2018

Confirmation that mrub_1751 is homologous to *E. coli* xylF, mrub_1752 is homologous to *E. coli* xylH, and mrub_1753 is homologous to *E. coli* xylG

Ben Price
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Confirmation that mrub_1751 is homologous to *E. coli* xylF, mrub_1752 is homologous to *E. coli* xylH, and mrub_1753 is homologous to *E. coli* xylG.

Introduction

It had been hypothesized that the ABC transporters that function in *E. coli* have homologs in a variety of other microbes. By studying the transporters in *E. coli,* we should be able to recognize the corresponding transporters in other genomes. These are important to locate because of the many substrates that they carry, ranging from the uptake of special sugars or antibiotic protection, secreting toxins, or even, in the case of some mutations, causing some genetic disorders (Moussatova *et. al.* 2008). For this particular study we are focusing on the D-xylose transporters, found in *E. coli,* and the proposed homologs in *Meiothermus ruber*, a Gram-negative bacterium discovered in hot springs ranging across northern Europe and Asia (Tindall *et. al.* 2010). This bacterium is being used as our study organism because it has not been well studied, though a majority of its genes (71.8%) have been given putative functions, which gives some idea of what each gene is most likely to be (Tindall *et. al.* 2010).

The genes in question that we are studying are the putative D-xylose transport genes of *M. ruber,* Mrub_1751, Mrub_1752, and Mrub_1753. The D-xylose system is important for the cell because it is used in the pentose phosphate pathway to produce D-xylulose 5-phosphate (Song and Park 1997). The corresponding genes, in order are xylF, xylH, and xylG in *E. coli,* and are part of an operon, as also discussed in Song and Park (1997). The transporter made from these three genes is embedded in the cell

membrane, with XylF filling the role of the solute binding protein, XylH the trans-membrane domain, and XylG being the nucleotide/ATP binding protein.

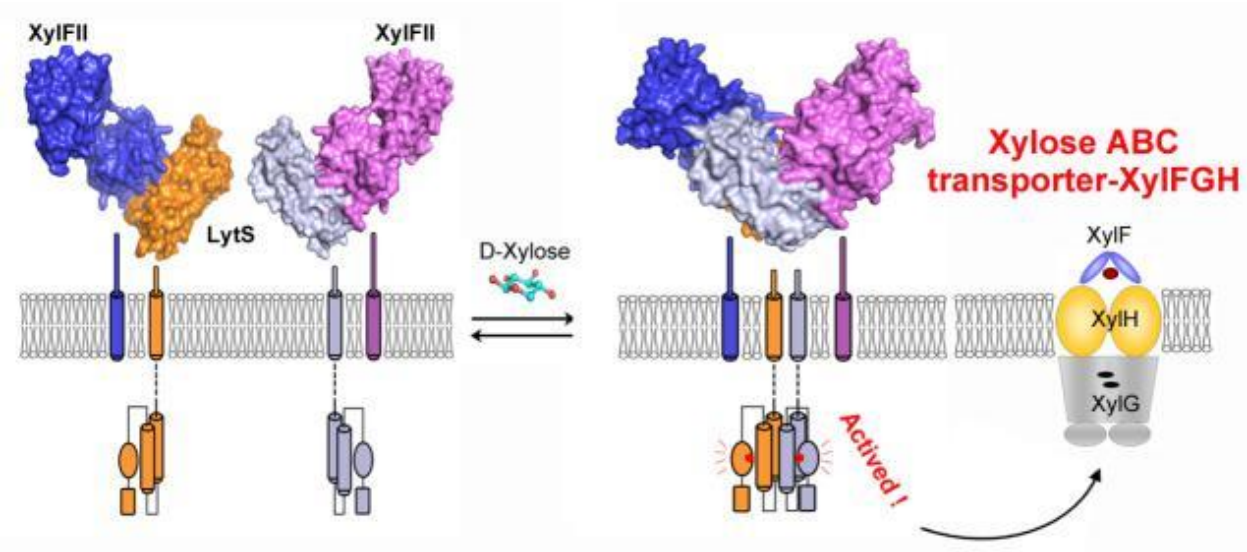

Figure 1. Activation and embedding of Xylose ABC transporter in cell membrane. Activation is initiated when a complex of the membrane-associated sensor protein XylFII and a transmembrane histidine kinase LytS senses d-xylose in the environment, activating the ABC transporter. Taken from Jia (2017).

Via Blast, it was noted that each gene had paralogs within the *M. ruber* genome, with Mrub_1751 only having 2 significant hits, Mrub_1752 having 6 significant hits, and Mrub_1753 having many, many hits (Altschul *et. al.* 1990). This study seeks to prove that Mrub_1751 is a homolog to *E. coli* xylF (b_3566), Mrub_1752 is homologous to *E. coli* xylH (b_3568), and Mrub_1753 is homologous to *E. coli* xylG (b_3567).

Materials and Methods

This study was completed by utilizing the *M. ruber* genome stored on KEGG (Kanehisa *et. al.* 2016). The gene in both nucleotide and amino acid chain form were obtained, which were then BLASTed (Altschul *et. al.* 1990) against the *E. coli* genome to discover likely homologs. The same process was repeated to discover homologs from other species, and fifteen homologous genes were selected. The results were

aligned by T-Coffee (Notredame *et. al.* 2000) and a Weblogo (Crooks *et. al.* 2004) was created to find highly conserved residues. This was followed up with TMHMM (Krogh *et. al.* 2001, Krogh and Rapacki 2016, Sonnhammer *et. al.* 1998) to determine the number of transmembrane helices, SignalP (Petersen *et. al.* 2012) to determine the probability of having a signal peptide, LipoP (Juncker *et. al.* 2003) to determine the most likely ending point of the protein if it had a signal peptide, PsortB (Yu *et. al.* 2010) to determine the most likely ending location of the protein, and Phobius for the same (Kall *et. al.* 2004, Kall *et. al.* 2007). The gene was then BLASTed again to determine the CDD (Marchler-Bauer *et. al.* 2015), TIGRFAM (Haft *et. al.* 2001), and PFAM (Finn *et. al.* 2014, Fin *et. al.* 2016) groups. The amino acid sequence was entered into the Protein Data base (PDB) (Berman *et. al.* 2000) to determine the closest crystallized known protein. Finally, IMG/M (Markowitz *et. al.* 2012) was consulted to determine the likelihood that the protein was part of an operon.

Results

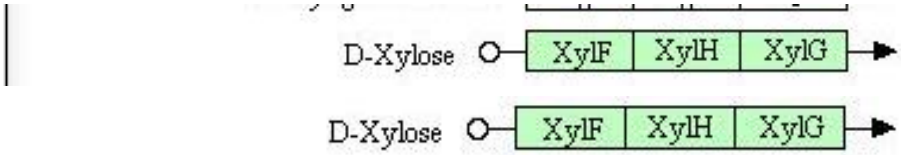

Figure 2. KEGG pathways monosaccharide ABC transporters for *M. ruber* (Top) and *E. coli* (Bottom). Note that both include the full D-xylose pathway.

The KEGG results showed that both organisms included the D-xylose transporters, the first piece of evidence that the genes may be homologs. This was followed by the BLAST of each gene, pictured below.

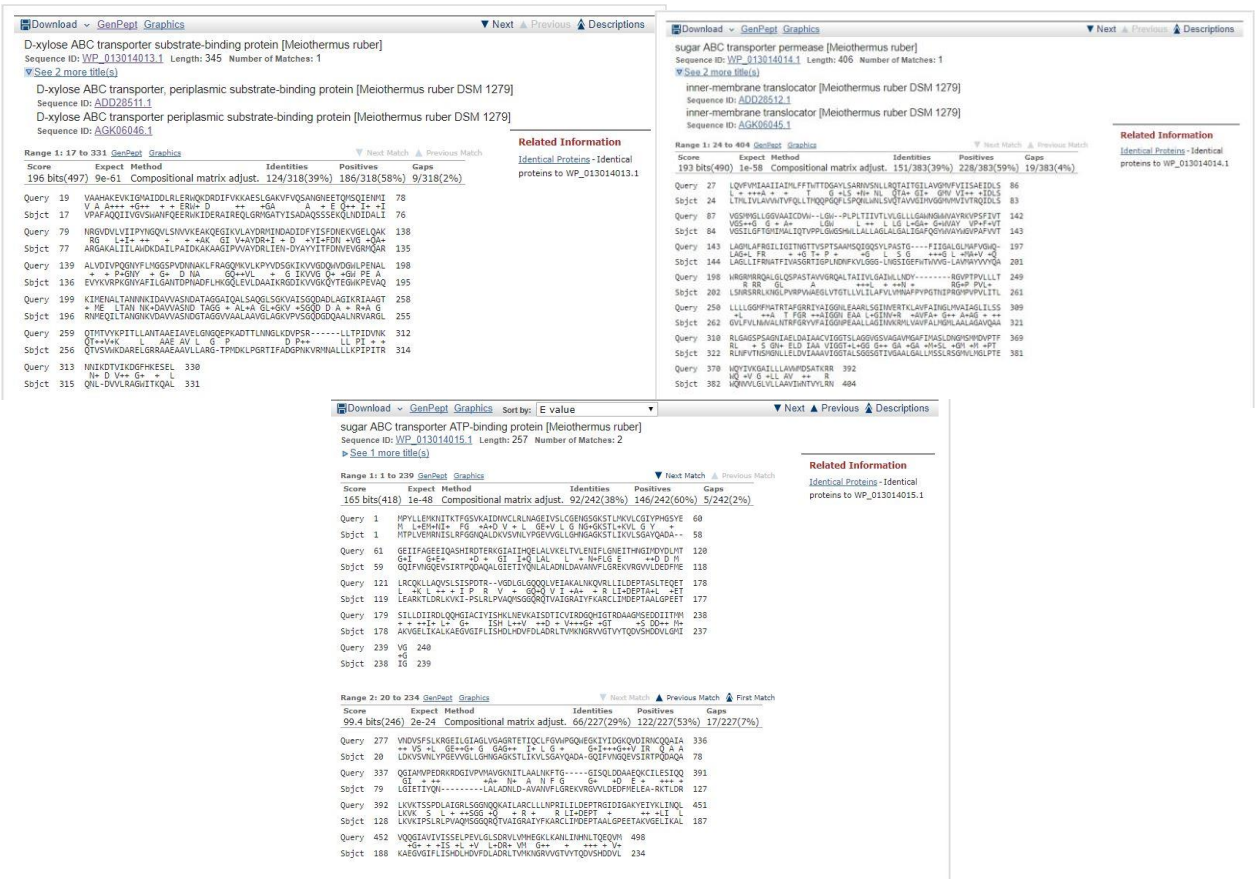

Figure 3. BLASTs of the *E. coli* amino acids chains against *M. ruber* genome. Top left being xylF and Mrub_1751, top right being xylH and Mrub_1752, and bottom being xylG and Mrub_1753.

The e-values for each of these BLASTs was, in order, 9e-61, 1e-58, and split gene at 1e-48 and 2e-24 for each section. Tables for each of the matchups below show the similarities between the genes.

Table 1. b_3566 and Mrub_1751

	B_3566	Mrub_1751
BLAST	9e-61	
CDD	COG4213 ABC-type xylose transport system, periplasmic component	COG4213 ABC-type xylose transport system, periplasmic component
Localization	Periplasmic	Periplasmic
TIGRFAM	TIGR02634 D-xylose ABC transporter, substrate binding protein	TIGR02634 D-xylose ABC transporter, substrate binding protein
PFAM	PF13407 Periplasmic binding domain	PF13407 Periplasmic binding domain
PDB	3M9W Open ligand free crystal structure of Xylose binding protein from *Escherichia coli*	3M9W Open ligand free crystal structure of Xylose binding protein from *Escherichia coli*

Table 2. b_3568 and Mrub_1752

	B_3568	Mrub_1752
BLAST	1e-58	
CDD	COG4214 ABC-type xylose transport system, permease component	COG4214 ABC-type xylose transport system, permease component
Localization	Cell Membrane	Cell Membrane
TIGRFAM	No matches	No matches
PFAM	PF02653 Branched-chain amino acid transport system/permease component	PF02653 Branched-chain amino acid transport system/permease component
PDB	No matches	No matches

Table 3. b_3567 and Mrub_1753

	B_3567	Mrub_1753
BLAST	1e-48 2e-24	
CDD	COG1129 ABC-type sugar transport system, ATPase component	COG 1129 ABC-type sugar transport system, ATPase component
Localization	Cytoplasmic Membrane	Cytoplasmic Membrane
TIGRFAM	TIGR02633 D-xylose ABC transporter, ATP-binding protein	TIGR02633 D-xylose ABC transporter, ATP-binding protein
PFAM	PF00005 ABC transporter	PF00005 ABC transporter
PDB	1G9X An ATP binding cassette of an ABC transporter	1G9X An ATP binding cassette of an ABC transporter

Based on the information included in these tables, it appears that we have genes that are homologous here. A point of interest related to Table 3 is that there were two hits for Mrub_1753 within the same gene when BLASTed against the *E. coli* genome. This is likely because *E. coli* xylG is a fused gene, containing two ATP-binding domains (Keseler *et. al.* 2013), which indicates that *M. ruber* may require two of Mrub_1753 to be translated to yield one complete protein complex, as opposed to the one likely required by *E. coli*. One additional point of interest within these tables is that b_3567 and Mrub_1753 are listed as predominantly in the cytoplasmic membrane. This is different from other ABC transporters, which are in the cytoplasm, which means that part of the protein may be partially embedded in the cell membrane. The PSORTb results that gave this answer are shown below.

PSORTb Results (Click here for an ex

```
SeqID: Mrub_1753 MTPLVEMRNISLRFGGNQALDKVSVNLY
  Analysis Report:
    CMSVM-            Unknown
    CytoSVM-          Cytoplasmic
    ECSVM-            Unknown
    ModHMM-           Unknown
    Motif-            Unknown
    OMPMotif-         Unknown
    OMSVM-            Unknown
    PPSVM-            Unknown
    Profile-          Unknown
    SCL-BLAST-        CytoplasmicMembrane
    SCL-BLASTe-       Unknown
    Signal-           Unknown
  Localization Scores:
    Cytoplasmic            2.11
    CytoplasmicMembrane    7.88
    Periplasmic            0.00
    OuterMembrane          0.00
    Extracellular          0.00
  Final Prediction:
    CytoplasmicMembrane    7.88
---------------------------------------------
```

PSORTb Results (Click here for an ex

```
SeqID: b_3567 MPYLLEMKNI TKTFGSVKAI DNVCLRLNA
  Analysis Report:
    CMSVM-            Unknown
    CytoSVM-          Cytoplasmic
    ECSVM-            Unknown
    ModHMM-           Unknown
    Motif-            Unknown
    OMPMotif-         Unknown
    OMSVM-            Unknown
    PPSVM-            Unknown
    Profile-          Unknown
    SCL-BLAST-        CytoplasmicMembrane
    SCL-BLASTe-       Unknown
    Signal-           Unknown
  Localization Scores:
    Cytoplasmic            2.11
    CytoplasmicMembrane    7.88
    Periplasmic            0.00
    OuterMembrane          0.00
    Extracellular          0.00
  Final Prediction:
    CytoplasmicMembrane    7.88
---------------------------------------------
```

Figure 4. PSORTb results for b_3567 and Mrub_1753. Contrary to expectations, the proteins appear to be majority embedded in the membrane, rather than in the cytoplasm. From Yu *et. al.* (2010).

Additionally, both sets of genes show evidence of being operons based on the position of the genes in relation to each other in the genome, as well as the fact that *E. coli* is confirmed to utilize the D-xylose operon. Images of the genomes around these genes are shown below.

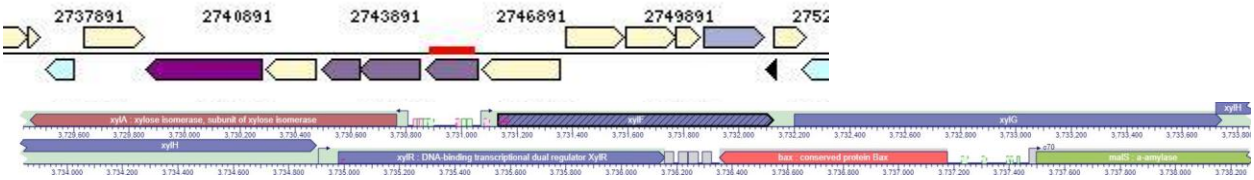

Figure 5. Genomes in locations of genes of interest in *M. ruber* (top) and *E. coli* (bottom) (taken from Keseler *et. al.* 2013). The genome for *E. coli* wraps around to the second line in the middle of the xylH gene, but it can still be seen that the genes are very close together, and even overlapping in *E. coli,* with an overlap of 22 nucleotides, as xylG ends at nucleotide position 3733742, and xylH starts at nucleotide position 3733720.

Additional research was done to examine whether the start codons for each of the genes examined were correct, yielding the following figure 6. This figure shows that for each gene it appears that the start codon has been correctly identified, as there was no evidence otherwise. The upstream region of each gene shows no likely replacements for the start codon, while the Weblogos show agreement in the starting amino acid. Of note is that one of the Weblogos for these genes has a large gap in the first row, due to one genome having a larger gene focused towards the beginning.

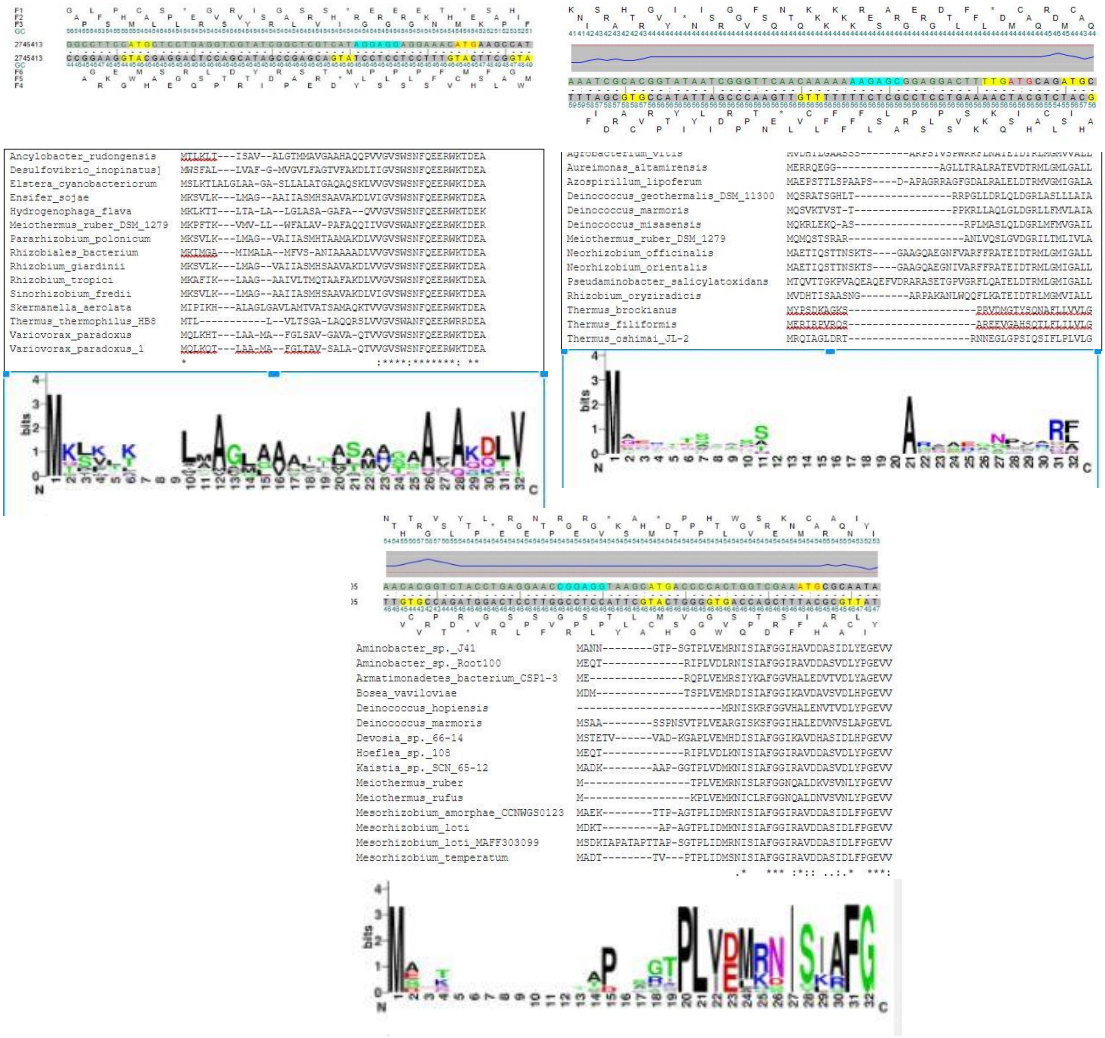

Figure 6. Start codon analysis of *M. ruber* genes. There is no evidence that would indicate incorrect start codon placement.

Finally, there was evidence that there were paralogs for each of the *M. ruber* genes researched, as shown below in figure 7. Of note is the relatively low number of paralogs for Mrub_1751 and Mrub_1752 compared to Mrub_1753.

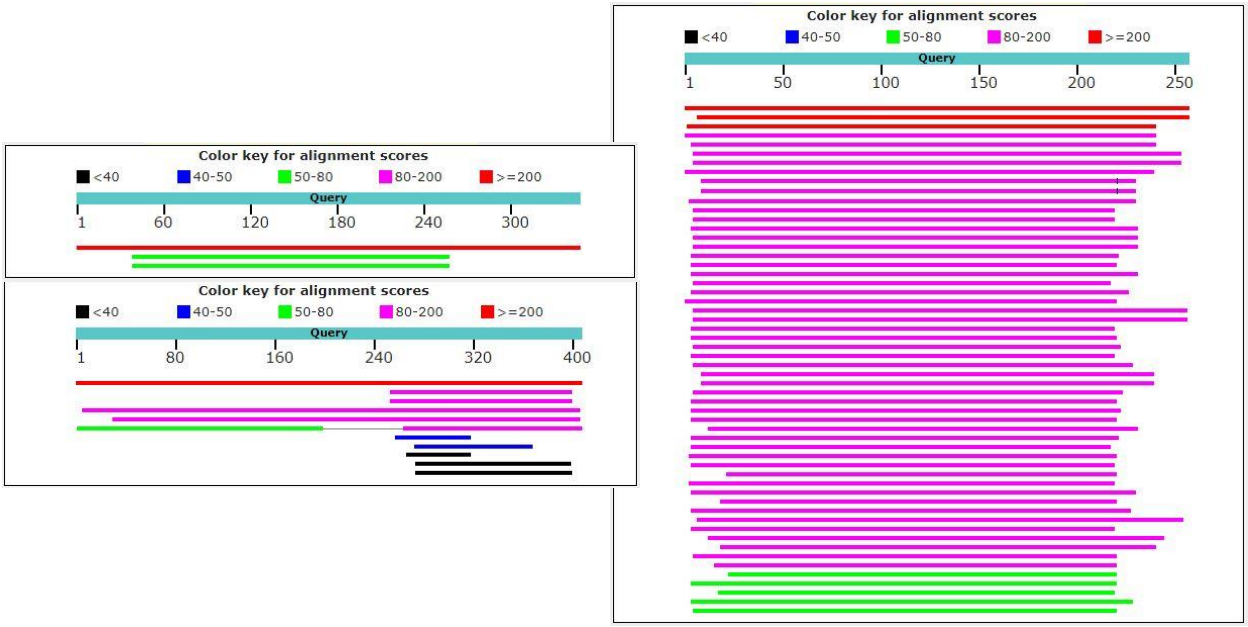

Figure 7. BLASTs of genes of interest against *M. ruber* genome. Displays paralogs of Mrub_1751 in the top left, Mrub_1752 in the bottom left, and Mrub_1753 on the right. Of note is the large number of paralogs for Mrub_1753, indicating a basic structure.

Conclusions

In conclusion, it would appear that our hypothesis is correct, as it appears that Mrub_1751 is homologous to b_3566, Mrub_1752 is homologous to b_3568, and Mrub_1753 is homologous to b_3567. Due to the low value of the e-values, which indicated the likelihood that two genes are similar due to random chance, with higher values indicating high likelihood of random chance, as well as the presence of paralogs within the genome that were similar to the *M. ruber* genes, but different enough from the *E. coli* genes, indicates these genes are orthologs. Additionally, both sets of genes are part of an operon, lending weight to the conclusion that they are orthologs.

Additionally, for Mrub_1751, the shared COG hit COG4213: ABC-type xylose transport system, periplasmic component, PFAM hit PF13407: Periplasmic binding domain, and TIGRFAM hit TIGR02634: D-xylose ABC transporter, substrate binding protein, shared with b_3566 all indicate that the genes are likely paralogs.

For Mrub_1752, the shared COG hit COG4214: ABC-type xylose transport system, permease component, PFAM hit PF02653: Branched-chain amino acid transport system/permease component, and lack of a TIGRFAM hit with b_3568 all indicate that the genes are likely paralogs.

For Mrub_1753, the shared COG hit COG1129: ABC-type sugar transport system, ATPase component, PFAM hit PF00005: ABC transporter, and TIGRFAM hit TIGR02633: D-xylose ABC transporter, ATP-binding protein with b_3567 all indicate that the genes are likely paralogs.

Finally, for a future study, I will propose the change of the proline at position 106 to an alanine in Mrub_1751 to determine whether that will change the functionality of the protein. This will likely cause some sort of conformational change as alanine is a very small amino acid, with an r-group composed of a single carbon, whereas proline is a complicated amino acid that connects the r-group to the n-side of the amino acid (Betts and Russel 2003). The figure detailing how we would do that is below.

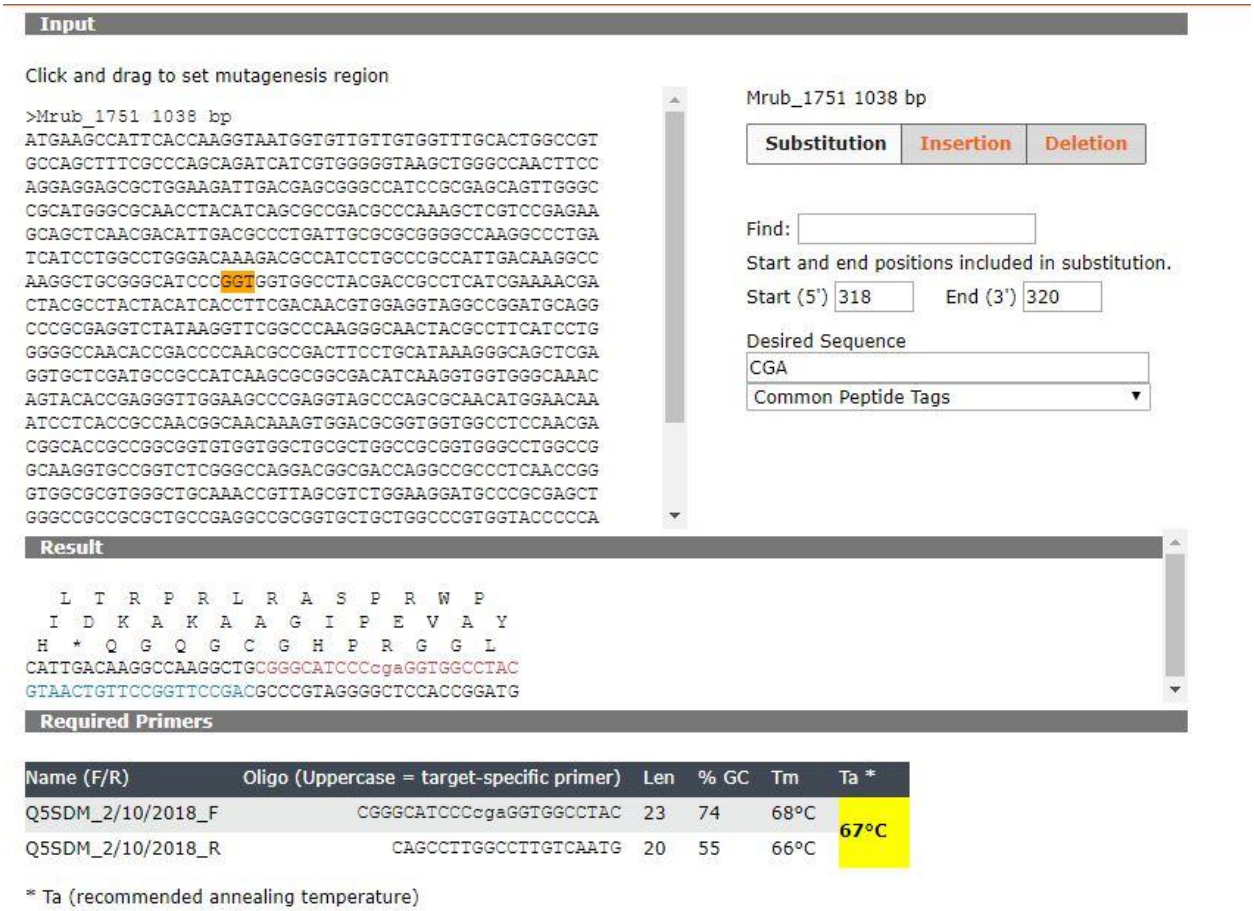

Figure 8. Changing of the highlighted GGT to a CGA would yield the change from a proline to an alanine, something that will change the conformation of the protein.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403-410.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. [Internet]. The Protein Data Bank; [cited 2018 Feb 6]. Available from: http://www.rcsb.org/.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank Nucleic Acids Research, 28: 235-242.

Betts MJ and Russell RB. 2003. Amino-Acid Properties and Consequences of Substitutions. Bioinformatics for Geneticists. 311–342.

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

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

Finn RD, Bateman A, Clements J, et al. 2014. Pfam: the protein families database. Nucleic Acids Research [Internet]. [cited 2018 Feb 6]. 42 (Database issue):D222-D230. Available from http://pfam.xfam.org/

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The Pfam protein families database: towards a more sustainable future: Nucleic Acids Res. [Internet]. [cited 2018 Feb 6] 44:D279-D285. 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.

Jia, L. 2017. Scientists Reveal the Mechanism of Environmental D-xylose Perception in Bacterial [Internet]. Chinese Academy of Sciences; [cited 2018 Feb 9]. Available from http://english.cas.cn/newsroom/research_news/201707/t20170725_181255.shtml.

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

Kall L, Krough A, Sonnhammer E. 2004. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 338(5):1027-1036.

Kall L, Krogh A, Erik LL. Sonnhammer E. 2007. Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. Nucleic Acids Res. 35:W429-32.

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

Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martinez C, Fulcher C, Huerta AM, 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 RP, Paulsen I, Karp PD. 2013. EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Res. 41:D605-612.

Krogh A, Larsson B, von Heijne G, Sonnhammer E. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J Mol Biol, 305(3):567-580.

Krogh A, Rapacki K. 2016. TMHMM Server, v. 2.0. Cbs.dtu.dk. [Internet]. Denmark: Technical University of Denmark. [cited 2018 Feb 6]. Available from http://www.cbs.dtu.dk/services/TMHMM/

Madden T. 2002 [Updated 2003 Aug 13]. The BLAST Sequence Analysis Tool [Internet] In: McEntyre J, Ostell J, editors. The NCBI Handbook [Internet]. Bethesda (MD): National Center for Biotechnology Information. 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. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res. [Internet]. [cited 2018 Feb 6] 43(Database issue):D222-2. Available from https://www.ncbi.nlm.nih.gov/pubmed/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 Res. 40(D1):D115-22. Available from: http://nar.oxfordjournals.org/content/40/D1/D115.full.

Moussatove A, Kandt C, O'Mara ML, Tieleman P. 2008. ATP-binding cassette transporters in *Escherichia coli*. Biochim & Biophys Acta. [cited 2018 Feb 8].;1778:1757-1771.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. [Internet]. [cited 2018 Feb 8];302(1):201-17. Available from http://www.tcoffee.org/Publications/Pdf/tcoffee.pdf.

Petersen T, Brunak S, von Heijne G, Nielsen H. 2011. Discriminating signal peptides from transmembrane regions. Nat Methods, 8:785-786. Available from: http://www.cbs.dtu.dk/services/SignalP.

Song, S, Park, C. 1997. Organization and Regulation of the D-Xylose Operons in *Escherichia coli* K-12: XylR Acts as a Transcriptional Activator. J Bacteriol, 179(22):7025-32. Available from http://jb.asm.org/content/179/22/7025.full.pdf+html

Sonnhammer E, von Heijne G, Krogh A. 1998. 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. Menlo Park, CA: AAAI Press. p. 175-182.

Tindall BJ, Sikorski J, Lucas S, Goltsman E, Copeland A, Del Rio TG, Nolan M, Tice

H, Cheng JF, Han C, Pitluck S, Liolios K, Ivanova N, Mavromatis K, Ovchinnikova G, Pati A, Fähnrich R, Goodwin L, Chen A, Palaniappan K, Land M, Hauser L, Chang YJ, Jeffries CD, Rohde M, Göker M, Woyke T, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk HP, Lapidus A. 2010. Complete genome sequence of Meiothermus ruber type strain (21T). Standards in Genomic Sciences, 3(1):26–36.

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