2-2016

# Comparison of genes in *Meiothermus ruber* and *Escherichia coli* in the Thiamine Biosynthesis Pathway

Erin E. Frye
*Augustana College, Rock Island Illinois*

Dr. Lori Scott
*Augustana College, Rock Island Illinois*

Erin Frye

Professor Scott

BIOL-375

2/6/16

Comparison of genes in *Meiothermus ruber* and *Escherichia coli* in the Thiamine Biosynthesis Pathway

**1. Background/Introduction**

It is important to study genes in the *M. ruber* because knowing more about this organism will allow us to better understand the roles that genes play in the organism and the mechanisms by which the organism thrives. In this case the genes of interest are specifically the thiD and thiE genes. The use of a positive control which was already known to have both the thiD and thiE genes as determined by functional evidence allows us to determine whether *M. ruber* has the same genes with the same function. This positive control was the b1203 and b3993 loci of *Escherichia coli* str. K-12 substr. MG1655. *Escherichia coli* was chosen as the positive control organism because of the vast amount of research already done on E. *coli* and knowledge already known about the organism, including the fact that it has the thiD and ThiE genes. In determining whether the Mrub_2046 locus and Mrub_2041 locus are in fact the thiD and thiE gene respectively it is necessary to use molecular genetics techniques and to conduct research that will lead to the conclusion of this paper.

The thiD gene (b2103) has been studied in the model organism (*E. coli*). The name of the protein/enzyme encoded for by the thiD gene is hydroxymethylpyrimidine kinase (also known as phosphohydroxymethylpyrimidine kinase). The protein includes has 266 amino acids. The function of this protein in the cell is to synthesize thiamine phosphate.The reaction it catalyses is

as follows: ATP + 4-amino-2-methyl-5-pyimidinemethanol→ ADP+ 4-amino-2-methyl-5phosphomethylpyrimidine + $H^{+}$ [1] . The pathway that hydroxymethylpyrimidine kinase is a part of is the thiamine biosynthesis pathway, identification number: eco00730 (KEGG). The cellular location of the protein is in the cytosol. The gene is part of an operon in *E. coli*. Functional evidence for the thiD gene can be seen in several studies including one where cloning and characterization were done on the thiD gene in *E. coli* cells and enzyme activity of the protein showed functionality [2] .

The thiE gene (b3993) has been studied in the model organism (*E. coli*). The name of the protein/enzyme encoded for by the thiE gene is thiamine phosphate synthase. One feature of the protein includes having 211 amino acids. The function of this protein in the cell is to synthesize thiamine. The protein does this by taking the compounds 4-methyl-5-(B-hydroxyethyl)thiazole phosphate and 4-amino-5-hydroxymethyl-2-methylpyrimidine-pyrophosphate and combining them in order to create thiamine phosphate [3]. The reaction it catalyses is as follows: 4-methyl-5 -(B-hydroxyethyl)thiazole phosphate + 4-amino-5-hydroxymethyl-2-methylpyrimidine -pyrophosphate +2H+ → thiamine phosphate +CO2 + diphosphate. The pathway that thiamine phosphate synthase is a part of is the thiamine biosynthesis pathway (KEGG). The cellular location of the protein is in the cytosol. The gene is part of an operon in *E. coli*. Functional evidence for the gene that encodes for thiamine phosphate synthase can be seen in several studies including one where complementation analysis and DNA sequencing were done in *E. coli* cells on genes responsible for thiamine synthesis, including thiE [4].

Bioinformatics is the study of biological data. More specifically, bioinformatics encompasses the collection of biological data, interpretation of that data, and comparison and analyzation of the data collected [5]. Bioinformatics relies heavily on computational techniques/

algorithms and database maintenance. Without the field of bioinformatics we would probably not have expansive knowledge of genome sequencing and would not be able to predict structure and/or function of various genes in the genome among many other things. Bioinformatics and bioinformatics based databases are integral in our analyses of whether the Mrub_2046 and Mrub_2041 loci are in fact the thiD and thiE gene through comparison of the loci to the b1203 and b3993 loci of *E. coli*.

Ultimately, if we compare Mrub_2046 and Mrub_2041 loci of *Meiothermus ruber* DSM 1279 to the b1203 and b3993 loci of *Escherichia coli* str. K-12 substr. MG1655, which we know correspond to the thiD and thiE genes, use bioinformatics tools, and if the genes are determined to be similar in genetic makeup, then the Mrub_2046 and Mrub_2041 loci of the *M. ruber* DSM 1279 must also contain the thiD and thiE gene, respectively.

**2. Methods**

The platform in this study was the Guiding Education through Novel Investigation – Annotation Collaboration Toolkit (GENI-ACT) site, specifically, the site set up by Dr. Lori Scott for her class BIOL 375 (Molecular Genetics) for the winter term (2015-16) with the intent of comparing *M. ruber* genes with *E. coli* genes [6]. Within each gene assignment category there were several modules to be completed. By completing these modules, comparison and determination of the *M. ruber* genes as the respective thiD and thiE genes could occur. Several modules of importance are listed in the table, along with the website of the bioinformatics program. The Sequence-based Similarity Data module will tell us information pertaining to the similarity of the sequences of *M. ruber* and *E. coli*. The Cellular Localization Data will tell us where in the cell the gene products are found. The Structure Based Evidence module will tell us the protein family corresponding to the four loci (*M. ruber* and *E. coli*). The Enzymatic Function

module contains pertinent information on the pathway and enzyme commission number. The modules within the GENI-ACT site were all completed using the GENI-ACT instructions with minimal deviations. The deviations include: using ecocyc instead of metacyc, not doing the paralog module, using 20 matches instead of ten on the Tcoffee bioinformatics site, performing the *E. coli* blast against the *M. ruber* genome, and excluding some species for the Tcoffee and Blast.

| **MODULES** | **BIOINFORMATICS PROGRAMS** |
|---|---|
| Basic Information | GENI-ACT: http://geni-act.org/ |
| Sequence-Based Similarity Data | NCBI BLAST: http://blast.ncbi.nlm.nih.gov/Blast.cgi<br>CCD: http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml T-Coffee: http://www.tcoffee.org/Projects/tcoffee/<br>WebLogo: http://weblogo.berkeley.edu/logo.cgi |
| Cellular Localization Data | TMHMM: http://www.cbs.dtu.dk/services/TMHMM-2.0/<br>SignalP: http://www.cbs.dtu.dk/services/SignalP/<br>LipoP: http://www.cbs.dtu.dk/services/LipoP/<br>PSORT-B: http://www.psort.org/psortb/<br>Phobius:http://phobius.sbc.su.se/ |
| Alternative Open Reading Frame | JGI IMG/EDU 6-Frame viewer: http://img.jgi.doe.gov/cgi--- bin/edu/main.cgi |
| Structure-Based Evidence | TIGRFAM: http://blast.jcvi.org/web---hmm/<br>Pfam: http://pfam.xfam.org/search<br>PDB: http://www.rcsb.org/pdb/home/home.do |
| Enzymatic Function | KEGG: http://www.genome.jp/kegg/<br>MetaCyc: http://metacyc.org/<br>EcoCyc:http://ecocyc.org/<br>ExPASy: http://enzyme.expasy.org/enzyme---search---ec.html |
| Horizontal Gene Transfer | Phylogeny.fr: http://www.phylogeny.fr/)<br>JGI IMG/EDU: http://img.jgi.doe.gov/cgi---bin/edu/main.cg |

## 3. Results

*3.1 E. coli b2103 and M. ruber Mrub_2046 (TABLE 1)*

Using TMHMM both b2103 and Mrub_2046 were predicted to have zero transmembrane helices (Figure A1a). Additionally, they were not likely to contain a signal peptide and were not found to have cleavage sites (FigureA1b). PSORT-B predicted that the location of the genes were unknown (p=2.0 ; p=2.0) and lipoP predicted that both of the proteins were cytoplasmic. However, Phobius predicted the location as non-cytoplasmic for both b2103 and Mrub_2046 (FigureA1c). A BLAST of b2103 and Mrub_2046 gave an alignment with a bit score of 192 bits, 47% identity, and an expect value of 5e-64 (Figure A2a). A BLAST of b2103 gave a top hit of hydroxymethyl pyrimidine for Shigella sp. with an expect value of 0.0 (FigureA2b). A blast of Mrub_2046 gave a top hit of hydroxymethylpyrimidine for Thermus oshimai with an expect value of 6e-92 (FigureA2c). KEGG pathway for both b1203 and Mrub_2046 show that the products belong to the same pathway (FigureA3a & FigureA3b). For CDD and CG analysis both gene products had the same top hit. The COG number was COG0351 with E=6.21e-116 and E=1.2e-136 for Mrub_2046 and b2103 respectively. The Pfam results showed PF08543 with E=1.1e-90 and E=4.5e-95 for Mrub_2046 and b2103 respectively. The TIGRfam results showed TIGR00097 with E=1.2e-127 and E=4.7e-174 for Mrub_2046 and b2103 respectively. The GC content for b2103 is 55% with a genomic GC content of 51%. The GC content for Mrub_2046 is 63% with a genomic GC content of 67%.

**Table 1: *E. coli* b2103 and Mrub 2046 are orthologs**

| Description of Evidence Collected | *M. ruber* (2046) | *E. coli* (b2103) |
|---|---|---|
| Cellular Localization | Cytoplasmic | |
| BLAST *E. coli* against *M.* | Score:192 bits | |

| | | |
|---|---|---|
| *ruber* | E-value: 5e-64 | |
| KEGG pathway | Thiamine Metabolism | |
| CDD | ThiD; Hydroxymethylpyrimidine/phosphomethylpyrimidine kinase | |
| | E-value: 6.21e-116 | E-value: 1.20e-136 |
| Pfam | PF08543<br>Phos_pyr_kin (phosphomethylpyrimidine kinase) | |
| | E-value: 1.1e-90 | E-value:4.5e-95 |
| TIGRfam | TIGR00097<br>HMP-P_kinase: phosphomethylpyrimidine kinase | |
| | E-value: 1.2e-127 | E-value: 4.7e-174 |
| E.C. Number | E.C.2.7.4.7 Phosphomethylpyrimidine kinase | |

*3.2 E. coli b3993 and M. ruber Mrub_2041 (TABLE 2)*

Using TMHMM both b3993 and Mrub_2041 were predicted to have zero transmembrane helices (FigureB1a). Additionally, they were not likely to contain a signal peptide and were not found to have cleavage sites (FigureB1b). PSORT-B predicted that the location of the genes were in the cytoplasm (p=8.96 ; p=8.96) and lipoP also predicted that both of the proteins were cytoplasmic. However, Phobius predicted the location as non-cytoplasmic for both b3993 and Mrub_2041 (FigureB1c). A BLAST of b3993 and Mrub_2041 gave an alignment with a bit score of 70.9 bits, 36% identity, and an expect value of 7e-20 (Figure B2a). A BLAST of b3993 gave a top hit of thiamine phosphate synthase for Shigella dysenteriae with an expect value of 2e-147 (FigureB2b). A blast of Mrub_2041 gave a top hit of thiamine phosphate synthase for Thermus

igniterrae with an expect value of 7e-90 (FigureB2c). KEGG pathway for both b3993 and Mrub_2041 show that the products belong to the same pathway (FigureA3a & FigureA3b). For CDD and CG analysis both gene products had the same top hit. The COG number was COG0352 with E=4.0E-202 and E=4.22E-63 for Mrub_2041 and b3993 respectively. The Pfam results showed PF02581 with E=3.3E-57 and E=2.4e-56 for Mrub_2041 and b3993 respectively. The TIGRfam results showed TIGR00693 with E=9.3e-88 and E=4.2e-90 for Mrub_2041 and b3993 respectively. The GC content for b3993 is 58% with a genomic GC content of 51%. The GC content for Mrub_2041 is 68% with a genomic GC content of 63%.

**Table 2: E. *coli* b3993 and Mrub_2041 are orthologs**

| Description of Evidence Collected | *M. ruber* (2041) | *E. coli* (b3993) |
|---|---|---|
| Cellular Localization | Cytoplasmic | |
| BLAST *E. coli* against *M. ruber* | Score: 70.9bits<br>E-value: 7e-20 | |
| KEGG pathway | Thiamine Metabolism | |
| CDD | ThiE; Thiamine monophosphate synthase | |
| | E-value: 4e-202 | E-value: 4.22e-63 |
| Pfam | PF02581<br>Thymine monophosphate synthase:TENI | |
| | E-value: 3.3e-57 | E-value:2.4e-56 |
| TIGRfam | TIGR00693<br>thiE: thiamine phosphate pyrophosphorylase | |
| | E-value: 9.3e-88 | E-value: 4.2e-90 |

| | |
|---|---|
| E.C. Number | E.C.2.5.1.3 Thiamine-Phosphate Synthase (Also accepted: thiamine-phosphate pyrophosphorylase) |

## 4. Conclusions

*4.1 E. coli b2103 and M. ruber Mrub_2046*

The results of the gene comparison suggest that the two genes are similar in functionality. There was a slight discrepancy in cellular localization determination in that P-SORT was unable to predict a cellular location, lipoP predicted a cytoplasmic location, and Phobius predicted a non-cytoplasmic location. Ultimately, a cytoplasmic location was determined as the products are found in the cytosol according to EcoCyc. Aside from the minor discrepancy, all other results point to similarity between the two genes. The BLAST results of b2103 and Mrub_2046 against each other's protein sequences resulted in a low expect value and the KEGG, CDD, Pfam, and TIGRfam results were identical between the two gene products. All of these results suggest functional relatedness. Horizontal gene transfer is not expected because there are no significant differences between the genomic and specific gene GC percentages.

Based on the striking similarities between the loci of *M. ruber* and *E. coli* we have supported the hypothesis that if we compare Mrub_2046 locus of *Meiothermus ruber* DSM 1279 to the b2103 locus of *Escherichia coli* str. K-12 substr. MG1655 and the bioinformatics tools show the genes are similar in genetic makeup, then the Mrub_2046 locus of the *M. ruber* DSM 1279 must also code for the thiD gene has been proven. In conclusion, our data strongly supports that Mrub_2046 locus of the M. *ruber* DSM 1279 codes for the thiD gene.

*4.2 E. coli b3993 and M. ruber Mrub_2041*

The results of the gene comparison suggest that the two genes are similar in functionality. There was a slight discrepancy in cellular localization determination in that P-SORT predicted a cytoplasmic location, lipoP predicted a cytoplasmic location, and Phobius predicted a non-cytoplasmic location. Ultimately, a cytoplasmic location was determined as the products are found in the cytosol according to EcoCyc. Aside from the minor discrepancy, all other results point to similarity between the two genes. The BLAST results of b3993 and Mrub_2041 against each other's protein sequences resulted in a low expect value and the KEGG, CDD, Pfam, and TIGRfam results were identical between the two gene products. All of these results suggest functional relatedness. Distant horizontal gene transfer is expected because there are is a significant difference (±5%) between the genomic and specific gene GC percentages.

Based on the striking similarities between the loci of *M. ruber* and *E. coli* we have supported the hypothesis that if we compare Mrub_2041 locus of Meiothermus *ruber* DSM 1279 to the b3993 locus of *Escherichia coli* str. K-12 substr. MG1655 and the bioinformatics tools show the genes are similar in genetic makeup, then the Mrub_2041 locus of the *M. ruber* DSM 1279 must also code for the thiE gene has been proven. In conclusion, our data strongly supports that Mrub_2041 locus of the *M. ruber* DSM 1279 does in fact code for the thiE gene.

**References:**


1. Mizote T, Nakayama H (1989). “Purification and properties of hydroymethylpyrimidine kinase from Escherichia coli.” Biochem Biophys Acta 991(1):109-13
2. Mizote t, Tsuda M, Smith DD, Nakayama H, Nakazawa T (1999). “Cloning and characterization of the thiD/J gene of Escherichia coli encoding a thiamin-synthesizing biufunctional enzyme, hydroxymethylpyrimidine kinase/phosohomethylpyrimidine kinase.” Microbiology 145(2):495-501
3. Backstrom AD, McMordie RAS, Begley TP (1995). “Biosynthesis of Thiamin I: The Function of the thiE Gene Product.” J Am Chem Soc 117:2351-352.
4. Vander Horn PB1, Backstrom AD, Stewart V, Begley TP (1993) “Structural genes for thiamine biosynthetic enzymes (thiCEDGH) in Escherichia coli K-12.” J Bacteriol 175(4):982-92.
5. Pujari, S. “Bioinformatics: A useful essay on bioinformatics & biotechnology.0” (Internet address:http://www.yourarticlelibrary.com/essay/bioinformatics-an-useful-essay-on-bioinformaticsbiotechnology/29374/
6. Meiothermus ruber genome analysis project . [accessed 2015 Dec]. http://www.geniact.org

Appendix A

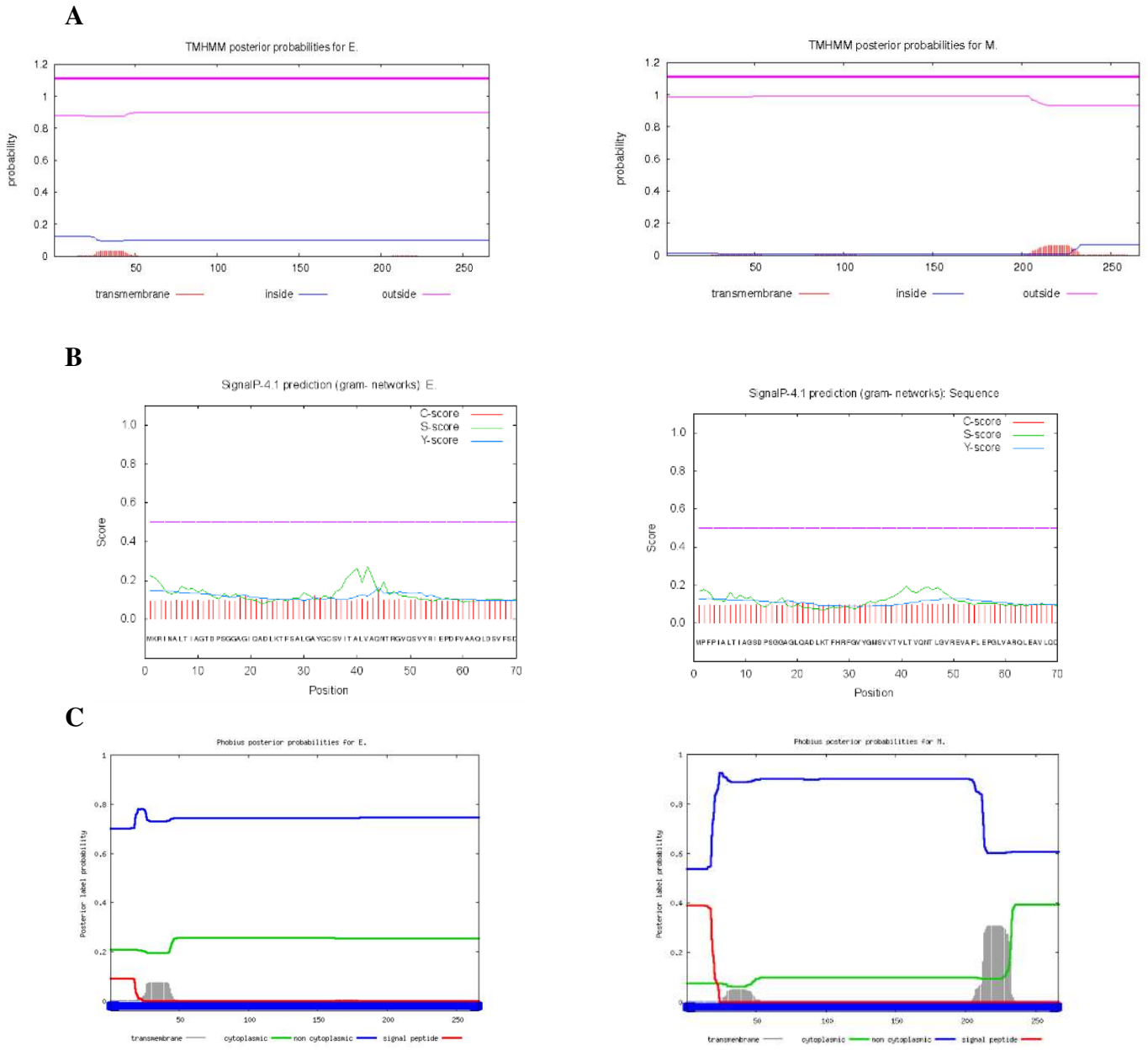



Fig. A.1 Bioinformatics tools used for cellular localization predict whether b2103 (left) and Mrub_2046 (right) are cytoplasmic or non-cytoplasmic. (a) TMHMM predicts neither b2103 or Mrub_2046 have transmembrane helices, (b) SignalP predicts neither b2103 or Mrub_2046 contain any signal peptide sequences, (c) Phobius predicts non-cytoplasmic gene products from both b2103 and Mrub_2046. The conflicting results could be due to having peripheral proteins that do not belong to either categories: cytoplasmic free- floating or transmembrane integral proteins.

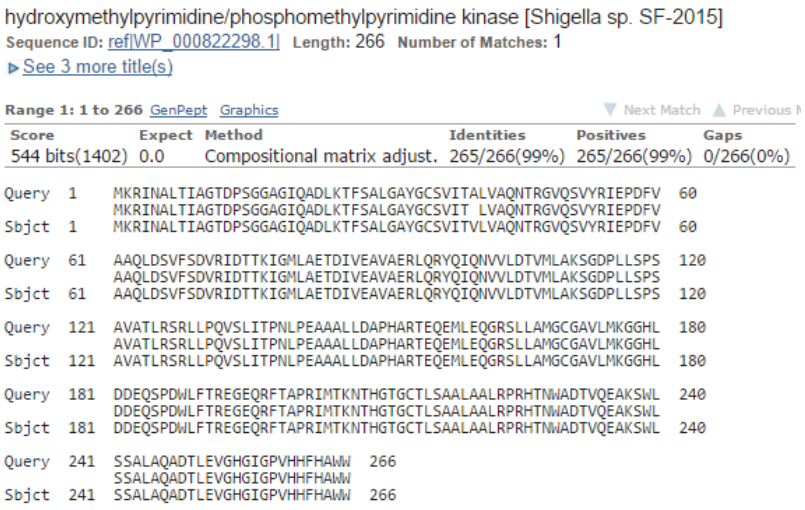
**A**

Range 1: 6 to 245 Graphics — Next Match — Previous Match

| Score | Expect | Method | Identities | Positives | Gaps |
|---|---|---|---|---|---|
| 192 bits(488) | 5e-64 | Compositional matrix adjust. | 114/242(47%) | 150/242(61%) | 2/242(0%) |

```
Query  6    ALTIAGTDPSGGAGIQADLKTFSALGAYGCSVITALVAQNTRGVQSVYRIEPDFVAAQLD  65
            ALTIAG+DPSGGAG+QADLKTF   G YG SV+T L  QNT GV+ V  +EP  VA QL+
Sbjct  6    ALTIAGSDPSGGAGLQADLKTFHRFGVYGMSVVTVLTVQNTLGVREVAPLEPGLVARQLE  65

Query  66   SVFSDVRIDTTKIGMLAETDIVEAVAERLQRYQIQNVVLDTVMLAKSGDPLLSPSAVATL  125
            +V  D      K G L +  IV ++A  L +  +  +V+D V++AKSGD LL+  A+  L
Sbjct  66   AVLQDPGAHAIKTGALGDAAIVHSIAPILAQTNLP-LVVDPVLVAKSGDWLLTEQALEAL  124

Query  126  RSRLLPQVSLITPNLPEAAALLDAPHARTEQEMLEQGRSLLAMGCGAVLMKGGHLDDEQS  185
            +S L P  +L+TPNLPEA ALL  P  R   +  E  R L  +G  AVL+KGGHL  E+S
Sbjct  125  KSELFPLATLLTPNLPEARALLGQP-IRDLADAREAARLLGGLGPRAVLLKGGHLAGEES  183

Query  186  PDWLFTREGEQRFTAPRIMTKNTHGTGCTLSAALAALRPRHTNWADTVQEAKSWLSSALA  245
             D L+       FTA +I + +THGTGCTLSAA+ AL  +     + V  AK +++ A+
Sbjct  184  TDVLWDGRKLHLFTAQKIPSSHTHGTGCTLSAAITALLAKGVALLEAVARAKRFVTRAIE  243

Query  246  QA  247
             A
Sbjct  244  TA  245
```

**B**



hydroxymethylpyrimidine/phosphomethylpyrimidine kinase [Shigella sp. SF-2015]

Sequence ID: ref|WP_000822298.1| Length: 266 Number of Matches: 1

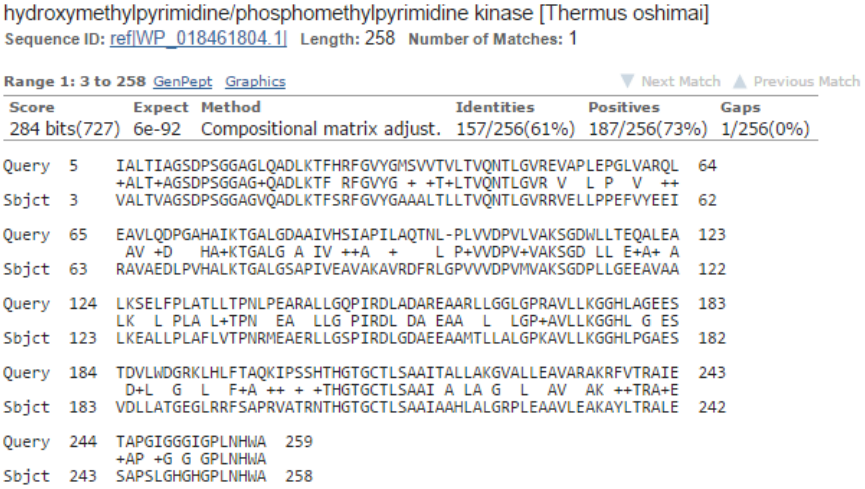
▶ See 3 more title(s)

Range 1: 1 to 266 GenPept Graphics — Next Match — Previous M

| Score | Expect | Method | Identities | Positives | Gaps |
|---|---|---|---|---|---|
| 544 bits(1402) | 0.0 | Compositional matrix adjust. | 265/266(99%) | 265/266(99%) | 0/266(0%) |

```
Query  1    MKRINALTIAGTDPSGGAGIQADLKTFSALGAYGCSVITALVAQNTRGVQSVYRIEPDFV  60
            MKRINALTIAGTDPSGGAGIQADLKTFSALGAYGCSVIT LVAQNTRGVQSVYRIEPDFV
Sbjct  1    MKRINALTIAGTDPSGGAGIQADLKTFSALGAYGCSVITVLVAQNTRGVQSVYRIEPDFV  60

Query  61   AAQLDSVFSDVRIDTTKIGMLAETDIVEAVAERLQRYQIQNVVLDTVMLAKSGDPLLSPS  120
            AAQLDSVFSDVRIDTTKIGMLAETDIVEAVAERLQRYQIQNVVLDTVMLAKSGDPLLSPS
Sbjct  61   AAQLDSVFSDVRIDTTKIGMLAETDIVEAVAERLQRYQIQNVVLDTVMLAKSGDPLLSPS  120

Query  121  AVATLRSRLLPQVSLITPNLPEAAALLDAPHARTEQEMLEQGRSLLAMGCGAVLMKGGHL  180
            AVATLRSRLLPQVSLITPNLPEAAALLDAPHARTEQEMLEQGRSLLAMGCGAVLMKGGHL
Sbjct  121  AVATLRSRLLPQVSLITPNLPEAAALLDAPHARTEQEMLEQGRSLLAMGCGAVLMKGGHL  180

Query  181  DDEQSPDWLFTREGEQRFTAPRIMTKNTHGTGCTLSAALAALRPRHTNWADTVQEAKSWL  240
            DDEQSPDWLFTREGEQRFTAPRIMTKNTHGTGCTLSAALAALRPRHTNWADTVQEAKSWL
Sbjct  181  DDEQSPDWLFTREGEQRFTAPRIMTKNTHGTGCTLSAALAALRPRHTNWADTVQEAKSWL  240

Query  241  SSALAQADTLEVGHGIGPVHHFHAWW  266
            SSALAQADTLEVGHGIGPVHHFHAWW
Sbjct  241  SSALAQADTLEVGHGIGPVHHFHAWW  266
```

**C**



hydroxymethylpyrimidine/phosphomethylpyrimidine kinase [Thermus oshimai]

Sequence ID: ref|WP_018461804.1| Length: 258 Number of Matches: 1

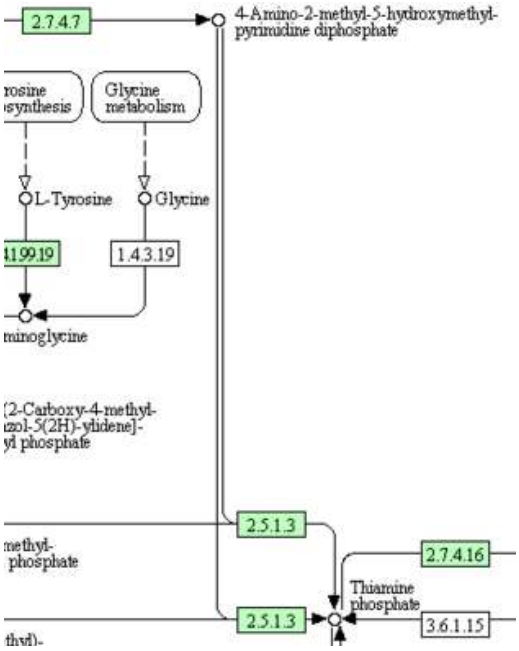
Range 1: 3 to 258 GenPept Graphics — Next Match — Previous Match

| Score | Expect | Method | Identities | Positives | Gaps |
|---|---|---|---|---|---|
| 284 bits(727) | 6e-92 | Compositional matrix adjust. | 157/256(61%) | 187/256(73%) | 1/256(0%) |

```
Query  5    IALTIAGSDPSGGAGLQADLKTFHRFGVYGMSVVTVLTVQNTLGVREVAPLEPGLVARQL  64
            +ALT+AGSDPSGGAG+QADLKTF RFGVYG + +T+LTVQNTLGVR V  L P  V  ++
Sbjct  3    VALTVAGSDPSGGAGVQADLKTFSRFGVYGAAALTLLTVQNTLGVRRVELLPPEFVYEEI  62

Query  65   EAVLQDPGAHAIKTGALGDAAIVHSIAPILAQTNL-PLVVDPVLVAKSGDWLLTEQALEA  123
             AV +D   HA+KTGALG A IV ++A  +    L P+VVDPV+VAKSGD LL E+A+ A
Sbjct  63   RAVAEDLPVHALKTGALGSAPIVEAVAKAVRDFRLGPVVVDPVMVAKSGDPLLGEEAVAA  122

Query  124  LKSELFPLATLLTPNLPEARALLGQPIRDLADAREAARLLGGLGPRAVLLKGGHLAGEES  183
            LK  L PLA L+TPN  EA  LLG PIRDL DA EAA  L  LGP+AVLLKGGHL G ES
Sbjct  123  LKEALLPLAFLVTPNRMEAERLLGSPIRDLGDAEEEAAMTLLALGPKAVLLKGGHLPGAES  182

Query  184  TDVLWDGRKLHLFTAQKIPSSHTHGTGCTLSAAITALLAKGVALLEAVARAKRFVTRAIE  243
             D+L  G  L  F+A ++ + +THGTGCTLSAAI A LA G  L  AV  AK ++TRA+E
Sbjct  183  VDLLATGEGLRRFSAPRVATRNTHGTGCTLSAAIAAHLALGRPLEAAVLEAKAYLTRALE  242

Query  244  TAPGIGGGIGPLNHWA  259
            +AP +G G GPLNHWA
Sbjct  243  SAPSLGHGHGPLNHWA  258
```

Fig.A.2 BLAST alignments for b2103 and Mrub_2046. (a) NCBI BLAST of b2103 against Mrub_2046, (b) top BLAST hit for b2103, (c) top BLAST hit for Mrub_2046.

A




B

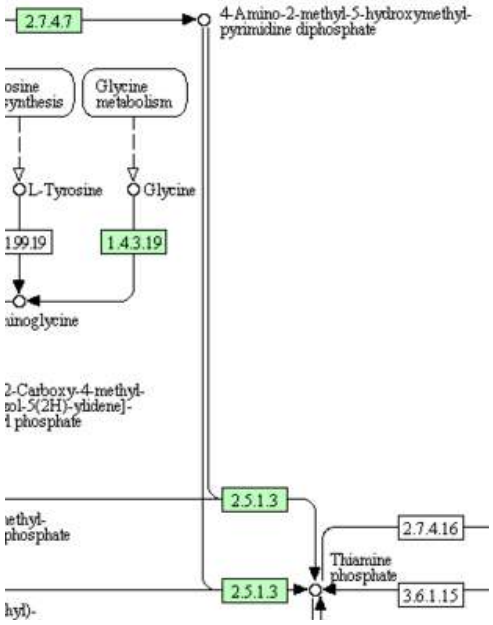
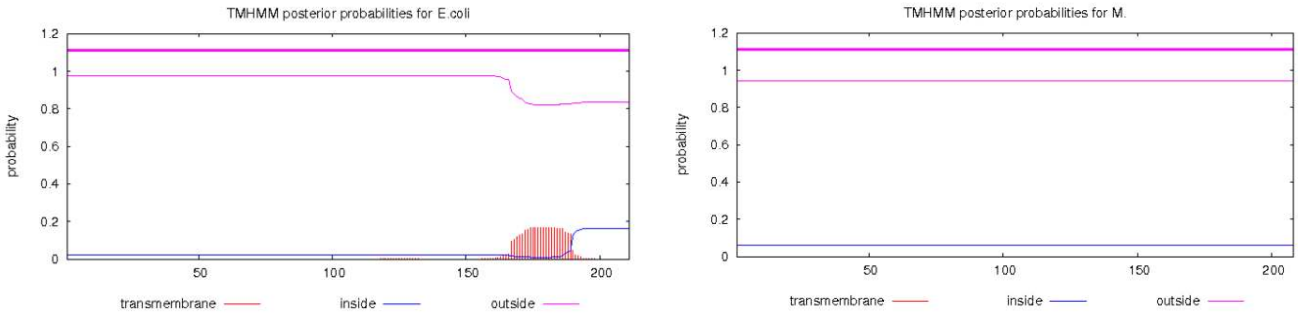


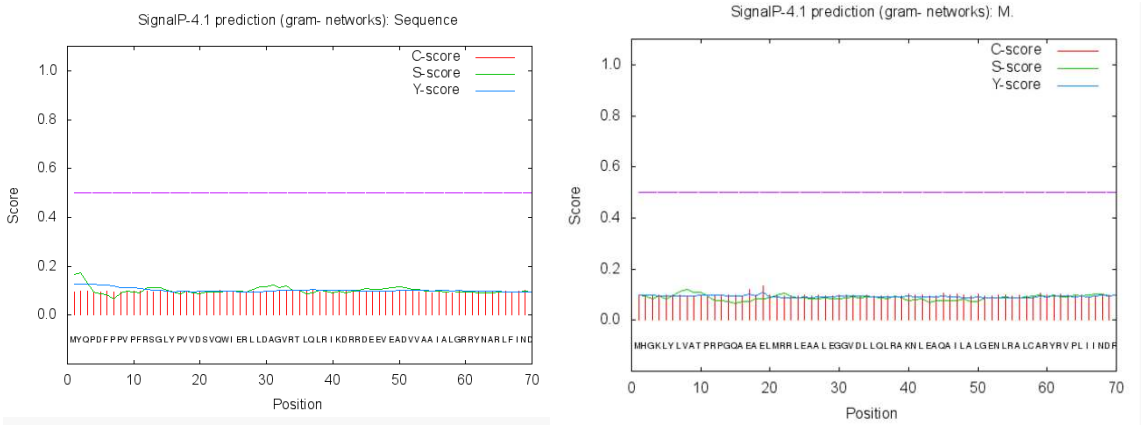

Fig. A.3 The genes being studied belong to the same KEGG pathway. (a) Partial KEGG pathway for *E. coli* and E.C. 2.7.4.7 corresponds to b2103 and E.C. 2.5.1.3 corresponds to b3993 and (b) Partial KEGG pathway for *M. ruber* and E.C 2.7.4.7 corresponds to Mrub_2046 and E.C. 2.5.1.3 corresponds to Mrub_2041.

Appendix B

**A**




**B**




**C**

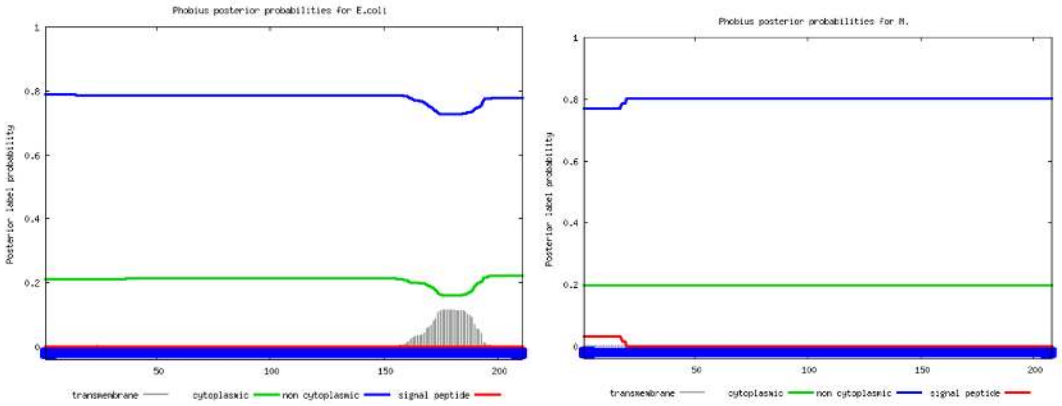



Fig. B.1 Bioinformatics tools used for cellular localization predict whether b3993 (left) and Mrub_2041 (right) are cytoplasmic or non-cytoplasmic. (a) TMHMM predicts neither b3993 or Mrub_2041 have transmembrane helices, (b) SignalP predicts neither b3993 or Mrub_2041 contain any signal peptide sequences, (c) Phobius predicts non-cytoplasmic gene products from both b3993 and Mrub_2041. The conflicting results could be due to having peripheral proteins that do not belong to either categories: cytoplasmic free- floating or transmembrane integral proteins.

**A**

**Range 1: 24 to 192** Graphics  Next Match  Previous M

| Score | Expect | Method | Identities | Positives | Gaps |
|---|---|---|---|---|---|
| 70.9 bits(172) | 7e-20 | Compositional matrix adjust. | 62/173(36%) | 80/173(46%) | 7/173(4%) |

```
Query  25   IERLLDAGVRTLQLRIKDRRDEEVEADVVAAIALGRRYNARLFINDYWRLAIKHQAYGVH  84
            +E  L+ GV  LQLR K+   + + A      AL  RY   L IND   LA   +A+GVH
Sbjct  24   LEAALEGGVDLLQLRAKNLEAQAILALGENLRALCARYRVPLIINDRPDLAALLEAHGVH  83

Query  85   LGQEDLQATDLNAIRAAGLRLGVSTHDD---MEIDVALAARPSYIALGHVFPTQTKQMPS  141
            LGQ DL      A R     +G STH+    +    AL   P Y+++G V+ T TK  P
Sbjct  84   LGQGDLNVA--QARRFFSGWIGRSTHEPEQALREQAALEGGPGYLSVGPVWETPTK--PG  139

Query  142  APQGLEQLARHVERLADYPTVAIGGISLARAPAVIATGVGSIAVVSAITQAAD  194
              P       R   +    P  AIGGI     P V+  G   +AVV +I  A D
Sbjct  140  RPAAGLAYVRWAAQNLRVPWFAIGGIDEHTLPQVLEAGARRVAVVRSILDAPD  192
```

**B**

thiamine phosphate synthase [Shigella dysenteriae]

Sequence ID: ref|WP_024250506.1| Length: 211 Number of Matches: 1

**Range 1: 1 to 211** GenPept Graphics  Next Match  Previous M

| Score | Expect | Method | Identities | Positives | Gaps |
|---|---|---|---|---|---|
| 421 bits(1082) | 2e-147 | Compositional matrix adjust. | 210/211(99%) | 211/211(100%) | 0/211(0%) |

```
Query  1    MYQPDFPPVPFRSGLYPVVDSVQWIERLLDAGVRTLQLRIKDRRDEEVEADVVAAIALGR  60
            MYQPDFPPVPFRSGLYPVVDSVQWIERLLDAGVRTLQLRIKDRRDEEVEADVVAAIALGR
Sbjct  1    MYQPDFPPVPFRSGLYPVVDSVQWIERLLDAGVRTLQLRIKDRRDEEVEADVVAAIALGR  60

Query  61   RYNARLFINDYWRLAIKHQAYGVHLGQEDLQATDLNAIRAAGLRLGVSTHDDMEIDVALA  120
            RYNARLFINDYWRLAIKHQAYGVHLGQEDLQATDLNAIRAAGLRLGVSTHDDMEIDVALA
Sbjct  61   RYNARLFINDYWRLAIKHQAYGVHLGQEDLQATDLNAIRAAGLRLGVSTHDDMEIDVALA  120

Query  121  ARPSYIALGHVFPTQTKQMPSAPQGLEQLARHVERLADYPTVAIGGISLARAPAVIATGV  180
            ARPSYIALGHVFPTQTKQMPSAPQGLEQLARHVERLADYPTVAIGGISLARAPAVIATGV
Sbjct  121  ARPSYIALGHVFPTQTKQMPSAPQGLEQLARHVERLADYPTVAIGGISLARAPAVIATGV  180

Query  181  GSIAVVSAITQAADWRLATAQLLEIAGVGDE  211
            GSIAVVS+ITQAADWRLATAQLLEIAGVGDE
Sbjct  181  GSIAVVSSITQAADWRLATAQLLEIAGVGDE  211
```

**C**

thiamine phosphate synthase [Thermus igniterrae]

Sequence ID: ref|WP_018110353.1| Length: 205 Number of Matches: 1

**Range 1: 1 to 204** GenPept Graphics  Next Match  Previous M

| Score | Expect | Method | Identities | Positives | Gaps |
|---|---|---|---|---|---|
| 275 bits(702) | 7e-90 | Compositional matrix adjust. | 145/207(70%) | 164/207(79%) | 3/207(1%) |

```
Query  1    MHGKLYLVATPRPGQAEAELMRRLEAALEGGVDLLQLRAKNLEAQAILALGENLRALCAR  60
            M G+LYLV TPRPG ++ + + R E AL GGV++LQLRAK+ EA+AIL LGE +RAL  R
Sbjct  1    MQGRLYLVVTPRPGWSQEKTLERTERALAGGVEVLQLRAKDWEARAILELGERMRALAWR  60

Query  61   YRVPLIINDRPDLAALLEAHGVHLGQGDLNVAQARRFFSGWIGRSTHEPEQALREQAALE  120
            Y VP ++NDRPDLAALLEA GVHLGQGDL   +ARRFFSG +GRSTH PEQAL+   ALE
Sbjct  61   YGVPFVLNDRPDLAALLEADGVHLGQGDLTPQEARRFFSGLVGRSTHAPEQALK---ALE  117

Query  121  GGPGYLSVGPVWETPTKPGRPAAGLAYVRWAAQNLRVPWFAIGGIDEHTLPQVLEAGARR  180
             G  YLSVGPVWETPTKPGR AAGL YVRWAA +LR PWFAIGGID   L QVLEAGARR
Sbjct  118  EGADYLSVGPVWETPTKPGRKAAGLGYVRWAAAHLRAPWFAIGGIDLANLDQVLEAGARR  177

Query  181  VAVVRSILDAPDPEKAARHMRRWLDGL  207
            V VVR+ILDA DPE+AAR  R  L G+
Sbjct  178  VVVVRAILDAEDPERAARAFRERLYGV  204
```

Fig.B.2 BLAST alignments for b3993and Mrub_2041. (a) NCBI BLAST of b2103 against Mrub_2041, (b) top BLAST hit for b3993, (c) top BLAST hit for Mrub_2041.