaaa ttacaggcgaaacatcagcgcaccattgtctttattt cccacgatcttgatgaagccatgcgtattggcgac cgaattgccattatgcaaaat ggtgaagtggtacaggtcggcacaccggatgaa attctcaataatccggcgaatgattat gtccgtaccttcttccgtggcgttgatattagtcagg tattcagtgcgaaagatattgcc	atgagttttatacgtgtagaaaacctatacaagatcttcg gcccaaaggccggacaagcc ctggaaatggtgcagggggcaccgataaagacacg ctttttcaaaagacccgccacgtgctgggcctgaacag gatcaacctggaggtgaagcagggcgaatttttgtgat catggggctttcggggtcgggcaagtccaccctgcttc gggtgctcaaccgcctgatcgagcccacagcaggtcg ggttttggtcggtgataccgaggtaaccaccctcccgc acaaagagcttctggtttccgccaggacaccttcggtat ggttttccagcactttgctttgcttcctcactacaacattct gcgcaacgtggctttcccgctggagctcaaagggcttt cccgtaaggagcgggaggagcagggcatggcctggt tagagcgggtggggctttccggctatgagaagcattac ccagggcagttgtctggtggacagaaacagcgggttg gcctggcgcgggccctttgcgcaaaccctcccatcctg ctcatggacgaggccttcagcgcgctggatcccctgat ccgcaaggagatgcaggacgaacttttgcgtctgcag caagagttaaaaaagaccatcgtctttgtaacccacgac ctggatgaggccatgcgcctgggagaccgtatcgcca tcatgcgggacggggaggtggtgcaggtaggaaccg cggaggagattctggcccgccctgcagacgattatgtg gccgcctttttgtccggtgttaatcccgccaaaatctaca aggtggaggagctggtgcaggaacccgtgaccgtgg tgctggaacgggagggcctgcgctcagccctgcgca agatgggccaggccggtgctgtgaatgcctatgtggta aatcgtagcggatttttcaggggatggtgcgagctgaa aagttggccgaagcgcttaaggccgaaggggagcgt ggtgggctggagagcctcctggaacccctacccgcgc

	cgccggacaccgaatggcttaattcgtaaaaccc ctggcttcggcccacgttcggcactgaaattattgc aggatgaagatcgcgaatatggctacgttatcgaa cgcggtaataagttt gtcggcgcagtctccatcgattcgcttaaaaccgc gttaacgcagcagcaaggtcttgat gcggcgctgattgatgcgccgttagcagtcgatg cacaaacgcctcttagcgagttgctc tctcatgtcggacaggcaccctgtgcggtgcccgt ggtcgacgaggaccaacagtatgtcggcatcattt cgaaaggaatgctgctgcgcgctttagatcgtga gggggtaaataatggctga	tttcgcccggtcagaccctggaagaggccctgccgct gttcagtgaaaccgcgctgcccttgcccatactggacg agaaagggcggctcctaggggtggtgacgcggggcc ggttgatcgcggccatggcggggcgttacgtgcctca atag
Sequence Length	1203 nt	1194 nt
Protein Sequence	MAIKLEIKNLYKIFGEHPQRAFK YIEQGLSKEQILEKTGLSLGVKD ASLAIEEGEIFVIMGLSGSGKST MVRLLNRLIEPTRGQVLIDGVDI AKISDAELREVRRKKIAMVFQSF ALMPHMTVLDNTAFGMELAGI NAEERREKALDALRQVGLENY AHSYPDELSGGMRQRVGLARA LAINPDILLMDEAFSALDPLIRTE MQDELVKLQAKHQRTIVFISHD LDEAMRIGDRIAIMQNGEVVQV GTPDEILNNPANDYVRTFFRGV DISQVFSAKDIARRTPNGLIRKTP GFGPRSALKLLQDEDREYGYVI ERGNKFVGAVSIDSLKTALTQQ QGLDAALIDAPLAVDAQTPLSE LLSHVGQAPCAVPVVDEDQQY VGIISKGMLLRALDREGVNNG	MSFIRVENLYKIFGPKAGQALEMV QGGTDKDTLFQKTRHVLGLNRINL EVKQGEFFVIMGLSGSGKSTLLRVL NRLIEPTAGRVLVGDTEVTTLPHKE LLRFRQDTFGMVFQHFALLPHY NILRNVAFPLELKGLSRKEREEQGM AWLERVGLSGYEKHYPGQLSGGQ KQRVGLARALCANPPILLMDEAFS ALDPLIRKEMQDELLRLQQELKKTI VFVTHLDEAMRLGDRIAIMRDGEV VQVGTAEEILARPADDYVAAFLSG VNPAKIYKVEELVQEPVTVVLERE GLRSALRKMGQAGAVNAYVVNRS GFFQGMVRAEKLAEALKAEGERG GLESLLEPLPALSPGQTLEEALPL FSETALPLPILDEKGRLLGVVTRGR LIAAMAGRYVPQ
Protein Sequence Length	400 aa	397 aa

BLASTp result for b2677 and Mrub_1283 is showed in figure 2. This is the initial BLAST we perform before doing all other informatics tool to establish out hypothesis. The first hit with lowest E-value of 3e-116 was Mrub_1283. The low E-value indicate the similarity due to evolutionary, not due to random chance.

glycine betaine/L-proline ABC transporter ATP-binding protein [Meiothermus ruber]
Sequence ID: WP_013013564.1 Length: 397 Number of Matches: 1
▸ See 3 more title(s)

Range 1: 4 to 390 GenPept Graphics ▼ Next Match ▲ Previous Ma

Score	Expect	Method	Identities	Positives	Gaps
342 bits(877)	3e-116	Compositional matrix adjust.	178/391(46%)	257/391(65%)	7/391(1%)

```
Query  5    LEIKNLYKIFGEHPQRAFKYIEQGLSKEQILEKTGLSLGVKDASLAIEEGEIFVIMGLSG  64
            + ++NLYKIFG    +A + ++ G  K+ + +KT   LG+   +L +++GE FVIMGLSG
Sbjct  4    IRVENLYKIFGPKAGQALEMVQGGTDKDTLFQKTRHVLGLNRINLEVKQGEFFVIMGLSG  63

Query  65   SGKSTMVRLLNRLIEPTRGQVLIDGVDIAKISDAELREVRRKKIAMVFQSFALMPHMTVL  124
            SGKST++R+LNRLIEPT G+VL+   ++  +   EL   R+    MVFQ FAL+PH  +L
Sbjct  64   SGKSTLLRVLNRLIEPTAGRVLVGDTEVTTLPHKELLRFRQDTFGMVFQHFALLPHYNIL  123

Query  125  DNTAFGMELAGINAEERREKALDALRQVGLENYAHSYPDELSGGMRQRVGLARALAINPD  184
             N AF +EL G++ +ER E+ +  L +VGL  Y   YP +LSGG +QRVGLARAL  NP
Sbjct  124  RNVAFPLELKGLSRKEREEQGMAWLERVGLSGYEKHYPGQLSGGQKQRVGLARALCANPP  183

Query  185  ILLMDEAFSALDPLIRTEMQDELVKLQAKHQRTIVFISHDLDEAMRIGDRIAIMQNGEVV  244
            ILLMDEAFSALDPLIR EMQDEL++LQ + ++TIVF++HDLDEAMR+GDRIAIM++GEVV
Sbjct  184  ILLMDEAFSALDPLIRKEMQDELLRLQQELKKTIVFVTHDLDEAMRLGDRIAIMRDGEVV  243

Query  245  QVGTPDEILNNPANDYVRTFFRGVDISQVFSAKDIARRTPNGLIRKTPGFGPRSALKLLQ  304
            QVGT +EIL  PA+DYV  F  GV+ ++++  +++ +     ++ +    G RSAL+ +
Sbjct  244  QVGTAEEILARPADDYVAAFLSGVNPAKIYKVEELVQEPVTVVLERE---GLRSALRKMG  300

Query  305  DEDREYGYVIERGNKFVGAVSIDSLKTALT---QQQGLDAALIDAPLAVDAQTPLSELLS  361
                   YV+ R   F G V  + L  AL    ++ GL++ L   P     QT L E L
Sbjct  301  QAGAVNAYVVNRSGFFQGMVRAEKLAEALKAEGERGGLESLLEPLPALSPGQT-LEEALP  359

Query  362  HVGQAPCAVPVVDEDQQYVGIISKGMLLRAL  392
               +    +P++DE  + +G++++G L+ A+
Sbjct  360  LFSETALPLPILDEKGRLLGVVTRGRLIAAM  390
```

Figure 2. BLASTp result for b2677 against *M. ruber* genome was showed in figure 2. The first hit with lowest E-value of 3e-116 was Mrub_1283. The low E-value indicate the similarity due to evolutionary, not due to random chance.

Table 2 shows the KEGG data for Mrub_1284 and b2678. From the data, we see these two genes have the same gene name of *proW*. Their sequence lengths are also very close to each other.

Table 2. KEGG data for Mrub_1284 and *E. coli* b2678. From the data, we can see the similarity between these two genes in term of gene name and sequence length etc.

	b2678	Mrub_1284
Gene Name	*proW*	
KEGG map	map02010 – ABC transporters	
DNA coordinates	2806010..2807074	1308988..1309827
DNA Sequence	atggctgatcaaaataatccgtgggataccacgccagcggcggacagtgccgcgcaatccgcagacgcctggggtacaccgacgactgcaccgactgacggcggtggtgctgactggctgaccagtacgcctgcgccaaacgtcgagcattttaatattctcgatccgttccataaaacgctgatcccgctcgacagttgggtcactgaagggatcgactgggtcgttacccatttccgtcccgtcttccagggcgtgcgcgttccggttgattatatcctcaacggtttccagcaattgctgctgggtatgcccgcaccggtggcgattatcgttttcgctctcatcgcctggcagatttccggggtcggaatgggtgtggcgacgctggtttcgctgattgccatcggcgcaatcggtgcctggtcgcaggcaatggtgactctggcgctggtgttaaccgccctgctgttctgtatcgtcatcggtttgccgttggggatatggctggcgagaagtccgcgagcggcgaaaattattcgtccactgcttgatgccatgcagaccacgccagcgtttgtttatctggtgccaatcgtcatgctatttggtatcggtaacgtgccgggcgtggtggtgacgatcatctttgctctgccgccgattatccgtctgaccattctggggattaaccaggttccggcggatctgattgaagcctcgcgctcattcggtgccagcccgcgccagatgctgttcaaagttcagttaccgctggcgatgccgaccattatggcgggcgttaaccagacgctgatgctggccctttctatggtggtcatcgcctcgatgattgccgtcggcgggttgggtcagatggtacttcgcggtatcggtcgtctggatatggggcttgccaccgttggcggcgtcggg	atggatcttgcggaggcaatcaatgcctttgtgcgctggctggttcaaaactacggagagacctttgaggcgatttctcagggcctcctgagcttcctctctgttctttgaggggttgttgcgggatctttcctggttctgggtagccggcttggtgttctctggcgggctggtggttgagccgccgcctggtctttgccctgggcatggggcttggcgtgtggctgatagaggcgctgggtctgtgggacaaaggcatgcagaccctggccctggtgctagctgcggtggcggtttcggtaattatcggcctccctctgggaatcctgatggggcggagcgaccgcttccgcgggttcatgctgccaattctggacgccatgcagaccatgcccagtttcgtgtatctgattccggctctgctgctctttggtctgggaaaggttccagccctgatcgccacggtcatctatgcggttcccccccatgatccgccttaccgaccttgggctgcgcatggtgcagcgggaggttatggaggctgccgaggccttcggggccacttcgtggcagcggctgcttaaggtggagctgcctctggcctttgcccaacctcctggcagggttgaaccagaccaccatgatggccctggcgatggtggttatcgcctctatgattggggctcgaggtctcggggaggaggttcttctgggaatccagcgcctggatgtgggccggggcgcggtggcaggggtggccattgtggccctggccatcgtgctggatcgactgattcaggcagccgggcaacgggccgttaaacgttaccgggaggagcgatga

	attgtgatcctcgccattatcctc gatcgtctgacgcaggccgttgggcgcgactcacg cagtcgcggcaaccgtcgctggtac accactggccctgttggtctgctgacccgcccattca ttaagtaa	
Sequence Length	1065 nt	840 nt
Protein Sequence	MADQNNPWDTTPAADSAAQSAD AWGTPTAPTDGGGADWLTSTPAP NVEHFNILDPFHKTLIPLDSWVTE GIDWVVTHFRPVFQGVRVPVDYI LNGFQQLLLGMPAPVAIIVFALIA WQI SGVGMGVATLVSLIAIGAIGAWS QAMVTLALVLTALLFCIVIGLPLG IWLARSPRAAKIIRPLLDAMQTTP AFVYLVPIVMLFGIGNVPGVVVTI IFALPPIIRLTILGINQVPADLIEA SRSFGASPRQMLFKVQLPLAMPTI MAGVNQTLMLALSMVVIASMIA VGGLGQMVLRGIGRLDMGLATV GGVGIVILAIILDRLTQAVGRDSRS RGNRRWYTTGPVGLLTRPFIK	MDLAEAINAFVRWLVQNYGETFE AISQGLLSFLLFFEGLLRDLSWFW VAGLVFLAGWWLS RRLVFALGMGLGVWLIEALGLW DKGMQTLALVLAAVAVSVIIGLP LGILMGRSDRFRGFM LPILDAMQTMPSFVYLIPALLLFG LGKVPALIATVIYAVPPMIRLTDL GLRMVQREVMEA AEAFGATSWQRLLKVELPLALPN LLAGLNQTTMMALAMVVIASMI GARGLGEEVLLGIQR LDVGRGAVAGVAIVALAIVLDRL IQAAGQRAVKRYREER
Protein Sequence Length	354 aa	279 aa

BLASTp result for b2678 and Mrub_1284 was presented in figure 3. The first hit with lowest E-value of 1e-58 was Mrub_1284. The low E-value indicate the similarity due to evolutionary, not due to random chance.

ABC transporter permease [Meiothermus ruber]
Sequence ID: WP_013013565.1 Length: 279 Number of Matches: 1
See 3 more title(s)

binding-protein-dependent transport systems inner membrane component [Meiothermus ruber DSM 1279]
Sequence ID: ADD28046.1
binding-protein-dependent transport system inner membrane protein [Meiothermus ruber DSM 1279]
Sequence ID: AGK04516.1
binding-protein-dependent transport system inner membrane protein [Meiothermus ruber H328]
Sequence ID: GAO74992.1

Range 1: 11 to 279 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
189 bits(480)	1e-58	Compositional matrix adjust.	129/274(47%)	181/274(66%)	13/274(4%)

```
Query  72   IDWVVTHFRPVFQGVRVPVDYILNGFQQLLLGMP-APVAIIVFALIAWQIS-----GVGM  125
            + W+V ++    F+ +    +    L  F+ LL  +    VA +VF L  W +S      +GM
Sbjct  11   VRWLVQNYGETFEAISQGLLSFLLFFEGLLRDLSWFWVAGLVF-LAGWWLSRRLVFALGM  69

Query  126  GVATLVSLIAIGAIGAWSQAMVTLALVLTALLFCIVIGLPLGIWLARSPRAAKIIRPLLD  185
            G+   +    I A+G W + M TLALVL A+   ++IGLPLGI + RS R    + P+LD
Sbjct  70   GLGVWL----IEALGLWDKGMQTLALVLAAVAVSVIIGLPLGILMGRSDRFRGFMLPILD  125

Query  186  AMQTTPAFVYLVPIVMLFGIGNVPGVVVTIIFALPPIIRLTILGINQVPADLIEASRSFG  245
            AMQT P+FVYL+P ++LFG+G VP ++ T+I+A+PP+IRLT LG+  V  +++EA+ +FG
Sbjct  126  AMQTMPSFVYLIPALLLFGLGKVPALIATVIYAVPPMIRLTDLGLRMVQREVMEAAEAFG  185

Query  246  ASPRQMLFKVQLPLAMPTIMAGVNQTLMLALSMVVIASMIAVGGLGQMVLRGIGRLDMGL  305
            A+  Q L KV+LPLA+P ++AG+NQT M+AL+MVVIASMI   GLG+ VL GI RLD+G
Sbjct  186  ATSWQRLLKVELPLALPNLLAGLNQTTMMALAMVVIASMIGARGLGEEVLLGIQRLDVGR  245

Query  306  ATVGGVGIVILAIILDRLTQAVGRDS--RSRGNR  337
              V GV IV LAI+LDRL QA G+ +  R R  R
Sbjct  246  GAVAGVAIVALAIVLDRLIQAAGQRAVKRYREER  279
```

Figure 3. BLASTp result of b2678 against *M. ruber* genome was showed in figure 3. The first hit with lowest E-value of 1e-58 was Mrub_1284. The low E-value indicated the similarity due to evolutionary, not due to random chance.

Table 3 shows the KEGG data for Mrub_1285 and b2679. From the data, we see these two genes have the same gene name of *proX*. Their sequence lengths are also very close to each other.

Table 3. KEGG data for Mrub_1285 and *E. coli* b2679. From the data, we can see the similarity between these two genes in term of gene name and sequence length etc.

	b2679	Mrub_1285
Gene Name	*proX*	
KEGG map	map02010 – ABC transporters	

DNA coordinates	2807132..2808124	1309888..1310892
DNA Sequence	atgcgacatagcgtacttttgcgacagcgtttgccac gcttatctctacacaaactttt gctgccgatctgccgggcaaaggcattactgttaatc cagttcagagcaccatcactgaa gaaaccttccagacgctgctggtcagtcgtgcgctg gagaaattaggttataccgtcaac aaacccagcgaagtagattacaacgttggctacacc tcgcttgcttccggcgatgcaacc ttcaccgccgtgaactggacgccactgcatgacaac atgtacgaagctgccggtggcgat aagaaattttatcgtgaaggggtatttgttaacggcgc ggcacagggttacctgatcgat aagaaaaccgccgaccagtacaaaatcaccaacat cgcacaactgaaagatccgaagatc gccaaactgttcgataccaacggcgacggaaaagc ggatttaaccggttgtaaccctggc tggggctgcgaaggtgcgatcaaccaccagcttgc cgcgtatgaactgaccaacaccgtg acgcataatcaggggaactacgcagcgatgatggc cgacaccatcagtcgctacaaagag ggcaaaccggtgttttattacacctggacgccgtact gggtgagtaacgaactgaagccg ggcaaagatgtcgtctggttgcaggtgccgttctccg cactgccgggcgataaaaacgcc gataccaaactgccgaatggtgcgaattatggcttcc cggtcagcaccatgcatatcgtt gccaacaaagcctgggccgagaaaaacccggcag cagcgaaactgtttgccattatgcag ttgccagtggcagatattaacgcccagaacgccatta tgcatgacggcaaagcctcagaa ggcgatattcagggacacgttgatggttggatcaaa gcccaccagcagcagttcgatggc tgggtgaatgaggcgctggcagcgcagaagtaa	atgcgaggaaaacttgttctactcagtctcgtcgtgg cctttggcactgctatgggccag caatgcgaggtaaatcggcccatcgtctttgccgact acgattgggaaagcgcccgcgtg cataaccggattgctcagttcatcctagagaagggct acgggtgtaagacagacgcccta ccgggcacttccatcccgctgatcaccggactgggc cgggggggacatcgatgtatccatg gaaatctggtacaacctgacccgcgacgtggttact caactggaaacggaggggaagata cagcgccttggggtaacctttcccgatgcggtgcag ggatggtttgtacccacttacgtg attaagggcgattcccaaaggggtatcaggcccatg gcgcccgacctgaagtccgttttt gaccttccaaagtacaagacgcttttccgcgacccc gaggagcccagcaaagggcgcttc tacaacggggtgctgggttggttcgcggaaagggtt aacaccaaaaagctcaaagcctac ggcctcgaggcccacttcaccaacttccgccccgg cacctccgatgccctggtggcggcc attgcttcggcctacgagcgggggcgtcccatcgtc ttttactactgggggcctacctgg gttctgggtaaatacgacctgaccatgctggaagaa ccctcctatgatgccgagacttgg aatgcccttatagggcaggacaacccctccaaggc caccgccttccccatggaaacggtt tacaacgcagtcaatacacgtctagcccgtgaggct ccttccgtggtggagttcctaaag aagtaccgcacctccaacgccctaaccagcgagct gctggcctacatggaggaaaaccgg gccaaggaggaggaggtggcccgccactttctgaa aacccatccagagctctggacggcc tgggtgcctgctgaagttgctgaaagagtgaagcga gcgctctaa
Sequence Length	993 nt	1005 nt

Protein Sequence	MRHSVLFATAFATLISTQTFAADL PGKGITVNPVQSTITEETFQTLLVS RALEKLGYTVNKPSEVDYNVGYT SLASGDATFTAVNWTPLHDNMY EAAGGDKKFYREGVFVNGAAQG YLIDKKTADQYKITNIAQLKDPKI AKLFDTNGDGKADLTGCNPGWG CEGAINHQLAAYELTNTVTHNQG NYAAMMADTISRYKEGKPVFYY TWTPYWVSNELKPGKDVVWLQV PFSALPGDKNADTKLPNGANYGF PVSTMHIVANKAWAEKNPAAAK LFAIMQLPVADINAQNAIMHDGK ASEGDIQGHVDGWIKAHQQQFD GWVNEALAAQK	MRGKLVLLSLVVAFGTAMGQQC EVNRPIVFADYDWESARVHNRIA QFILEKGYGCKTDALPGTSIPLITG LGRGDIDVSMEIWYNLTRDVVTQ LETEGKIQRLGVTFPDAVQGWFV PTYVIKGDSQRGIRPMAPDLKSVF DLPKYKTLFRDPEEPSKGRFYNG VLGWFAERVNTKKLKAYGLEAH FTNFRPGTSDALVAAIASAYERGR PIVFYYWGPTWVLGKYDLTMLEE PSYDAETWNALIGQDNPSKATAF PMETVYNAVNTRLAREAPSVVEF LKKYRTSNALTSELLAYMEENRA KEEEVARHFLKTHPELWTAWVP AEVAERVKRAL
Protein Sequence Length	330 aa	334 aa

BLASTp result of the last pair of genes is presented in figure 4. The first hit with lowest E-value of 1e-12 was Mrub_1285. The low E-value indicate the similarity due to evolutionary, not due to random chance.

Download GenPept Graphics

ABC transporter substrate-binding protein [Meiothermus ruber]

Sequence ID: WP_013013566.1 Length: 334 Number of Matches: 1

See 2 more title(s)

Range 1: 69 to 329 GenPept Graphics Next Match Previous Mat

Score	Expect	Method	Identities	Positives	Gaps
65.5 bits(158)	1e-12	Compositional matrix adjust.	71/274(26%)	111/274(40%)	29/274(10%)

```
Query  72   TSLASGDATFTAVNWTPLHDNMYEAAGGDKKFYREGVFVNGAAQGYLIDK---KTADQYK  128
            T L  GD   +   W  L  ++      + K  R GV    A QG+ +     K   Q
Sbjct  69   TGLGRGDIDVSMEIWYNLTRDVVTQLETEGKIQRLGVTFPDAVQGWFVPTYVIKGDSQRG  128

Query  129  ITNIA-QLKD----PKIAKLF-DTNGDGKADLTGCNPGWGCEGAINHQLAAYELTNTVTH  182
            I  +A  LK     PK   LF D     K        GW  E     +L AY L    T+
Sbjct  129  IRPMAPDLKSVFDLPKYKTLFRDPEEPSKGRFYNGVLGWFAERVNTKKLKAYGLEAHFTN  188

Query  183  NQ-GNYAAMMADTISRYKEGKPVFYYTWTPYWVSNELKPGKDVVWLQVP------FSALP  235
             + G   A++A   S Y+ G+P+ +Y W P WV  +     D+  L+ P      ++AL
Sbjct  189  FRPGTSDALVAAIASAYERGRPIVFYYWGPTWVLGKY----DLTMLEEPSYDAETWNALI  244

Query  236  GDKNADTKLPNGANYGFPVSTMHIVANKAWAEKNPAAAKLFAIMQLPVADINAQNAIMHD  295
            G  N     P+ A   FP+ T++   N   A + P+  +     +   A  +   A M +
Sbjct  245  GQDN-----PSKAT-AFPMETVYNAVNTRLAREAPSVVEFLKKYRTSNALTSELLAYMEE  298

Query  296  GKASEGDIQGHVDGWIKAHQQQFDGWVNEALAAQ  329
             +A E ++  H   ++K H + +  WV   +A +
Sbjct  299  NRAKEEEVARH---FLKTHPELWTAWVPAEVAER  329
```

Figure 4. BLASTp result of b2679 against *M. ruber* genome was showed in figure 4. The first hit with lowest E-value of 1e-12 was Mrub_1285. The low E-value indicated the similarity due to evolutionary, not due to random chance.

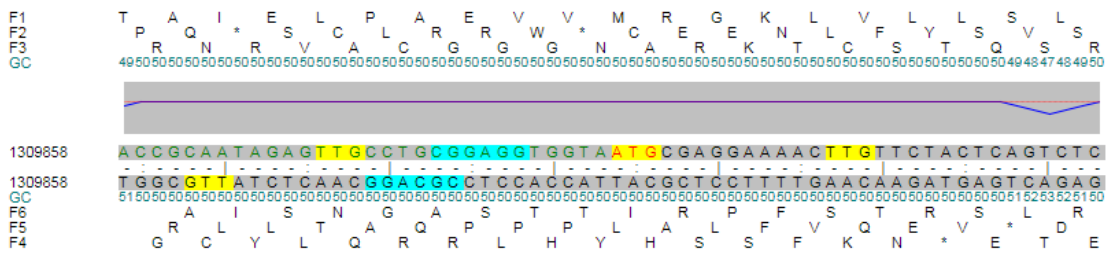
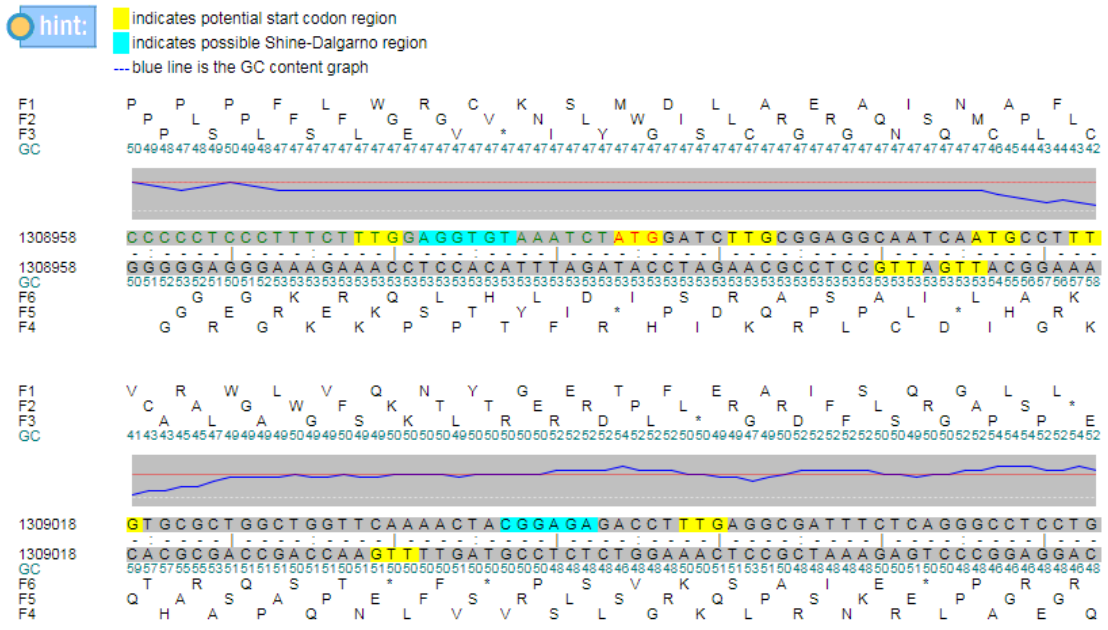
Section II . Alternate Open Reading Frame for *M. ruber* genes

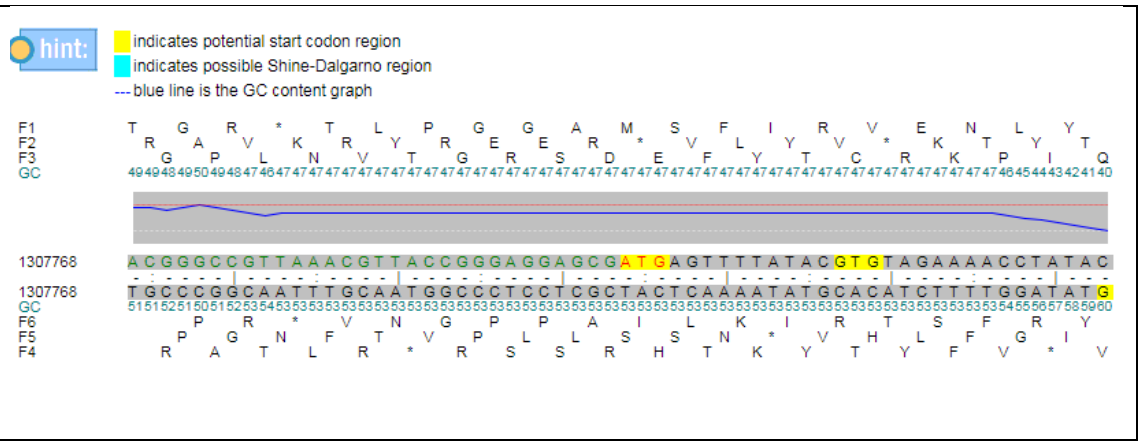
While the reading frame for all *E. coli* genes is very well studied, the reading frame for *Meiothermus ruber* are less familiar and become interesting to take a closer look. To determine whether the starting codon of *M. ruber* genes are called correctly, we have two alternate approaches, one using the relative distance of the starting codon vs potential Shine-Dalgarno sequence and the other looking at the start codon of the WebLogo created by 15-20 species from a different genus. The results of the first approach are presented in table 4 below while the results of the WebLogo approach are presented in table 5.

In table 4, the codon triplet in yellow shade indicates potential start codon while the sequence in cyan shade indicates potential Shine-Dalgarno sequences (SDs). The triple in red is

the start codon that is used. Since we expect our start codon to be at the distance of 6-10 nucleotides from the SDs, we expect to see a SD about 6-10 nucleotides upstream of the triplet in red. Base on the relative position with the SDs, we see the start codon of Mrub_1285 and Mrub_1284 are called correctly. However, in Mrub_1283, we don't see any SD near the start codon that is called. Therefore, we need to use a different approach to verify the start codon of Mrub_1283.

Table 4. Alternate Open Reading Frame suggestion using Shine-Dalgarno Sequence.

M. ruber gene	Relative location of the gene with potential Shine-Dalgarno sequences
Mrub_1285	
Mrub_1284	

Mrub_1283	

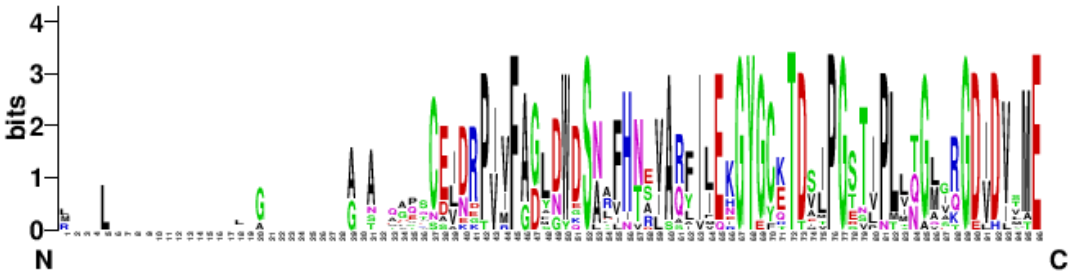
Another approach to find start codon is using WebLogo. In this section, we create a Weblogo use the multi-alignment of 15-20 species from various strains. The multi-alignment of these species is created by T-coffee. The first line of the multi-alignment created by T-coffee is shown in table 5. The first line of WebLogo for all *M. ruber* GOIs are also showed in table 5. Since we expect the start codon M to be highly conserved between these various species, the expect to see the M amino acid in the first place of multi-alignment as well as in the first position of the WebLogo. As we can see, the start codons of both Mrub_1285 and Mrub_1284 are not conserved throughout different species so we cannot use the WebLogo to conclude about the start codon for these two genes. However, for Mrub_1283, the M start codon is highly conserve so we expect that the called start codon we have is correct for this gene based on WebLogo.

Table 5. Alternate Open Reading Frame suggestion using T-coffee and WebLogo

M. ruber gene	T-coffee multi-alignment and WebLogo

Mrub_1285	`CLUSTAL W (1.83) multiple sequence alignment`
	
	`Alphaproteobacteria_bacterium  ------------------------------------CELNRPIIFAGLDW`

```
Balneatrix_alpica                -------------------------------------CEVDKPVRFAGMNW
Chelatococcus_daeguensis         ----------------------------------QSCEVDRPVVFGDLDW
Desulfotomaculum_geothermicum    LRNKLPLLVLITALFFTLAGVAGCSSGEADNA-GESNSAKETIVFADYNW
Gammaproteobacteria_bacterium    MI-GLSVL---ISS-GLLIGQS------AVAEEETKCDIERPIVFAGSDW
Marinithermus_hydrothermalis     RV-QLGL----IAL-ALTLG-V------AFA-QVPECELDRPVVFAGLDW
Meiothermus_ruberDSM_1279        M-------------------AF------GTA-MGQQCEVNRPIVFADYDW
Neomegalonema_perideroedes       RV-ALGFF--------SILGVS------AAS-QAEACELNRPIVFAGLDW
Nesiotobacter_exalbescens        -----------------------------------QCEIDRPVVFAGLDW
Parvibaculum_lavamentivorans     ----------------------------GAA-AAPSCAIDRPVMFGGLDW
Planifilum_fulgidum              --------------------------------------DDPIIFADAGW
Pseudovibrio_hongkongensis       L---------------PLVATT------AQA-AGPVCEIDRPVVFAGLDW
Rhodobiaceae_bacterium           ----------------------------GVA-TAQTCEIDRPIIFGGLDW
Sphaerobacter_thermophilus       ------------------------------TA-PGSSDLDGPIVFADFGW
Thalassospira_profundimaris      ----------------------------NA-QDATCEIDRPVMFAGLNY
                                                                          . .: *.. .:
```

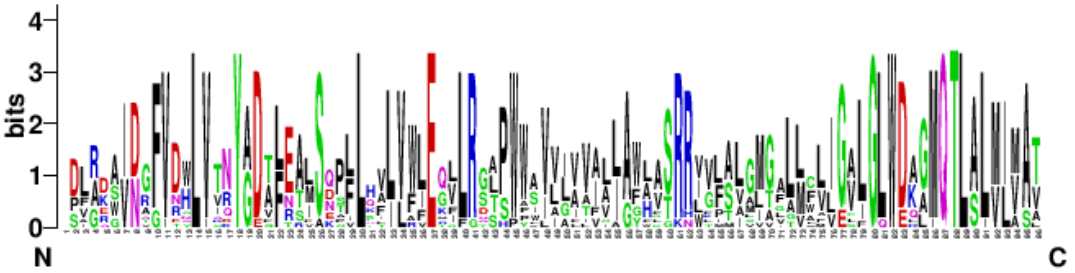


Mrub_1284

```
CLUSTAL W (1.83) multiple sequence alignment

Achromobacter_xylosoxidans       -----AIDGFVDHLVTNYADTLETLSQPVLHALVWLEQVLRSSPWWAVVG
Advenella_kashmirensis           -DVRKSIDGFVDHLVTNYADALNAMSEPFLHLLVWIEKILRGAPWWSVLI
Advenella_mimigardefordensis     -DVRKSIDGFVDHLVTNYADALNAMSEPFLHVLVWLEKVLRGAPWWSVII
Bordetella_ansorpii              ------IDTFVDHLVTQYADTLESLSKPFLHVLVWLEQLLRSAPWWAVVL
Bordetella_parapertussis         -----AIDGFVDHLVTNYADTLEAMSQPFLSVLVWLEQVLRQSSWWAVVA
Kerstersia_gyiorum               -DLRRAIDGFVENLVSRYADTLETLSKPFLHVLVWLEQVLRGTPWWVVVL
Marinithermus_hydrothermalis     LPLGDWVDAFVNWLVIQYGDAFEALSNSLLFVLVRLERFLGTLPWWSVVL
```

```
Meiothermus_ruber                MDLAEAINAFVRWLVQNYGETFEAISQGLLSFLLFFEGLLRDLSWFWVAG
Neomegalonema_perideroedes       --IADGVNQFVRYLIVNYGDGFEAVSNFILKILLFIETGLRDLHPAILLI
Oligella_ureolytica              -DARRSIDGFVDGLVVKYADALTAMSQPFLKTLVWIEMVLRSAPWWSIVI
Pseudomonas_psychrotolerans      -SFADAVNRGVDWLVTRYGDVFRAISDTLLQAIVWLEGLLRGTPWWAILL
Pseudomonas_psychrotolerans_1    -SFADAVNRGVDWLVTRYGDVFRTISDTLLQAIVWLEGLLRGTPWWAILL
Saccharospirillum_impatiens      --LRQGIDGFVNQLVIDYADTLERLSQPLLNFLVFVEQLLRNSPWYLVIA
Thalassospira_lucentensis        --LGKWVNRFVDWLVVNYGDAFEAFADSLLTVLVWLEQILRGTPWWIVVI
Thermus_thermophilus             IPLGEWVNVFITWLVRNYGDFFESLSNALLQFILAFEGFFRSLSWPWVAG
                                       ::  :  *:  *.: :  .:. .*  :: .*  :       :
```

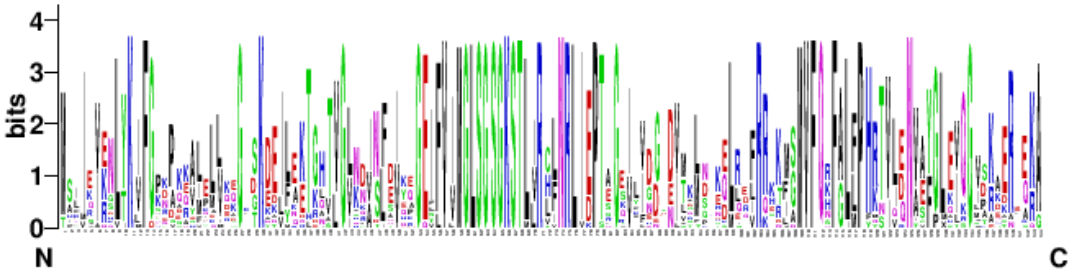


Mrub_1283

```
CLUSTAL W (1.83) multiple sequence alignment

Bacillus_thermozeamaize                  --IIRVRQLTKIFGPQPERALQLLRQGKDKAEIFKELEATVGVNQATFDV
Chlamydia_abortus                        MAVISVKNLTKMFGSEVGRAFPLLEQGLSKKEIYEKTKITVGVNRISFDI
Desulfobacteraceae_bacterium             ---MQVRNLYKVFGASPKEALELLQQGAGKDEILEKTGQSVGLADINFDI
Desulfonispora_thiosulfatigenes          MIQIKVENLYKVFGSNPQRAIKLIEEGMDKDEILNKTGLAVGVGGVSFEV
Endozoicomonas_sp.S-B4-1U                --LIAIRNLYKVFGRKPEQAMTKVKAGVGKDELLAEHDHTLGLKDINLTI
Firmicutes_bacterium_HGW-Firmicutes-14   MSKIEVKNLYKIFGKQPDKAVELLQKGESKENILNQTGQVVGLRNVSFSI
Halanaerobium_kushneri                   ---IKAENLYKIFDGKGQKEIEMLKSGKSKDDILEKTGATVGINNASFEV
Marinithermus_hydrothermalis             MAKIEVRHLYKIFGPRPKEVLALLKKGVGKEEVYRKTGHTVGLHDVSFSV
Meiothermus_ruber                        MSFIRVENLYKIFGPKAGQALEMVQGGTDKDTLFQKTRHVLGLNRINLEV
Methanosarcina_flavescens                ---IEIKNLIKIFGKNPQEVLSLLQEGRTKNEIFEKTKQTVGLNNININV
Neptuniibacter_caesariensis              --LIEIESLYKLFGNNPAKYMPLVHEGKSKDEILAETGHTLGLKDINLQI
Oleiphilus_sp.HI0122                     T-FIEIKSLYKLFGEDASKHMDLVYKGLSKTEILEKTGHTLGLKDINLDI
Ornithinibacillus_halophilus             MSIIKIENLYKVFGKEPKAAIELVEQGYTKEEIMEKTGNTVGINNVSFEV
Paenibacillus_daejeonensis               MPIIEVKKLTKIFGPDAKRAIPLLNDGWSKEKILKETKLTVGVNQASFSI
```

```
Paenibacillus_sp.yr247                    MHIVEVNNLTKVFGQDPQKALLLLDKGWSKKKIYEETKHTVGVNKVNLAI
Photobacterium_sanguinicancri             --LISVKNLYKVFGPNDKKVLEQVKAGKSKDDILADTGHTVGLNDINLDV
Salisediminibacterium_haloalkalitolerans  MSEIKVEGLTKIFGKRPKQGLKLLDEGKTKDEILEETGLTVGVNKADFEV
Shewanella_sp.GutCb                       --LIQIRDLYKVFGKKPANVMPMVKEGLSKDEILAKTGHTVGLKAINLDI
Thermus_thermophilus                      MSYIRIEGVYKIFGPRAKQVLEEVRRGAGKDEVFQKTRHVVGLKNVNLEI
Thiohalospira_halophila                   ---IVVENLYKIFGPDPDEAFELMDQGQDKDAIFEKTGNTVGVKDANFAI
Vibrio_hangzhouensis                      --IIQIKNLYKIFGPKDKAYLQAVKDGETKDELLARTGHTLGLQDINLDV
                                             :  . : *:*.      .  :  *  *  :       :*:    : :
```



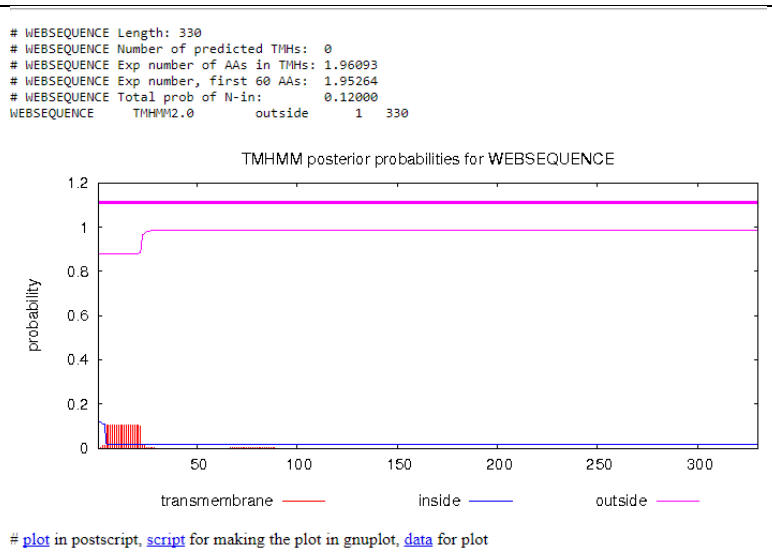
In conclusion, the start codon of each gene Mrub_1285, Mrub_1284 and Mrub_1283 are confirmed by either the Shine-Dalgarno sequence or by the multi-alignment of sequences presented in the WebLogo. However, none of the gene has its start codon confirmed by both approaches. Therefore, to be certain about the reading frame of all our *M. ruber* GOIs, we expect to have an additional test to identify the start codon of each genes.

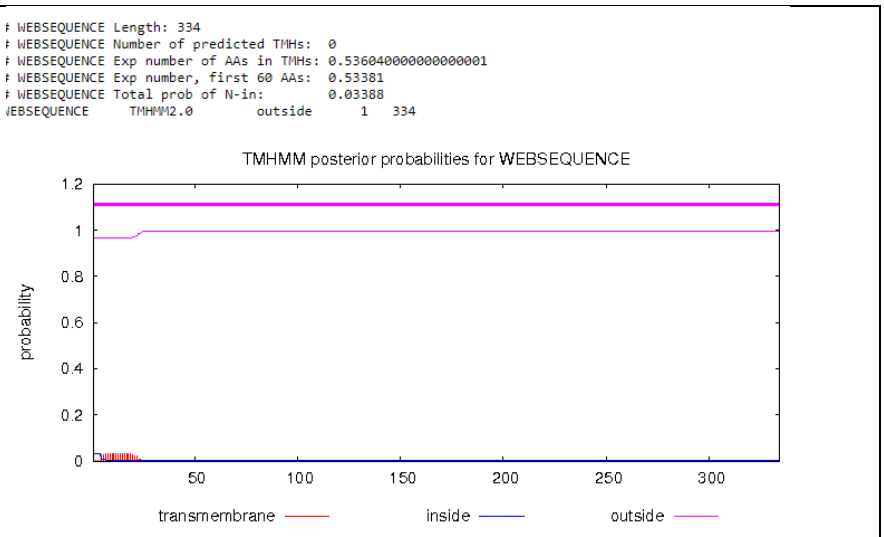
Section III – Cellular Localization Data

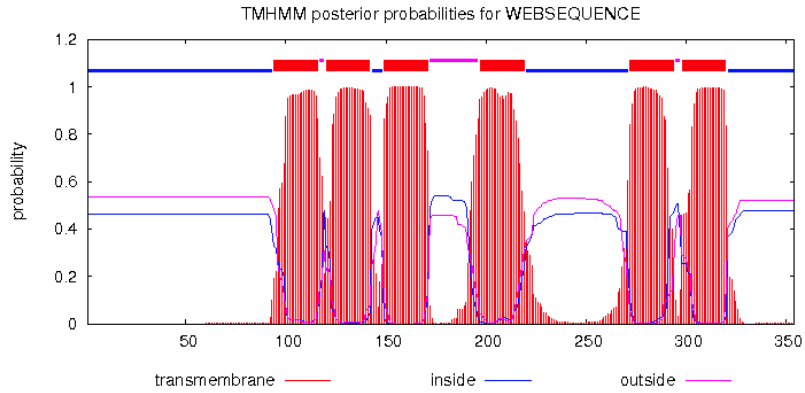
Cellular localization data is also helpful for us to determine the function of *M. ruber* gene and their orthologue with *E. coli* genes. In this section, we report data from various sources regard of location of the protein inside the cell. For the purpose of confirming orthologue, data for each *M. ruber* gene will be put next to data of the *E. coli* gene that is potentially orthologous.

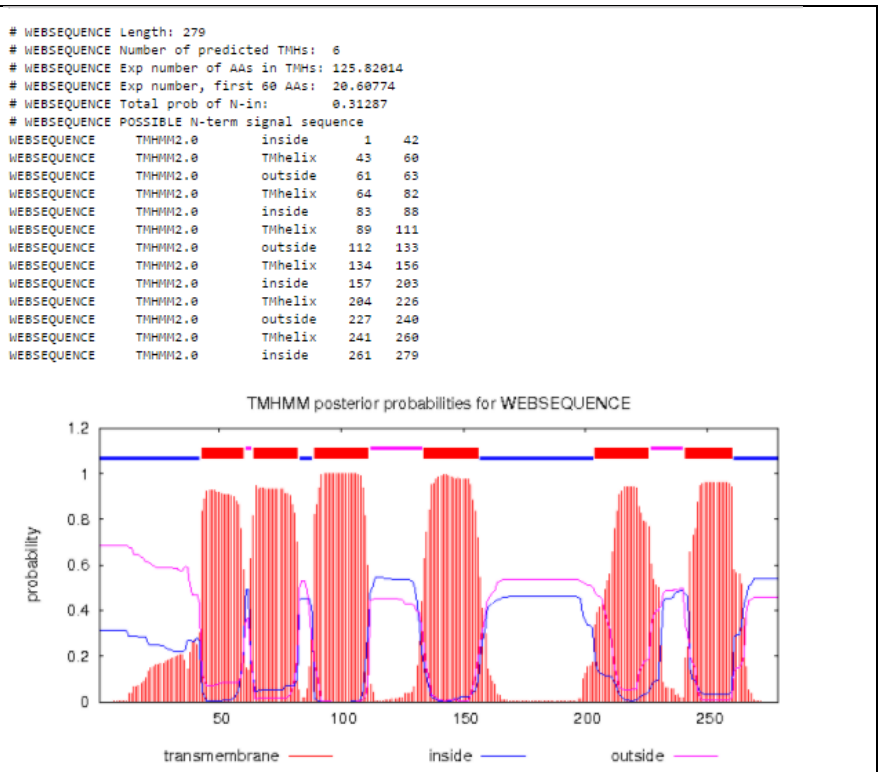
Table 6 below shows the result of TMHMM, the tool that can predict the number of transmembrane helices from the sequence of protein.

Table 6. TMHMM results of GOIs

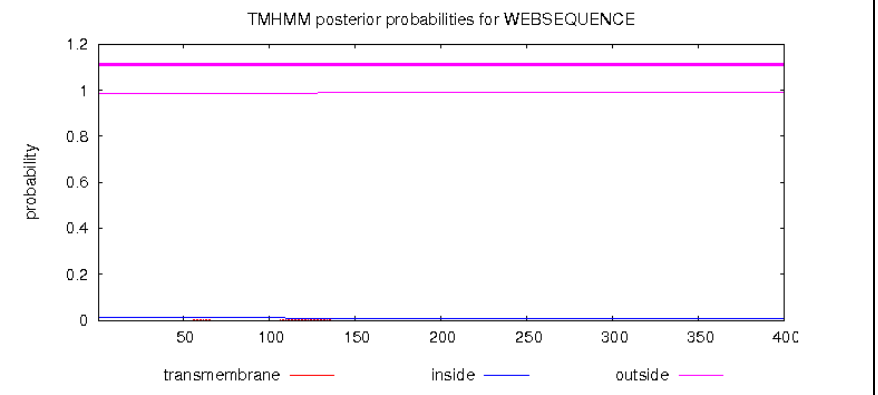
	TMHMM	
	# of transmem-brane helices	**Transmembrane topology graph**
b2679	**0**	 # WEBSEQUENCE Length: 330 # WEBSEQUENCE Number of predicted TMHs: 0 # WEBSEQUENCE Exp number of AAs in TMHs: 1.96093 # WEBSEQUENCE Exp number, first 60 AAs: 1.95264 # WEBSEQUENCE Total prob of N-in: 0.12000 WEBSEQUENCE TMHMM2.0 outside 1 330 # plot in postscript, script for making the plot in gnuplot, data for plot

Mrub1285	**0**	
		`# WEBSEQUENCE Length: 334`
		`# WEBSEQUENCE Number of predicted TMHs: 0`
		`# WEBSEQUENCE Exp number of AAs in TMHs: 0.536040000000000001`
		`# WEBSEQUENCE Exp number, first 60 AAs: 0.53381`
		`# WEBSEQUENCE Total prob of N-in: 0.03388`
		`WEBSEQUENCE TMHMM2.0 outside 1 334`

b2678	**6**	
		`# WEBSEQUENCE Length: 354`
		`# WEBSEQUENCE Number of predicted TMHs: 6`
		`# WEBSEQUENCE Exp number of AAs in TMHs: 135.43334`
		`# WEBSEQUENCE Exp number, first 60 AAs: 0.00061`
		`# WEBSEQUENCE Total prob of N-in: 0.46316`
		`WEBSEQUENCE TMHMM2.0 inside 1 93`
		`WEBSEQUENCE TMHMM2.0 TMhelix 94 116`
		`WEBSEQUENCE TMHMM2.0 outside 117 119`
		`WEBSEQUENCE TMHMM2.0 TMhelix 120 142`
		`WEBSEQUENCE TMHMM2.0 inside 143 148`
		`WEBSEQUENCE TMHMM2.0 TMhelix 149 171`
		`WEBSEQUENCE TMHMM2.0 outside 172 196`
		`WEBSEQUENCE TMHMM2.0 TMhelix 197 219`
		`WEBSEQUENCE TMHMM2.0 inside 220 271`
		`WEBSEQUENCE TMHMM2.0 TMhelix 272 294`
		`WEBSEQUENCE TMHMM2.0 outside 295 297`
		`WEBSEQUENCE TMHMM2.0 TMhelix 298 320`
		`WEBSEQUENCE TMHMM2.0 inside 321 354`
		# plot in postscript, script for making the plot in gnuplot, data for plot

Mrub1284	**6**	

```
# WEBSEQUENCE Length: 279
# WEBSEQUENCE Number of predicted TMHs:  6
# WEBSEQUENCE Exp number of AAs in TMHs: 125.82014
# WEBSEQUENCE Exp number, first 60 AAs:  20.60774
# WEBSEQUENCE Total prob of N-in:        0.31287
# WEBSEQUENCE POSSIBLE N-term signal sequence
WEBSEQUENCE     TMHMM2.0        inside       1    42
WEBSEQUENCE     TMHMM2.0        TMhelix     43    60
WEBSEQUENCE     TMHMM2.0        outside     61    63
WEBSEQUENCE     TMHMM2.0        TMhelix     64    82
WEBSEQUENCE     TMHMM2.0        inside      83    88
WEBSEQUENCE     TMHMM2.0        TMhelix     89   111
WEBSEQUENCE     TMHMM2.0        outside    112   133
WEBSEQUENCE     TMHMM2.0        TMhelix    134   156
WEBSEQUENCE     TMHMM2.0        inside     157   203
WEBSEQUENCE     TMHMM2.0        TMhelix    204   226
WEBSEQUENCE     TMHMM2.0        outside    227   240
WEBSEQUENCE     TMHMM2.0        TMhelix    241   260
WEBSEQUENCE     TMHMM2.0        inside     261   279
```


b2677	**0**	

```
# WEBSEQUENCE Length: 400
# WEBSEQUENCE Number of predicted TMHs:  0
# WEBSEQUENCE Exp number of AAs in TMHs: 0.15549
# WEBSEQUENCE Exp number, first 60 AAs:  0.00147
# WEBSEQUENCE Total prob of N-in:        0.01411
WEBSEQUENCE     TMHMM2.0        outside      1   400
```

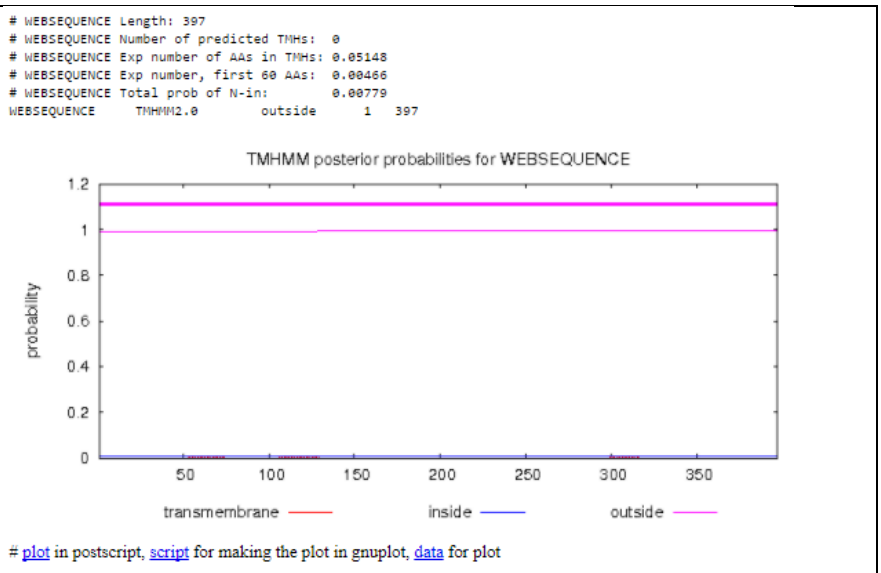
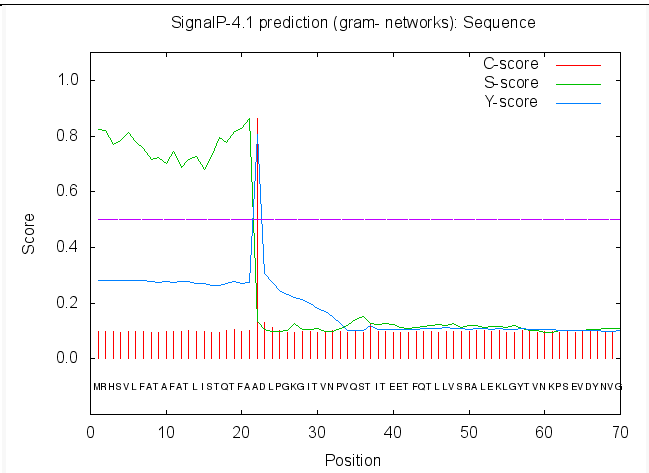
Mrub1283	**0**	 # WEBSEQUENCE Length: 397 # WEBSEQUENCE Number of predicted TMHs: 0 # WEBSEQUENCE Exp number of AAs in TMHs: 0.05148 # WEBSEQUENCE Exp number, first 60 AAs: 0.00466 # WEBSEQUENCE Total prob of N-in: 0.00779 WEBSEQUENCE TMHMM2.0 outside 1 397 # plot in postscript, script for making the plot in gnuplot, data for plot

Table 7 below shows the result of SignalP, the tool to predict whether or not the protein is a signal protein.

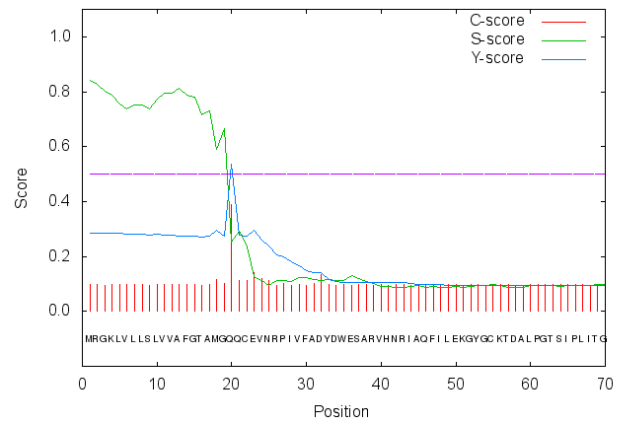
Table 7. SignalP data for all GOIs.

	SignalP	
	Signal peptide probability	Signal peptide graph

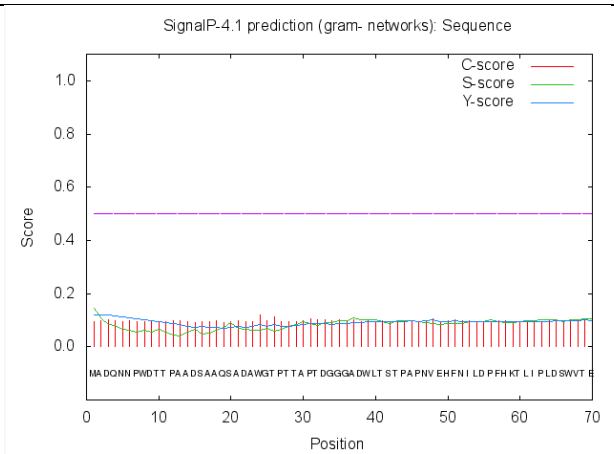
<table>
<tr>
<td>b2679</td>
<td>0.787

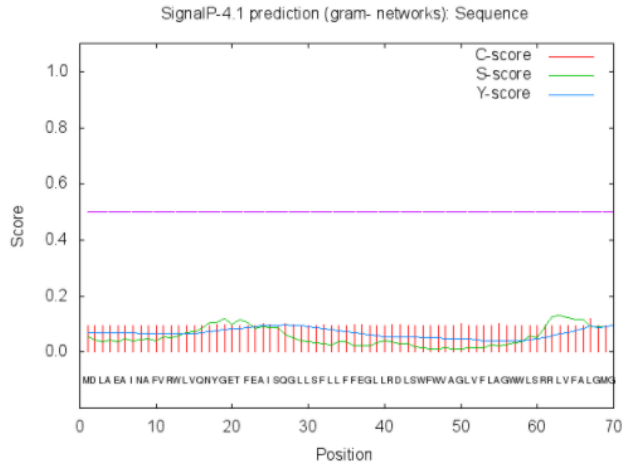
Cleavage 21-22</td>
<td>


<pre>
Measure Position Value   Cutoff   signal peptide?
  max. C   22   0.863
  max. Y   22   0.807
  max. S   21   0.862
  mean S   1-21   0.764
       D   1-21   0.787   0.570   YES
Name=Sequence   SP='YES' Cleavage site between pos. 21 and 22: TFA-AD D=0.787 D-cutoff=0.570 Networks=SignalP-noTM
</pre>
</td>
</tr>
<tr>
<td>Mrub1285</td>
<td>0.642

Cleavage 19-20</td>
<td>


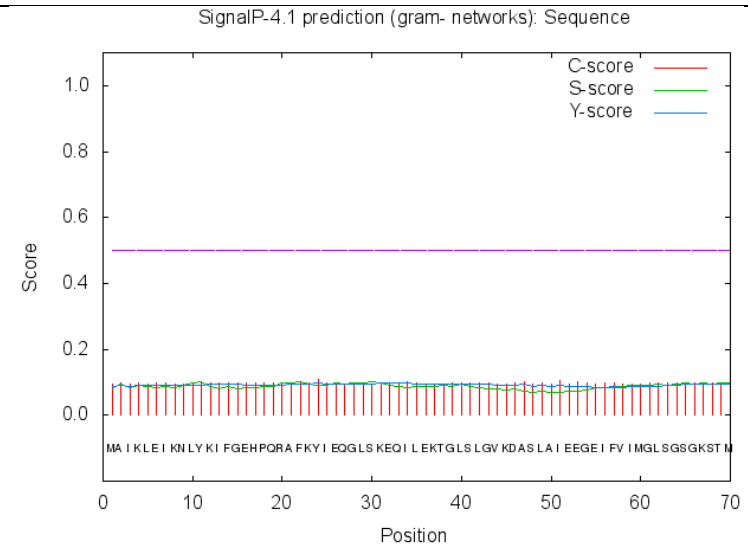
<pre>
Measure Position Value   Cutoff   signal peptide?
  max. C   20   0.388
  max. Y   20   0.537
  max. S   1   0.841
  mean S   1-19   0.760
       D   1-19   0.642   0.570   YES
Name=Sequence   SP='YES' Cleavage site between pos. 19 and 20: AMG-QQ D=0.642 D-cutoff=0.570 Networks=SignalP-noTM
</pre>
</td>
</tr>
</table>

b2678	0.091
	
	Not a signal
	
	peptide
		
		`# Measure Position Value Cutoff signal peptide?`
		`max. C 24 0.120`
		`max. Y 48 0.102`
		`max. S 1 0.145`
		`mean S 1-47 0.078`
		`D 1-47 0.091 0.570 NO`
		`Name=Sequence SP='NO' D=0.091 D-cutoff=0.570 Networks=SignalP-noTM`
		`# data`
		`# gnuplot script`

Mrub1284	0.088
	
	Not a signal
	
	peptide
		
		`Measure Position Value Cutoff signal peptide?`
		`max. C 67 0.120`
		`max. Y 27 0.099`
		`max. S 63 0.129`
		`mean S 1-26 0.070`
		`D 1-26 0.088 0.510 NO`
		`lame=Sequence SP='NO' D=0.088 D-cutoff=0.510 Networks=SignalP-TM`

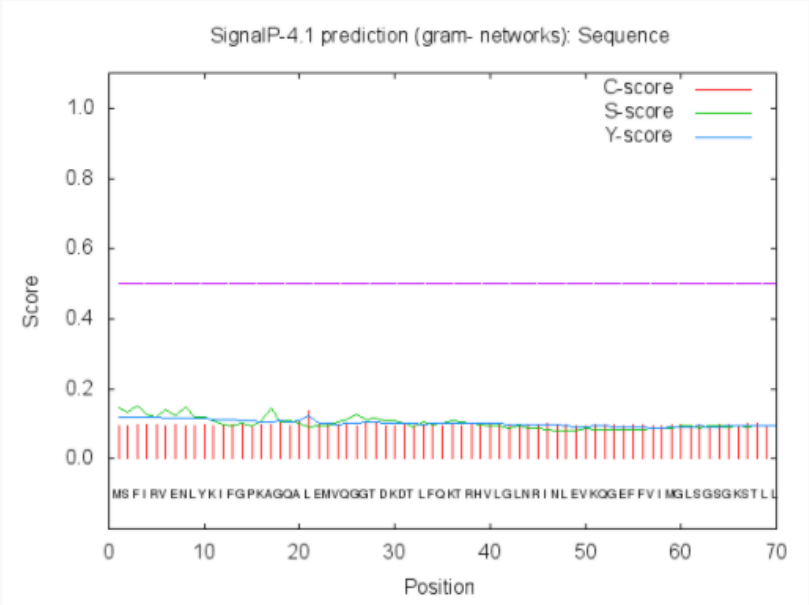
<table>
<tr>
<td>B2677</td>
<td>0.094

Not a signal

peptide</td>
<td>


<pre>
Measure Position Value   Cutoff   signal peptide?
  max. C   24   0.108
  max. Y   24   0.098
  max. S   22   0.102
  mean S   1-23   0.089
       D   1-23   0.094   0.570   NO
Name=Sequence   SP='NO' D=0.094 D-cutoff=0.570 Networks=SignalP-noTM
</pre>
</td>
</tr>
<tr>
<td>Mrub1283</td>
<td>0.121

Not a signal

peptide</td>
<td>
<pre>
SignalP-4.1 gram- predictions
>Sequence
</pre>


<pre>
Measure Position Value   Cutoff   signal peptide?
  max. C   21   0.138
  max. Y   21   0.123
  max. S   3   0.153
  mean S   1-20   0.120
       D   1-20   0.121   0.570   NO
Name=Sequence   SP='NO' D=0.121 D-cutoff=0.570 Networks=SignalP-noTM
# data
# gnuplot script
</pre>
</td>
</tr>
</table>

Table 8 below shows the result of LipoP tool, the tool that categorizes the type of signal proteins. Since b2678, b2677, Mrub_1284 and Mrub_1283 are not signal protein, we will only report the result of LipoP for b2679 and Mrub_1285. According to LipoP, both b2679 and Mrub_1285 are signal proteins of type SP1, each protein has the cleavage site very proximate to the site of each other. Therefore, LipoP data can be used to confirm the orthologue between b2679 and Mrub_1285.

Table 8. LipoP data for b2679 and Mrub_1285

	LipoP	
	Best Prediction	Cleavage site after AA#
b2679	SP1	21
Mrub_1285	SP1	19

To confirm the cellular localization of proteins products of genes, we also collect the data from PSORT-B, a tool that predict the localization of the protein along with a probabilistic score. Table 9 below will report the PSORT-B final prediction for each protein product as long as the score for that prediction. We can see that from PSORT-B data, all pairs of predicted orthologous genes are predicted to locate in the same cellular site. Therefore, we can use PSORT-B data to support our hypothesis about the orthologue between *M. ruber* and *E. coli* genes.

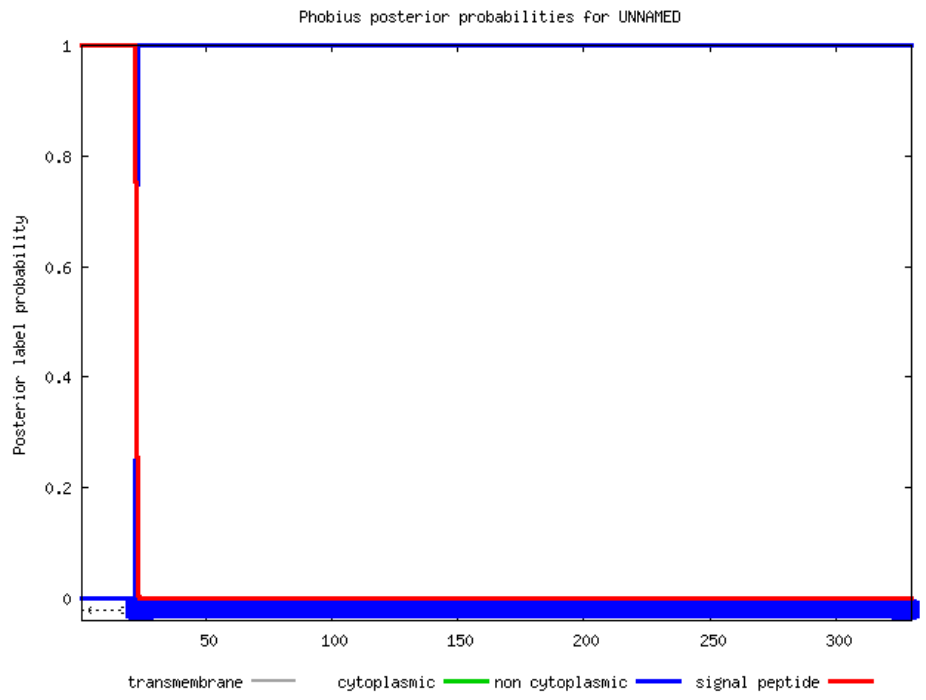
Table 9. PSORT-B prediction of protein products along with score

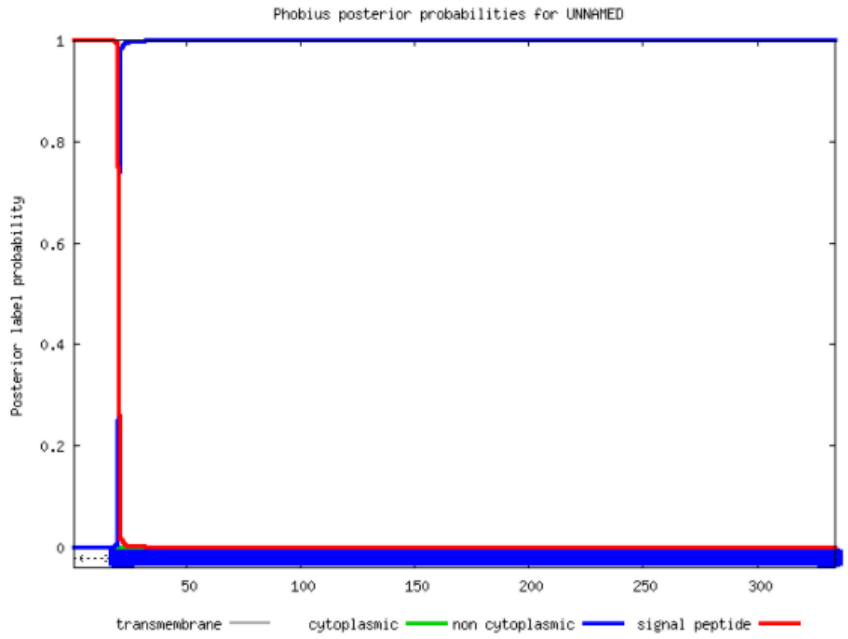
	PSORT-B	
	Final prediction	Score
b2679	Periplasmic space	10.00

Mrub_1285	Periplasmic space	9.76
b2678	Cytoplasmic Membrane	10.00
Mrub_1284	Cytoplasmic Membrane	10.00
b2677	Cytoplasmic Membrane	10.00
Mrub_1283	Cytoplasmic Membrane	9.99

Finally, to have an overall look at the location of the protein in cell, we collect the data from Phobius and report in Table 10 below. Again, all pairs of predicted orthologous genes show very similar Phobius probability graph indicate the evolutionary relation.

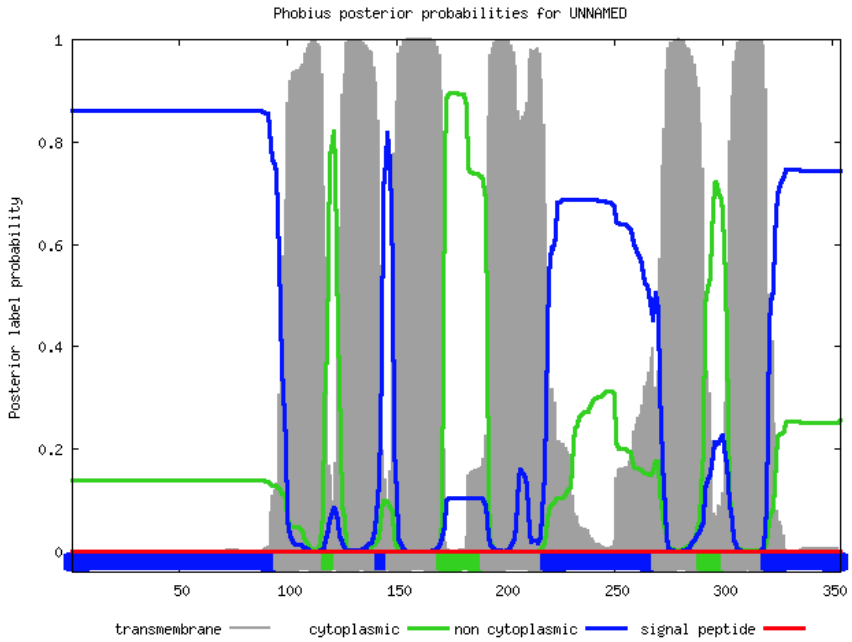
Table 10. Data from Phobius about probability of protein in different cell location

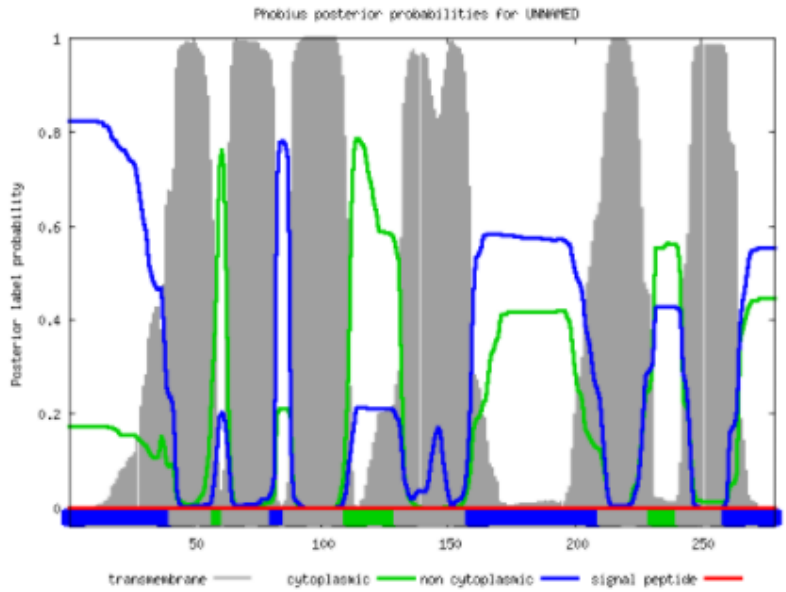
	Phobius probability graph
b2679	ID UNNAMED FT SIGNAL 1 21 FT REGION 1 4 N-REGION. FT REGION 5 16 H-REGION. FT REGION 17 21 C-REGION. FT TOPO_DOM 22 330 NON CYTOPLASMIC. // 

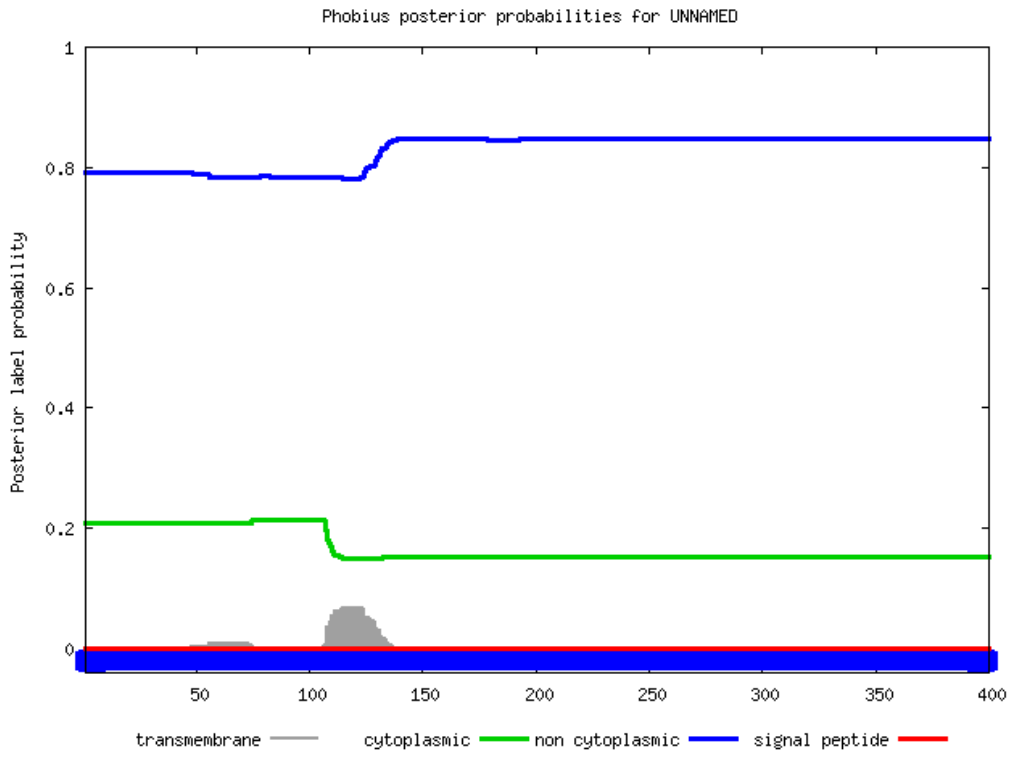
Mrub_1285	**Prediction of UNNAMED** ID UNNAMED FT SIGNAL 1 19 FT REGION 1 4 N-REGION. FT REGION 5 14 H-REGION. FT REGION 15 19 C-REGION. FT TOPO_DOM 20 334 NON CYTOPLASMIC. // 

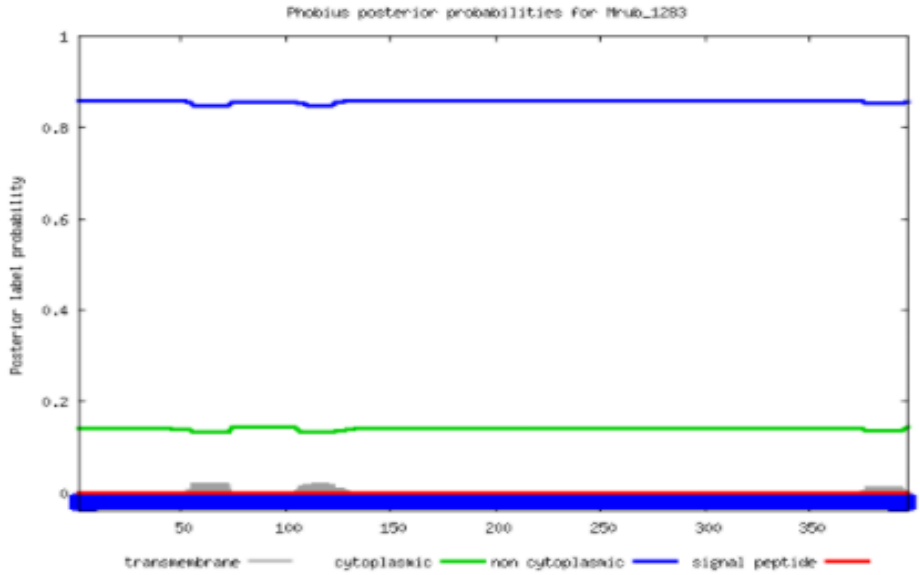
b2678

```
ID   UNNAMED
FT   TOPO_DOM      1     96       NON CYTOPLASMIC.
FT   TRANSMEM     97    118
FT   TOPO_DOM    119    124       CYTOPLASMIC.
FT   TRANSMEM    125    143
FT   TOPO_DOM    144    148       NON CYTOPLASMIC.
FT   TRANSMEM    149    171
FT   TOPO_DOM    172    191       CYTOPLASMIC.
FT   TRANSMEM    192    219
FT   TOPO_DOM    220    270       NON CYTOPLASMIC.
FT   TRANSMEM    271    291
FT   TOPO_DOM    292    302       CYTOPLASMIC.
FT   TRANSMEM    303    320
FT   TOPO_DOM    321    354       NON CYTOPLASMIC.
//
```



Mrub_1284	**Prediction of UNNAMED** ID UNNAMED FT TOPO_DOM 1 42 NON CYTOPLASMIC. FT TRANSMEM 43 59 FT TOPO_DOM 60 63 CYTOPLASMIC. FT TRANSMEM 64 82 FT TOPO_DOM 83 87 NON CYTOPLASMIC. FT TRANSMEM 88 111 FT TOPO_DOM 112 131 CYTOPLASMIC. FT TRANSMEM 132 159 FT TOPO_DOM 160 211 NON CYTOPLASMIC. FT TRANSMEM 212 231 FT TOPO_DOM 232 242 CYTOPLASMIC. FT TRANSMEM 243 260 FT TOPO_DOM 261 279 NON CYTOPLASMIC. // 

b2677	**Prediction of UNNAMED** ID UNNAMED FT TOPO_DOM 1 400 NON CYTOPLASMIC. // 

Mrub_1283	**Phobius prediction** **Prediction of Mrub_1283** ID Mrub_1283 FT TOPO_DOM 1 397 NON CYTOPLASMIC. //  The probability data used in the plot is found here, and the gnuplot script is here

Section IV – Structure-based Evidence Module

In this section, we collect several structural based evidence to prove that Mrub_1285, Mrub_1284, and Mrub_1283 are orthologous to *proX*, *proW* and *proV* in *E. coli* genomes, respectively. We collect data from different protein family database like CDD, TIGRFAM, PFAM, PDB. The data is reported below in three separate tables for three pairs of orthologous genes.

In table 11 below, b2679 and Mrub_1285 are orthologous due to several different structural data from a variety of bioinformatics tools such as CDD, PFAM. The CDD searches for both sequences result in the same COG2113 result. COG2113 is the ABC-type proline/glycine betaine transport system, periplasmic component, which is consistent with the

function of b2679 in literature. Since Mrub_1285 is orthologous to b2679 as shown by several informatics data, Mrub_1285 is expected to encode for the periplasmic component of ABC transporter as well. There are no TIGRFAM hits for both sequence so TIGRFAM data cannot be used to confirm the orthologue between two genes. The PFAM searches of both sequences result in the same PF04069 hit, which is the substrate binding domain of ABC-type glycine betaine transport system. From PFAM result, we predict that Mrub_1285 encode for the substrate binding domain of the ABC transporter. CL0177 – periplasmic binding protein clan is found as the first hit of the search for both sequences. This Clan result also supports our hypothesis that Mrub_1285 is orthologs of b2679. All the hits are associated with very low E-value which indicates the significance of the data found.

Table 11: b2679 and Mrub_1285 are orthologous according to structural-data from a variety of bioinformatics tools. For this pair of gene, there are no TIGRFAM hits with reasonable E-value so no TIGRFAM result was reported.

Categories	*E. coli* b2679 (*proX*)	*M. ruber* Mrub_1285
CDD data	COG2113 ABC-type proline/glycine betaine transport system, periplasmic component	
	Score: 668.57 E-value: 0e+00	Score: 215 E-value: 4.0e-57
PFAM – Protein family	PF04069	

	Substrate binding domain of ABC-type glycine betaine transport system	
	E-value: 4.3e-63	E-value: 3.3e-70
Clan	CL0177 Periplasmic binding protein clan	
	E-value: 5.4e-56	E-value: 9.6e-60
Highly conserved amino acids (HMM logo)	G28, G49, Y143	G28., G49, Y143
PDB protein	1R9L ProX in complex with glycine betaine	
	E-value: 0.0	E-value: 1.28099e-10

In table 12 below, b2678 and Mrub_1284 are orthologous due to several different structural data from a variety of bioinformatics tools such as TIGRFAM and PFAM. The CDD searches for both sequences result from no reasonable COG value so we can use CDD data to support our hypothesis. The TIGRFAM searches of both sequences result in the same TIGR03416 hit, which is the ABC_choXWV_perm: choline ABC transporter family. The TIGRFAM of choline ABC transporter seems unrelated to the function as glycine betaine transporter. However, in bacteria, the primary role of choline is the precursor of glycine/betaine so in most bacteria, the transport mechanism of choline and glycine betaine are similar (Wargo *et al*, 2013). Hence, we can still use the TIGR03416 as evidence to understand the function and structure of the protein encoded by Mrub_1284. The PFAM searches of both sequences result in

the same PF00528 hit, which is the BPD_transp_1: ABC transporter, permease protein. From PFAM result, we predict that Mrub_1285 encode for the transmembrane domain of the ABC transporter. CL0404 – BpD_transp_1 clan are found as the first hits of the searches for both sequences. This Clan result also supports our hypothesis that Mrub_1284 is orthologs of b2678. All the hits are associated with very low E-value which indicates the significance of the data found.

Table 12: b2678 and Mrub_1284 are orthologous according to structural data from a variety of bioinformatics tools. There are no COG found from CDD search so COG data is not included in the table.

Categories	*E. coli* b2678 (*proW*)	*M. ruber* Mrub_1283
TIGRFAM – Protein family	TIGR03416 ABC_choXWV_perm: choline ABC transporter	
	E-value: 6.2e-95	E-value: 5e-97
PFAM – Protein family	PF00528 BPD_transp_1: ABC transporter, permease protein	
	E-value:4.8e-92	E-value: 3.3e-70
Clan	CL0404 BpD_transp_1	
	E-value:3.1e-26	E-value: 4.3e-27
Highly conserved amino acids	G5, A9, P31	G5, P31, G91

(HMM logo)		
PDB protein	3DHW methionine importer MetNI	No PDB structure found
	E-value: 7.04674E-5	

In table 13 below, b2677 and Mrub_1283 are orthologous due to several different structural data from a variety of bioinformatics tools such as CDD, TIGRFAM, and PFAM. The CDD searches for both sequences result in the same COG4175 result. COG4175 is ProV protein, which is consistent with the function of b2677 in literature. Since Mrub_1283 is orthologous to b2677 as shown by several informatics data, Mrub_1283 is expected to encode for the ProV protein, the ATP-binding domain of ABC transporter as well. The TIGRFAM searches of both sequences result in the same TIGR01186 hit, which is the proV: Glycine betaine/L-protein transport A. From TIGRFAM result, we predict that Mrub_1283 encode for the ATP binding domain of the ABC transporter. The PFAM searches of both sequences result in the same PF00005 hit, which is the ATP binding domain of ABC-type transport system. Again, PFAM result is consistent with the predicted function for Mrub_1283 protein. CL0023 – P-loop NTPases clan is found as the first hit of the searches for both sequences. This Clan result also supports our hypothesis that Mrub_1283 is an ortholog of b2677. All the hits are associated with very low E-value which indicates the significance of the data found.

Table 13: b2677 and Mrub_1283 are orthologous according to structural-data from a variety of bioinformatics tools.

Categories	*E. coli* b2677 (*proV*)	*M. ruber* Mrub_1283

TIGRFAM – Protein family	TIGR01186 proV: Glycine betaine/L-protein transport A	
	E-value: 4.2e-278	E-value: 2.3e-137
CDD search	COG 4175 ProV	
	E-value: 2.73e-178	E-value: 2.73e-178
PFAM – Protein family	PF00005 ATP-binding domain of ABC transporter	
Clan	CL0023 P-loop_NTPase	
	E-value:2.1e-34	E-value: 9.8e-31
Highly conserved amino acids (HMM logo)	G18, G21, G23	G18, G21, G23
PDB protein	2D62 Crystal structure of multiple sugar binding transport ATP-binding protein	2IT1 Structure of PH0203 protein from Pyrococcus horikoshii
	E-value: 5.43883E-49	E-value: 2.69755E-52

Section V – Operon Module

This section is to show that Mrub_1285, Mrub_1284 and Mrub_1283 are three components of an operon that all encode for the ABC transporter that transport glycine/betaine. If we can verify these three genes are in the same operon just like *proX*, *proW*, *proV*, this data can also support our hypothesis about orthologue between *M. ruber* and *E. coli* genes.

We see that *proX*, *proW*, *proV* are parts of an operon by the pathway in EcoCyc (Keseler *et al*, 2013), the database with data for *E. coli*. The pathway is represented in figure 5 below.

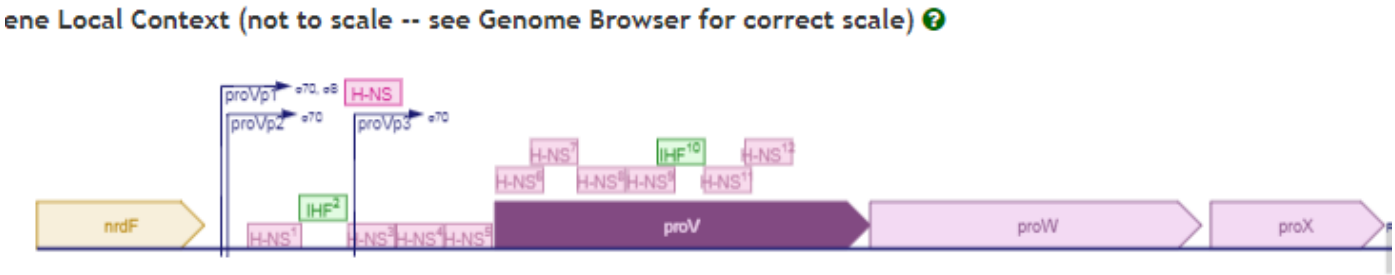


Figure 5. *proX*, *proW* and *proV* are part of an operon due to data from EcoCyc (Keseler *et al*, 2013). In the figure, they are located next to each other and work in a same pathway to encode the ABC transporter.

Data from Color-by-KEGG map from IMG/M also confirm that these three *E. coli* genes are parts of an operon. The map is presented in figure 6.

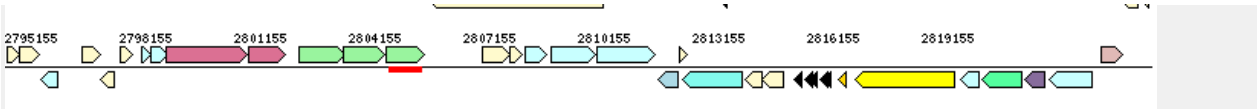


Figure 6. *proX*, *proW* and *proV* are part of an operon due to data from IMG/M Color-by-KEGG map. In the figure, the gene are located next to each other. They have the

same color which indicate the same function so they are very likely to be in an operon and serve in a same pathway.

On the other hand, Mrub_1285, Mrub_1284 and Mrub_1283 are parts of an operon as well. This fact can be confirmed also using the Color-by-KEGG map from IMG/M in figure 7.

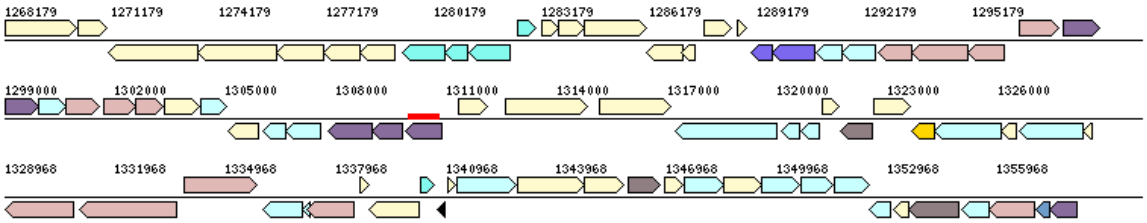


Figure 7. Mrub_1285, Mrub_1284 and Mrub_1283 are parts of an operon by IMG/M Color-by-KEGG. In the picture, we can see that these three genes are located next to each other and have the same color: indicate the same function and in the same pathway as an operon.

Another way to show that these *M. ruber* genes are in an operon is to look at other organisms that are evolutionarily close-related to *M. ruber*. Figure 8 shows the neighborhood regions with the same top COG hit with the GOI from *M. ruber*.

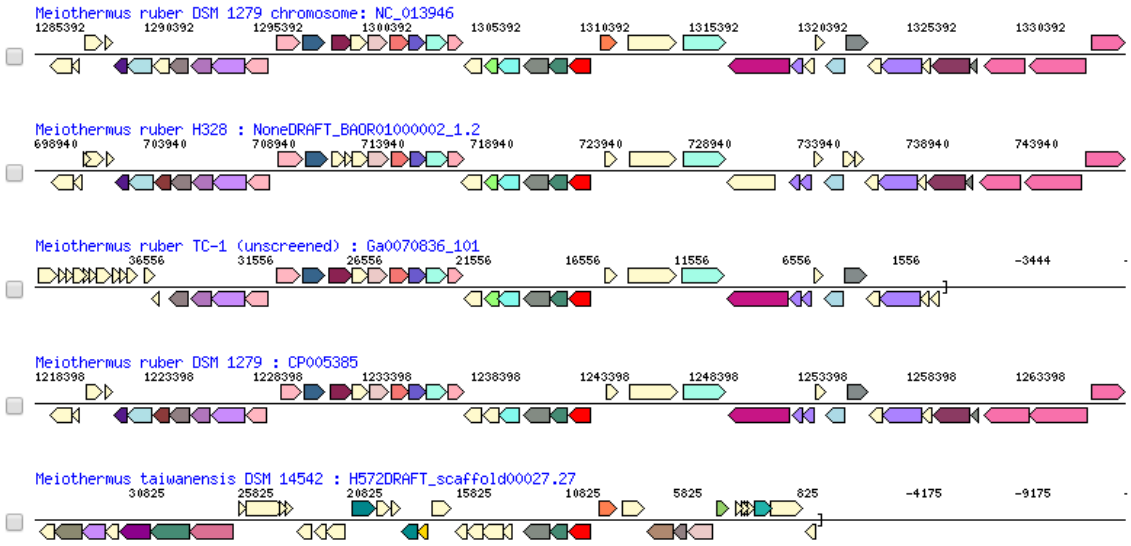


Figure 8. Mrub_1285, Mrub_1284 and Mrub_1283 are parts of an operon by IMG/M. In the picture, we can see that these all the gene with same top COG hits in other organism closely related to *M. ruber* are in an operon. Therefore, it is reasonable to claim that Mrub_1285, Mrub_1284 and Mrub_1283 are parts of an operon as well.

Section VI. Duplication and Degradation Module

In this section, we are looking for duplication of the gene in *M. ruber* and *E. coli* genomes (paralogs). However, for all *E. coli* and *M. ruber* genes, there are no paralogs found in the BLAST of the gene against its own genome. The DB search in KEGG also results in none genes with reasonable E-value. Therefore, we conclude that there are no paralogs of all our *M. ruber* GOIs.

Section VII. Horizontal transfer Module

In this section, we are looking for horizontal gene transfer between phylogenetic-related organisms. However, there are no HGT suspected so we don't have any further analyses in this module.

CONCLUSION

From all the results we have collected, we conclude that Mrub_1285, Mrub_1284 and Mrub_1283 are orthologous to *proX, proW* and *proV* in *E. coli* genomes, respectively. As the result, we predict that Mrub_1285 encodes for the substrate binding domain, Mrub_1284 encodes for the transmembrane domain (permease) and Mrub_1283 encodes for the ATP-binding domain. Furthermore, these three genes are in the same operon codes for the complete structure of glycine betaine ABC transporter.

As mentioned in the introduction, to confirm our hypothesis about the function of *M. ruber* genes, we must show that the *M. ruber* genes of our interests are orthologous to genes in *E. coli* genome. To test our hypothesis, we collect evidence from various sources contained protein family like PFAM, TIGRFAM and conserved domain (CDD) as well as evidence about cell localization of the protein product of each gene from TMHMM, LipoP, SignalP, PSORT-B. The summary of evidence confirming that Mrub_1285 and *proX* (b2679) are presented in table 14,

the evidence confirm that Mrub_1284 and *proW* (b2678) are presented in table 15 while the evidence confirm that Mrub_1283 and *proV* (b2677) are presented in table 16.

In table 14, we see in the first row of data shows the BLAST result of *E. coli* b2679 against *M. ruber* genome. From this result, we predict that there is only one ortholog of b2679 in *M. ruber* genome since there is only one hit with E-value in the acceptable range (<0.01). This BLAST has E-value very close to zero so we are certain that the two sequences do not align due to random chance but rather are orthologous and derive from the same ancestor gene. The CCD and PFAM pulled out the same COG number and PFAM for both sequences, indicate they have similar conserved domain, belong to one protein family that code for substrate binding domain of ABC-type transporter. Unfortunately, the search for TIGRFAM family for both sequences does not find any hits with reasonable E-value. All results are found with very small E-value so the similarity is certainly not due to chance. In term of cell localization, both protein products are showed by PSORT-B to be in the periplasmic space. They are both concluded by TMHMM to have no transmembrane domain. Their protein products are showed to be signal proteins by SignalP and LipoP with high probability. Both proteins contains same amino acids that highly conserved in the HMM logo, which indicated the two genes to have an evolutionary relation. The search in Protein Data Bank (PDB) also results in only one structure with relatively low E-value, indicates that the two genes encode for the same structure or the genes have a similar function. Finally, yet importantly, both of the genes are showed to be in an operon, which emphasize their close evolutionary relation.

Table 14: b2679 and Mrub_1285 are orthologous according to data from a variety of bioinformatics tools.

Categories	*E. coli* b2679 (*proX*)	*M. ruber* Mrub_1285
BLAST *E. coli* b2679 amino acid sequence against *M. ruber*	Score: 65.5 bits E-value: 1e-12	
CDD data	COG2113 ABC-type proline/glycine betaine transport system, periplasmic component	
	Score: 668.57 E-value: 0e+00	Score: 215 E-value: 4.0e-57
Cellular Localization (PSORT-B result)	Periplasmic Space	
TMHMM	0	0
LipoP	SP1	
SignalP	Probability: 0.787 Signal Protein	Probability: 0.642 Signal Protein
PFAM – Protein family	PF04069 Substrate binding domain of ABC-type glycine betaine transport system	
	E-value: 4.3e-63	E-value: 3.3e-70
Clan	CL0177 Periplasmic binding protein clan	
	E-value: 5.4e-56	E-value: 9.6e-60

Highly conserved amino acids (HMM logo)	G28, G49, Y143	G28., G49, Y143
PDB protein	1R9L ProX in complex with glycine betaine	
	E-value: 0.0	E-value: 1.28099e-10
Part of Operon?	Yes	Yes
KEGG pathway map	Map02010: ABC transporters	

In table 15, we see in the first row of data shows the BLAST result of *E. coli* b2678 against *M. ruber* genome. From this result, we predict that there is only one ortholog of b2678; in *M. ruber* genome since there is only one hit with E-value in the acceptable range (<0.01). This BLAST has E-value very close to zero so we are certain that the two sequences do not align due to random chance but rather are orthologous and derive from the same ancestor gene. The TIGRFAM and PFAM pulled out the same TIGRFAM number and PFAM for both sequences, indicate they have a similar conserved domain, belong to one protein family that code for substrate binding domain of ABC-type transporter. Even when TIGRFAM pulled out the same protein for both sequences, TIGR03416 hit, which is the ABC_choXWV_perm: choline ABC transporter family, the TIGRFAM of choline ABC transporter seems unrelated to the function as glycine betaine transporter. However, in bacteria, the primary role of choline is the precursor of glycine/betaine so in most bacteria, the transport mechanism of choline and glycine betaine are similar (Wargo *et al*, 2013). Hence, we can still use the TIGR03416 as evidence to understand the function and structure of the protein encoded by Mrub_1284. Unfortunately, the search for

COG from CDD search for both sequence do not find any hits with reasonable E-value. All results are found with very small E-value so the similarity is certainly not due to chance. In term of cell localization, both protein products are showed by PSORT-B to be in the cytoplasmic membrane. They are both concluded by TMHMM to have 6 transmembrane domains. Their protein products are showed not to be signal proteins by both SignalP and LipoP with high probability. Both proteins contains same amino acids that highly conserved in the HMM logo (G9 and A31), which indicated the two genes to have an evolutionary relation. The search in Protein Data Bank (PDB) cannot be used to conclude about orthology in this situation since the search for *E. coli* gene results in an unrelated protein while there is no hit found for *M. ruber* gene. Finally, yet importantly, both of the genes are showed to be in an operon, which emphasize their close evolutionary relation.

Table 15: b2678 and Mrub_1284 are orthologous according to data from a variety of bioinformatics tools.

Categories	*E. coli* b2678 (*proW*)	*M. ruber* Mrub_1283
BLAST *E. coli* b2678 amino acid sequence against *M. ruber*	Score: 189 bits E-value: 1e-58	
Cellular Localization (PSORT-B result)	Cytoplasmic Membrane	
TMHMM	6	6
SignalP	Probability: 0.091 Not a signal protein	Probability: 0.088 Not a signal protein

LipoP	Cytoplasm	
TIGRFAM – Protein family	TIGR03416 ABC_choXWV_perm: choline ABC transporter	
	E-value: 6.2e-95	E-value: 5e-97
PFAM – Protein family	PF00528 BPD_transp_1: ABC transporter, permease protein	
	E-value:4.8e-92	E-value: 3.3e-70
Clan	CL0404 BpD_transp_1	
	E-value:3.1e-26	E-value: 4.3e-27
Highly conserved amino acids (HMM logo)	G5, A9, P31	G5, P31, G91
PDB protein	3DHW methionine importer MetNI	No PDB structure found
	E-value: 7.04674E-5	
Part of Operon?	Yes	Yes
KEGG pathway map	Map02010: ABC transporters	

In table 16, we see in the first row of data shows the BLAST result of *E. coli* b2677 against *M. ruber* genome. From this result, we predict that there is only one ortholog of b2678; in *M. ruber* genome since there is only one hit with E-value in the acceptable range (<0.01). This

BLAST has E-value very close to zero so we are certain that the two sequences do not align due to random chance but rather are orthologous and derive from the same ancestor gene. The CDD search, TIGRFAM and PFAM pulled out the same COG, TIGRFAM, and PFAM number for both sequences, indicate they have a similar conserved domain, belong to one protein family that code for substrate binding domain of ABC-type transporter. All results are found with very small E-value so the similarity is certainly not due to chance. In term of cell localization, both protein products are showed by PSORT-B to be in the cytoplasmic membrane. They are both concluded by TMHMM to have 0 transmembrane domains means that they are just anchored on the membrane instead of embedded in the membrane. Their protein products are showed not to be signal proteins by both SignalP and LipoP with high probability. Both proteins contains same amino acids that highly conserved in the HMM logo (G18, G21, G23), which indicated the two genes to have an evolutionary relation. The search in Protein Data Bank (PDB) cannot be used to conclude about orthology in this situation since the search for two sequences results in different structures and both structures seem to be unrelated to our protein as ATP-binding domain of ABC transporter. We need further data regard of the crystal structure/ structure of the protein products of Mrub_1285 and b2677 to conclude about their structural similarity. Finally, yet importantly, both of the genes are showed to be in an operon and encode for the complete structure of glycine betaine ABC transporter. This data from operon module emphasizes the close evolutionary relation between *E. coli* genes and our *M. ruber* genes of interest.

Table 16: b2677 and Mrub_1283 are orthologous according to data from a variety of bioinformatics tools.

Categories	*E. coli* b2677 (*proV*)	*M. ruber* Mrub_1283
BLAST *E. coli* b2677 amino acid sequence against *M. ruber*	Score: 342 bits E-value: 3e-116	
Cellular Localization (PSORT-B result)	Cytoplasmic Membrane	
TMHMM	0	0
SignalP	Probability: 0.094 Not a signal protein	Probability: 0.121 Not a signal protein
LipoP	Cytoplasm	
TIGRFAM – Protein family	TIGR01186 proV: Glycine betaine/L-protein transport A	
	E-value: 4.2e-278	E-value: 2.3e-137
CDD search	COG 4175 ProV	
	E-value: 2.73e-178	E-value: 2.73e-178
PFAM – Protein family	PF00005 ATP-binding domain of ABC transporter	
Clan	CL0023 P-loop_NTPase	
	E-value:2.1e-34	E-value: 9.8e-31

Highly conserved amino acids (HMM logo)	G18, G21, G23	G18, G21, G23
PDB protein	2D62 Crystal structure of multiple sugar binding transport ATP-binding protein	2IT1 Structure of PH0203 protein from Pyrococcus horikoshii
	E-value: 5.43883E-49	E-value: 2.69755E-52
Part of Operon?	Yes	Yes
KEGG pathway map	Map02010: ABC transporters	

References:

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." J. Mol. Biol. 215:403-410. PubMed

Balakrishnan L, Venter H, Shilling RA, van Veen, Hendrik W: Reversible transport by the ATP-binding cassette multidrug export pump LmrA: ATP synthesis at the expense of downhill ethidium uptake. The Journal of biological chemistry 2004, 279:11273-80.

Berman, H.M., J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne. (2000) The Protein Data Bank *Nucleic Acids Research*, 28: 235-242.

Betts, M. J., and Russell, R. B. 2003. Amino-Acid Properties and Consequences of Substitutions. Bioinformatics for Geneticists. 2003: 311–342.

Biolabs, N. E. Home - NEB | New England Biolabs. Home - NEB | New England Biolabs. https://www.neb.com/

Blattner FR, Plunkett G, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y: The complete genome sequence of Escherichia coli K-12. Science (New York, N.Y.) 1997, 277:1453-62.

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/.

Crooks, GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: A sequence logo generator, Genome Research, 14:1188-1190, (2004)

Euzéby JP. List of bacterial names with standing in nomenclature: A folder available on the Internet. Int J Syst Bacteriol 1997; 47:590-592. PubMed doi:10.1099/00207713-47-2-590

Finn RD, Bateman A, Clements J, et al. Pfam: the protein families database. Nucleic Acids Research. 2014;42 (Database issue):D222-D230. doi:10.1093/nar/gkt1223.

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.

Higgins CF, Hiles ID, Salmond GP, Gill DR, Downie JA, Evans IJ, Holland IB, Gray L, Buckel SD, Bell AW: A family of related ATPbinding subunits coupled to many distinct biological processes in bacteria. Nature 1986, 323:448-50.

Higgins CF: ABC transporters: from microorganisms to man. Annual review of cell biology 1992, 8:67-113

Holland IB, Blight MA: ABC-ATPases, adaptable energy generators fuelling transmembrane movement of a variety of molecules in organisms from bacteria to humans. Journal of molecular biology 1999, 293:381-99.

Jardetzky O: Simple allosteric model for membrane pumps. Nature 1966, 211:969-70.

Juncker, A., H. Willenbrock, G. von Heijne, H. Nielsen, S. Brunak and A. Krogh. 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 Lukas Käll, Anders Krogh and Erik L. L. Sonnhammer. A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology*, 338(5):1027-1036, May 2004. (doi) (PubMed)

Käll,L., Anders Krogh and Erik L. L. Sonnhammer. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res.*, 35:W429-32, July 2007 (doi) (PubMed)

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. Genera: fusing model organism databases with systems biology Nucleic Acids Research 41:D605-612.

Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. Journal of Molecular Biology, 305(3):567-580, January 2001. (PDF, 959503 bytes)

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

Lewinson O, Lee AT, Locher KP, Rees DC: A distinct mechanism for the ABC transporter BtuCD-BtuF revealed by the dynamics of complex formation. Nature structural & molecular biology 2010, 17:332-8

Li J, Jaimes KF, Aller SG: Refined structures of mouse Pglycoprotein. Protein science: a publication of the Protein Society 2014, 23:34-46.

Loginova LG, Egorova LA. Obligate thermophilicbacterium Thermus ruber in hot springs of Kamchatka. Mikrobiologiya 1975; 44:661-665.

Loginova LG, Egorova LA, Golovacheva RS, Seregina LM. Thermus ruber sp. nov., nom. rev. Int J Syst Bacteriol 1984; 34:498-499. doi:10.1099/00207713-34-4-498

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/ BLAST tool: BLASTp tool from https://BLAST.ncbi.nlm.nih.gov/BLAST.cgi

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. 2015 Jan 28;43(Database issue):D222-2. doi: 10.1093/nar/gku1221. Epub 2014 Nov 20. [PubMed PMID: 25414356]

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

Nobre MF, Trüper HG, Da Costa MS. Transfer of Thermus ruber (Loginova et al. 1984), Thermus silvanus (Tenreiro et al. 1995), and Thermus chliarophilus (Tenreiro et al. 1995) to Meiothermus gen. nov. as Meiothermus ruber comb. nov., Meiothermus silvanus comb. nov., and Meiothermus chliarophilus comb. nov., respectively, and emendation of the genus Thermus. Int J Syst Bacteriol 1996; 46:604-606. doi:10.1099/00207713- 46-2-604

Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. Journal of molecular biology. 2000.

Oldham ML, Chen J: Snapshots of the maltose transporter during ATP hydrolysis. Proceedings of the National Academy of Sciences of the United States of America 2011, 108:15152-6.

Paula S, Volkov AG, Van Hoek AN, Haines TH, Deamer DW: Permeation of protons, potassium ions, and small polar molecules through phospholipid bilayers as a function of membrane thickness. Biophysical journal 1996, 70:339-48

Petersen, Thomas, 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

Pires AL, Albuquerque L, Tiago I, Nobre MF, Empadinhas N, Veríssimo A, da Costa MS. Meiothermus timidus sp. nov., a new slightly thermophilic yellow-pigmented species. FEMS Microbiol Lett 2005; 245:39-45. PubMed doi:10.1016/j.femsle.2005.02.011

Saier MH, Reddy VS, Tamang DG, Västermark A: The transporter classification database. Nucleic acids research 2014, 42:D251-8.

Skerman VBD, McGowan V, Sneath PHA. Approved Lists of Bacterial Names. Int J Syst Bacteriol 1980; 30:225-420. doi:10.1099/00207713- 30-1-225

Sonnhammer, ELL., G. von Heijne, and A. Krogh.A hidden Markov model for predicting transmembrane helices in protein sequences. In J. Glasgow, T. Littlejohn, F. Major, R. Lathrop, D. Sankoff, and C. Sensen, editors, Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology, pages 175-182, Menlo Park, CA, 1998. AAAI Press.

Tindall, B. J., Sikorski, J., Lucas, S., Goltsman, E., Copeland, A., Glavina Del Rio, T., … Lapidus, A. (2010). Complete genome sequence of Meiothermus ruber type strain (21T). Standards in Genomic Sciences, 3(1), 26–36. http://doi.org/10.4056/sigs.1032748

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

Zhang XQ, Zhang WJ, Wei BP, Xu XW, Zhu XF, Wu M. Meiothermus cateniformans sp. nov., a slightly thermophilic species from north-eastern China. Int J Syst Evol Microbiol 2010; 60:840-844. PubMed doi:10.1099/ijs.0.007914-0

Wargo, M. 2013. Homeostasis and Catabolism of Choline and Glycine Betaine: Lessons from *Pseudomonas aeruginosa*. Applied and Environmental Microbiology, 79(7), 2112-2120