2-2016

Genomic Analysis of *Meiothermus ruber* Mrub_1907 and *Meiothermus ruber* Mrub_1844 with Potential Ortholog *Escherichia coli* b3774 IlvC and *Escherichia coli* b3771 IlvC Gene through Bioinformatics

Felipe A. Hernandez
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College - Rock Island

Felipe A. Hernandez

Dr. Scott

BIO 375-01

Genomic Analysis of *Meiothermus Ruber Mrub*_1907 and *Meiothermus Ruber Mrub*_1844 with Potential Ortholog *Escherichia coli* b3774 IlvC and *Escherichia coli* b3771 IlvC Genes through Bioinformatics

Introduction:

In this study, we compare two sets of genes that are believed to code for the same key enzymes in the isoleucine biosynthesis pathway. The first pair of genes come from the bacteria *Meiothermus ruber* and the second from *E. coli*. Comparing *M. ruber* to *E. coli* is important because so much information about E. coli is known. Therefore, *E. coli* serves as a great model to compare results to *M. ruber*. The method in which these two genes will be compared is through bioinformatics. Bioinformatics is a great tool to use when studying unknown features in an organism. With this special tool, scientists can analyze the numerous files of data from other known organisms find any similarity or contrast.

There is very little known about *Meiothermus ruber* (*M. ruber*) in the scientific community. For this reason, recent studies have pushed the boundaries on the knowledge of *M.ruber* to uncover any unknown secrets. Past studies have shown *M. ruber* as a Gram-negative bacterium that lives in relatively hot environments (Tindall et. al., 2010). However, there are many realms in the bacterial world that are still unknown to scientists.

The pathway of focus for the *M. ruber* genome is the L-isoleucine biosynthesis pathway (Figure 1). Since L-isoleucine is a non-polar amino acid, it is a very important to study one of the biggest structural backbones in all proteins (Betts et al. 2003). L-isoleucine biosynthesis is formed from a five-step process originating with the amino acid, threonine. The last four steps of this pathway are also in close ties with the pathway of valine biosynthesis. These entwined pathways are part of a super pathway of branched amino acid biosynthesis that generates not only isoleucine and valine, but also leucine (Umbarger 1957).

There is direct evidence of *E. coli* showing expression of L-isoleucine biosynthesis according to a study done by a team lead by M. Freundlich (Freundlich et al. 1962). The study was testing the relationship between the inhibitive properties of isoleucine, leucine, and valine (the super pathway of branched amino acids). Having evidence of this study shows proof that *E. coli* has the ability to express the L-isoleucine biosynthesis pathway.

The purpose of this study is to compare the two sets of *M. ruber* genes within the L-isoleucine biosynthesis pathway with their possible *E. coli* counterparts. The Mrub_1907 gene from *M. ruber* will be compared to the ilvC gene of E. coli, as well as the Mrub_1844 gene to ilvD using the tools of bioinformatics. It is hypothesized that Mrub_1907 and ilvC are orthologues, along with Mrub_1907 and ilvD respectively. If results are positive, then *M. ruber* will show evidence of possing the same ability as *E. coli.*

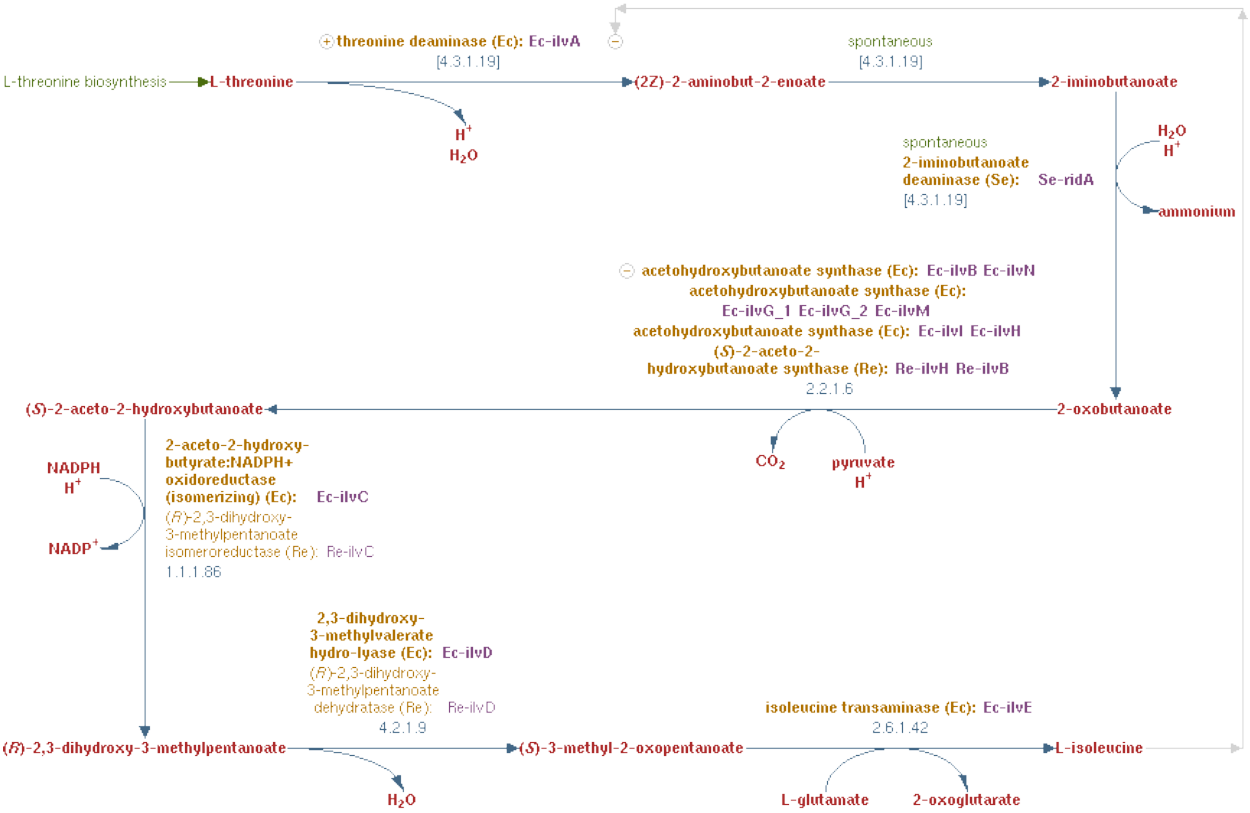


Figure 1. L-isoleucine biosynthesis pathway. Red= reactants, electron acceptors, and products. Gold= enzymes. Image from MetaCyc.

Methods:

Bioinformatics programs were derived from within the GEBI-ACT lab notebook. The web site for this lab notebook can be found at: http://www.geni-act.org/student/view_course/6fa2ae446a0244ad. Mostly all protocol was followed, but there were deviations present. An example was, the Tcoffee tool used 15 comparisons of amino strains rather than 10. There were no references to paralogs. Instead of using the metacyc search engine, ecocyc was used in its place. For T-coffee, BLAST of ilvC and ilvD had to exclude both E.coli and Salmonella e. because of insufficient amount of hits.

Results:

Table 1: E.coli ilvC and M. ruber_1907 are orthologues

Evidence collected	**E.coli**	**M. ruber**
Cellular Location (Module 3 in Lab notebook)	Cytoplasmic score: 10 (PSORT-B)	Cytoplasmic score: 9.97 (PSORT-B)
	Cytoplasmic	
BLAST E. coli against M.ruber	Score: 161 bits E-value: 3e-121; 36% identity	
KEGG pathway	L-isoleucine biosynthesis pathway	
E.C number	E.C #: 1.1.1.86 E.C Name: ketol-acid reductoisomerase	
Pfam- protein family	PF07991 IlvN, Acetohydroxy acid isomeroreductase, NADPH-binding domain Score= 197 bits ; E-value= 1.4e-58	PF07991 IlvN, Acetohydroxy acid isomeroreductase, NADPH-binding domain Score= 255.2 bits ; E-value= 1.7e-76
CDD (COG grouping)	COG #: COG0059 Ketol-acid reductoisomerase E-value: 2.96e-170	COG #: COG0059 Ketol-acid reductoisomerase E-value: 0.0
TIGRfam (protein family)	TIGR00465	TIGR00465

	E-value: 1.3e-149 ketol-acid reductoisomerase	E-value: 4.5e-138 ketol-acid reductoisomerase

Information regrarding the cellular location, protein sequence, protein family, and biosynthesis pathways were outlined using bioinformatics tools provided by the GENI-ACT lab notebook.

The genes of Mrub_1907 and E. coli ilvC were first compared with BLAST amino acid sequence analysis. The results of the BLAST search of the *M.ruber* gene (Mrub_1907) against E.coli yields are shown in Table 1. A pairwise alignment showed the presence of many conserved amino acids in similar locations (Fig. 2). The amino acids of Mrub_1907 that were seen to be conserved were G24, G26, G29, A57, P82, D83, G109, A129, P130, G134, G146, G173, G178. G44, G46, G49, A82, P107, D108, G133, A153, P154, G158, G170, G199, G204. For E. coli, G44, G46, G49, A82, P107, D108, G133, A153, P154, G158, G170, G199, G204 were similairly conserved.

A

```
#HMM      lkgkkiaviGyGsqGhaqalnlrdsgldvvglregs.....ksvkkAkeegfevltvaeavkkadvvmiliPDelqkevye
#MATCH    l+gkk++++G+G+qG +q+ln+rdsgld+ ++lr+++     +s++kA+e+gf+v t++e++ +ad+v++l+PD+++++v+
#PP       89*******************************988777767***********************************.
#SEQ      LQGKKVVIVGCGAQGLNQGLNMRDSGLDISYALRKEAiaekrASWRKATENGFKVGTYEELIPQADLVINLTPDKQHSDVV-
```

B

```
#HMM       lkgkkiaviGyGsqGhaqalnlrdsgldvvvglregsksvkkAkeegfevltvaeavkkadvvmiliPDelqkevye
#MATCH     +k+k++a++G+GsqGha+alnlrdsg++vvvglr gs++ +kA+++g+evl v eav++advvmil+PDe+q +vy+
#PP        89**********************************************************************
#SEQ       IKDKTVAILGFGSQGHAHALNLRDSGIKVVVGLRPGSRNEEKARKAGLEVLPVGEAVRRADVVMILLPDETQGAVYK
```

A

```
eeiepnlkegkalafahGfnihfglikppkdvdvvlvAPkgpGhlvreeyeegkGvpaliave..qdasgkakelalayakgiGatraG
++++p +k+g+al ++hGfni+    +++kd++vv+vAPk+pG++vreey++g+Gvp+liav+  +d +g+++++a+a+a ++G++raG
89**********************************************************988999*******************9
RTVQPLMKDGAALGYSHGFNIVEVGEQIRKDITVVMVAPKCPGTEVREEYKRGFGVPTLIAVHpeNDPKGEGMAIAKAWAAATGGHRAG
```

B

```
eeiepnlkegkalafahGfnihfglikppkdvdvvlvAPkgpGhlvreeyeegkGvpaliaveqdasgkakelalayakgiGatraG
+e+epnlkeg+alafahGfnihfg+ikp++d+dv++vAPkgpGhlvr+eye+g+Gvp+l+av+qdasg+a+ +alayak+ G+traG
*******************************************************************************9
AEVEPNLKEGAALAFAHGFNIHFGQIKPRRDLDVWMVAPKGPGHLVRSEYEKGSGVPSLVAVYQDASGSAFPTALAYAKANGGTRAG
```

Figure 2. E. coli ilvC and Mrub_1907 contain similar conserved amino acids. (A) = E. coli ilvC; (B) = Mrub_1907. #HMM = consensus protein sequence; #MATCH = match between bacteria and consensus sequence; #SEQ = bacteria amino acid sequence. Produced on Pfam (http://pfam.xfam.org/search).

The pathway that both organisms were also analyzed to confirm the presence of Mrub_1907 and ilvC. Figure 3 shows a KEGG map of both isoleucine pathways in E. coli (Fig.3A) and M.ruber (Fig. 3B). E.C numbers were also provided by KEGG (Table 1) to show that both genes containing the same number and name. CDD analysis reinforced the result of identical names by showing similar results with low E-values (Table 1). TIGRFAM analysis of Mrub_1907 and E. coli ilvC gave identical TIGRFAM numbers accompanied by small E-values (Table 1). Finally, Pfam analysis concluded the analysis of similar proteins by showing the same Pfam numbers, and protein family names (Table 1).

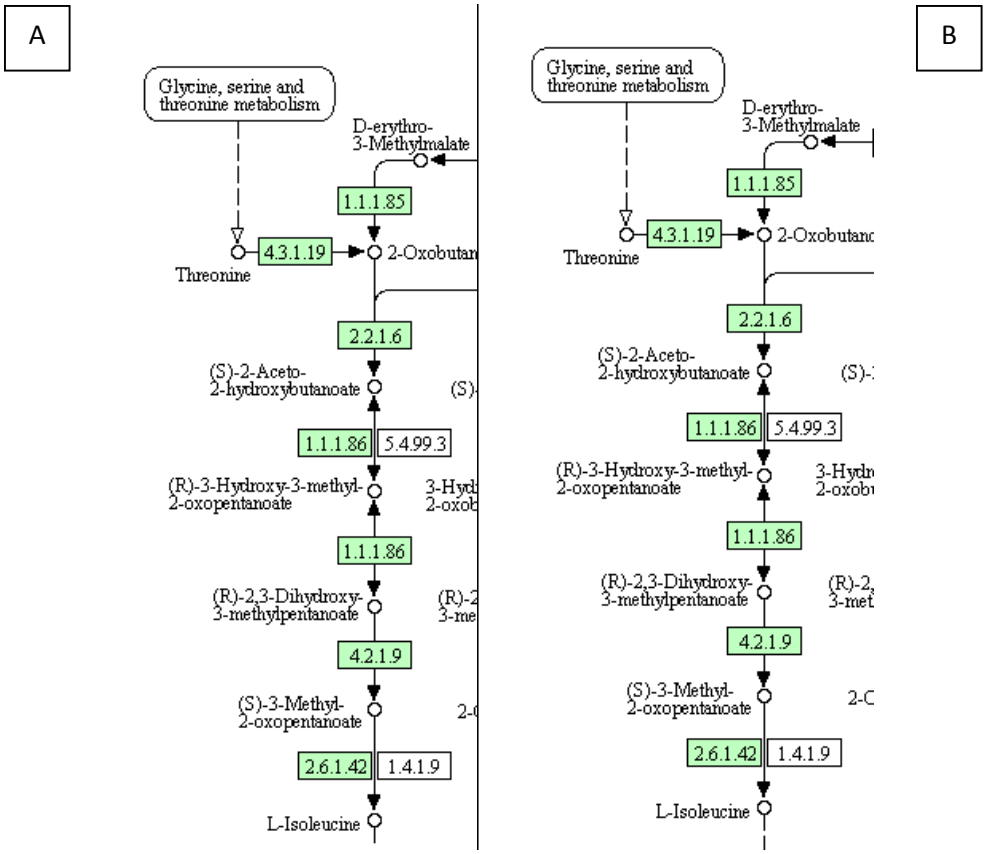


Figure 3. Both L-isoleucine biosynthesis pathways for E. coli and M. ruber shows the presence of ilvC (E.C. 1.1.86) and ilvD (E.C. 4.2.1.9). (A) = E. coli pathway, (B) = M. ruber pathway. Green represents proteins produced by the bacteria. Created by KEGG PATHWAY Database (http://www.genome.jp/kegg/pathway.html).

For cellular location analysis, the following bioinformatics tools were used. TMHMM analysis resulted in identical hydropathy plots for Mrub_1907 and E. coli ilvC (Fig. 4). Results of TMHMM data showed both Mrub_1907 and E. coli ilvC containing no transmembrane helices. This result was backed up by an inconclusive LipoP analysis. According to PSORT-B, the estimated cellular location was said to be in the cytoplasm (Table 1). Analysis of SignalP

results indicated no presence of a signal peptide on Mrub_1907 or E. coli ilvC concluding the cellular location analysis.

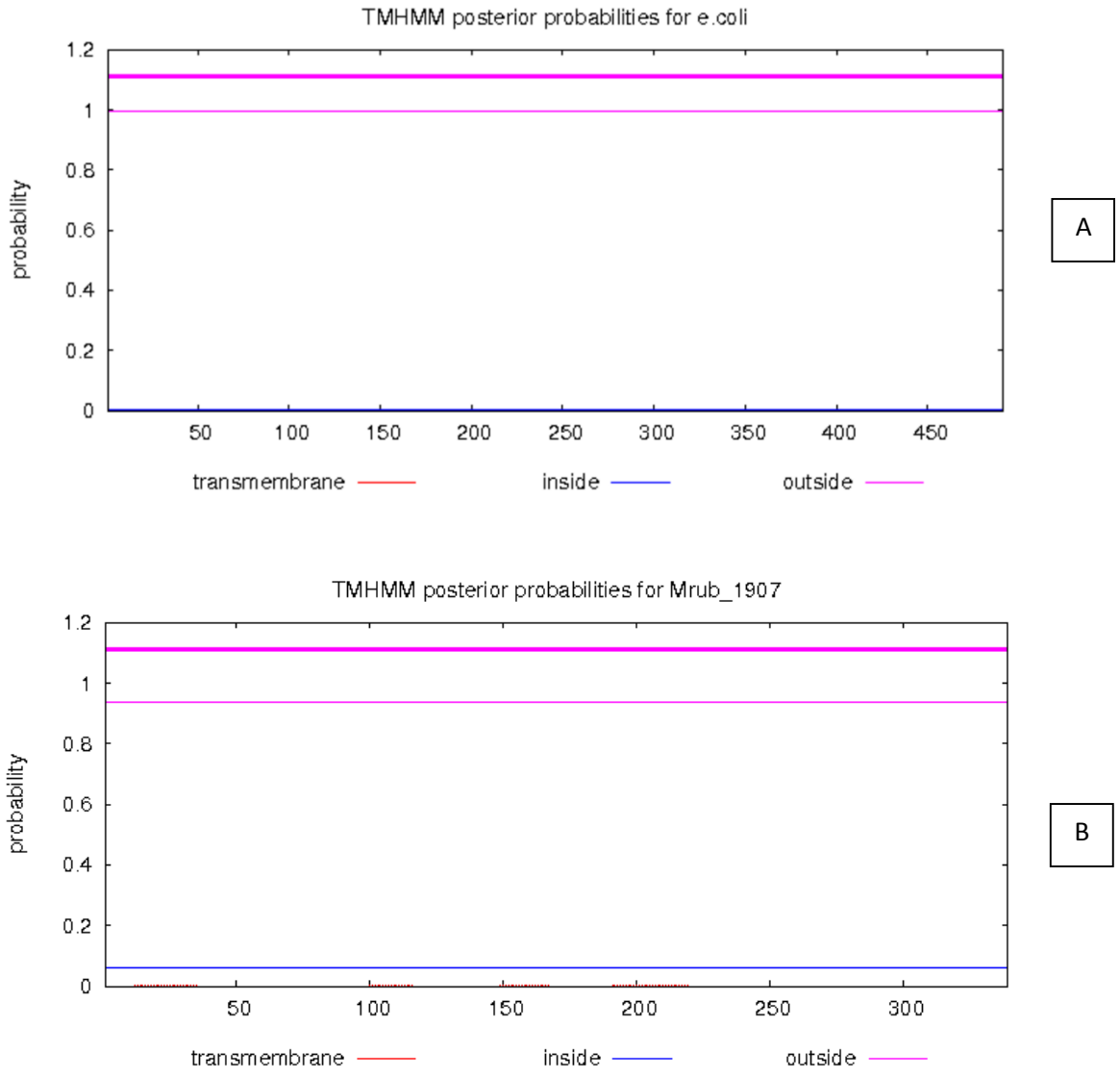
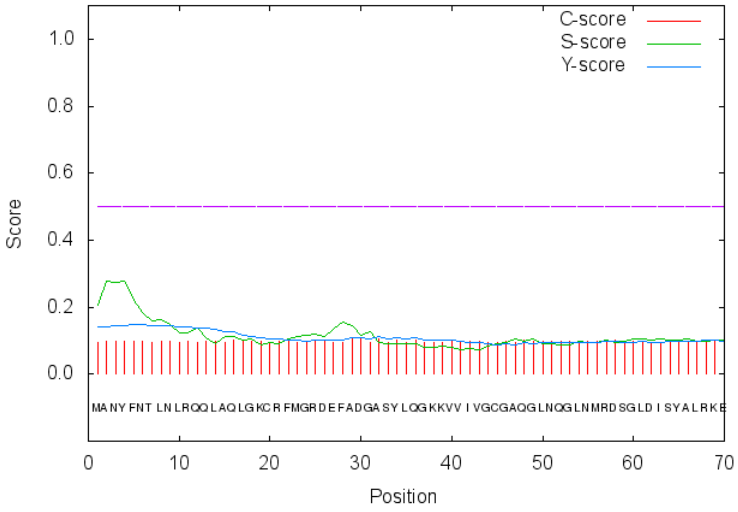
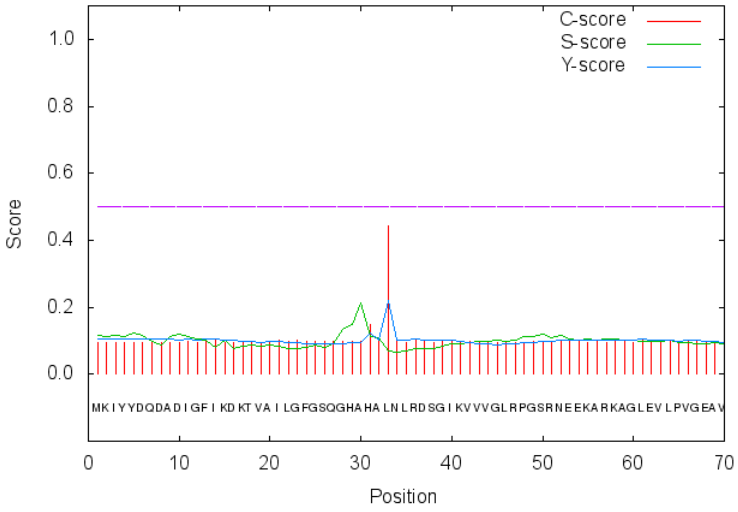


Figure 4. E. coli ilvC and Mrub_1907 do not contain TMH regions. Both enzymes are predicted to be in the cytoplasm. (A) = E. coli ilvC; (B) = Mrub_1907. Plots created by TMHMM Server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/).


C-score
S-score
Y-score
Score
Position
MANYFNTLNLRQQLAQLGKCRFMGRDEFADGASYLQGKKVVIVGCGAQGLNQGLNMRDSGLDISYALRKE

A


C-score
S-score
Y-score
Score
Position
MKIYYDQDADIGFIKDKTVAILGFGSQGHAHALNLRDSGIKVVVGLRPGSRNEEKARKAGLEVLPVGEAV

B

Figure 5. E. coli ilvC and Mrub_1907 do not contain signaling sequences (signal proteins). (A) = E. coli ilvC; (B) = Mrub_1907. E. coli ilvC D-score = 0.144; Mrub_1907 D-score = 0.218. Plots created by SignalP 4.1 Server (http://www.cbs.dtu.dk/services/SignalP).

Table 2: E.coli ilvD and M. ruber_1844 are orthologues

Evidence collected	**E.coli**	**M. ruber**
Cellular Location (Module 3 in Lab notebook)	Cytoplasmic score: 9.97 (PSORT-B)	Cytoplasmic score: 9.97 (PSORT-B)
	Cytoplasmic	
BLAST E. coli against M.ruber	Score: 372 bits E-value: 3e-121; 37% identity	

KEGG pathway	L-isoleucine biosynthesis pathway	
E.C number	E.C #: 4.2.1.9 E.C Name: dihydroxy acid dehydratase	
Pfam- protein family	PF00920 ILVD_EDD, Dehydratase family Score= 727.1 bits ; E-value= 9.6e-219	PF00920 ILVD_EDD, Dehydratase family Score= 662.7 bits ; E-value= 3.1e-199
CDD (COG grouping)	COG #: COG0129 Dihydroxyacid dehydratase/phosphogluconate dehydratase E-value: 4.79e	
TIGRfam (protein family)	TIGR00110 E-value: 0.0 ilvD: dihydroxy-acid dehydratase	TIGR00110 E-value: 7.5e-258 ilvD: dihydroxy-acid dehydratase

Information regrarding the cellular location, protein sequence, protein family, and biosynthesis pathways were outlined using bioinformatics tools provided by the GENI-ACT lab notebook.

The genes of Mrub_1844 and E. coli ilvD were first compared with BLAST amino acid sequence analysis. The results of the BLAST search of the M.ruber gene (Mrub_1844) against E.coli yields are shown in Table 2. A pairwise alignment showed the presence of many

conserved amino acids in similar locations (Fig. 6). The first ten amino acids of Mrub_1844 that were seen to be conserved were D82, L95, R98, D115, C123, D124, K125, P128, P140, G147 respectively. For ilvD, D78, R94, D111, C119, D120, K121, P124, P136, G143, G191 were shown to be conserved in a similar manner.

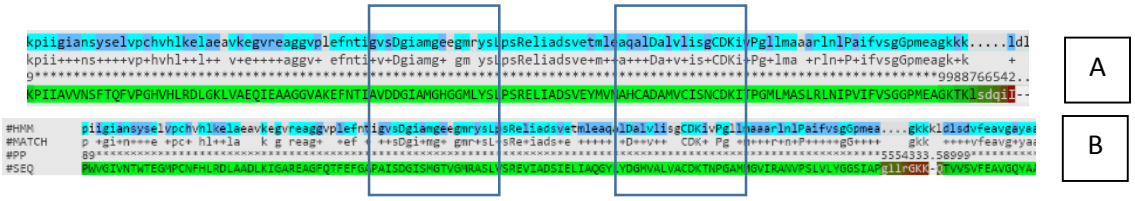
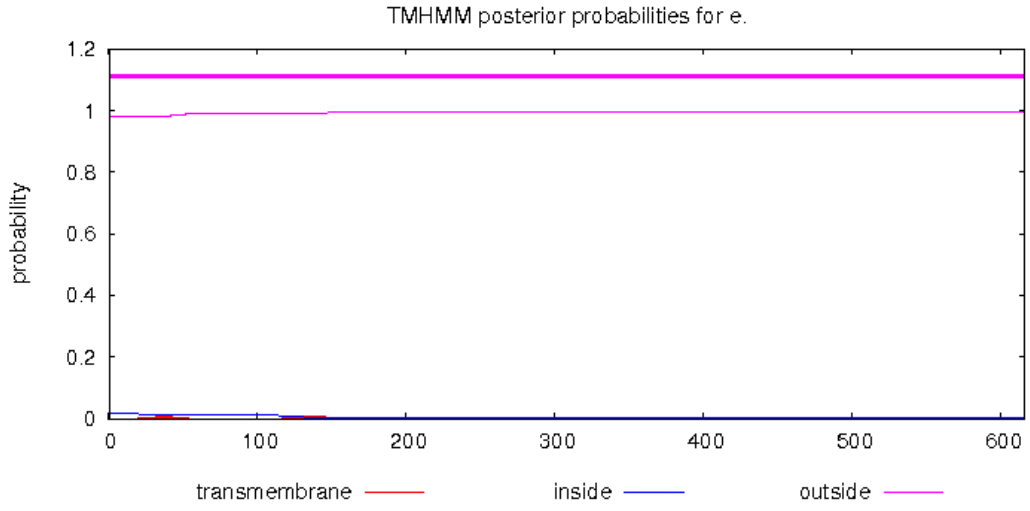


Figure 6. E. coli ilvD and Mrub_1844 contain somewhat similar conserved amino acids. Sequences are offset for better alignment. (A) = E. coli ilvD; (B) = Mrub_1844. #HMM = consensus protein sequence; #MATCH = match between bacteria and consensus sequence; #SEQ = bacteria amino acid sequence. Produced on Pfam (http://pfam.xfam.org/search).

The pathway that both organisms were also analyzed to confirm the presence of Mrub_1844 and ilvD. Figure 3 shows a KEGG map of both isoleucine pathways in E. coli (Fig.3A) and M.ruber (Fig. 3B). E.C numbers were also provided by KEGG (Table 2) to show that both genes containing the same number and name. CDD analysis reinforced the result of identical names by showing similar results with low E-values (Table 2). TIGRFAM analysis of Mrub_1844 and E. coli ilvD gave identical TIGRFAM numbers accompanied by small E-values (Table 2). PBD analysis revealed the same PDB code and protein name (Dihydroxyacid dehydratase/phosphogluconate dehydratase E-value: 4.79e) (Table 2). Finally, Pfam analysis

concluded the analysis of similar proteins by showing the same Pfam numbers, and protein family names (Table 2).

For cellular location analysis, the following bioinformatics tools were used. TMHMM analysis resulted in identical hydropathy plots for Mrub_1907 and E. coli ilvC. Results of TMHMM data showed both Mrub_1907 and E. coli ilvC containing no transmembrane helices (Fig. 7). This result was backed up by an inconclusive LipoP analysis. According to PSORT-B, the estimated cellular location was said to be in the cytoplasm (Table 2). Analysis of SignalP results indicated no presence of a signal peptide on Mrub_1907 or E. coli ilvC concluding the cellular location analysis (Fig. 8).



A

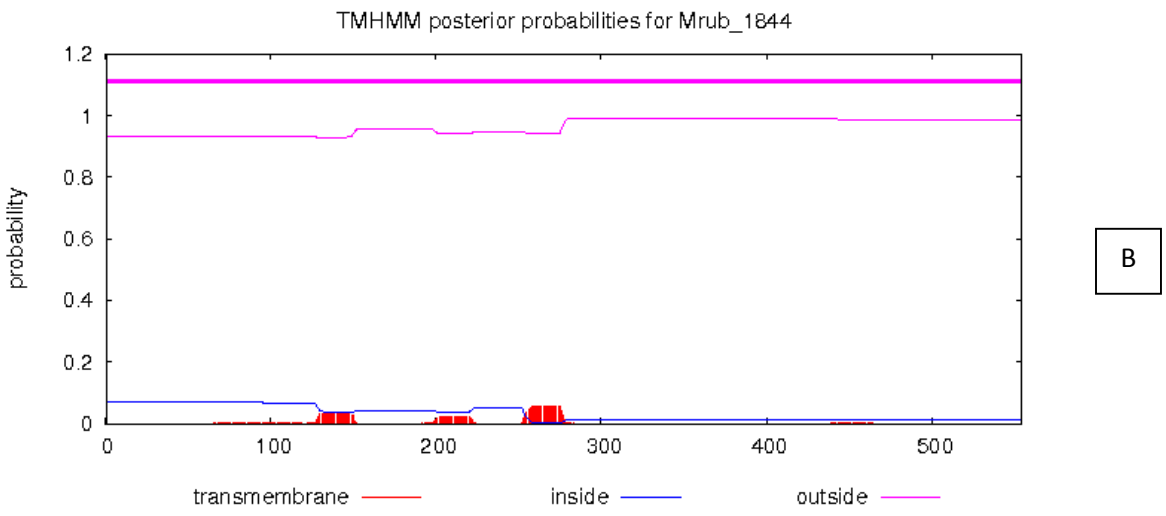
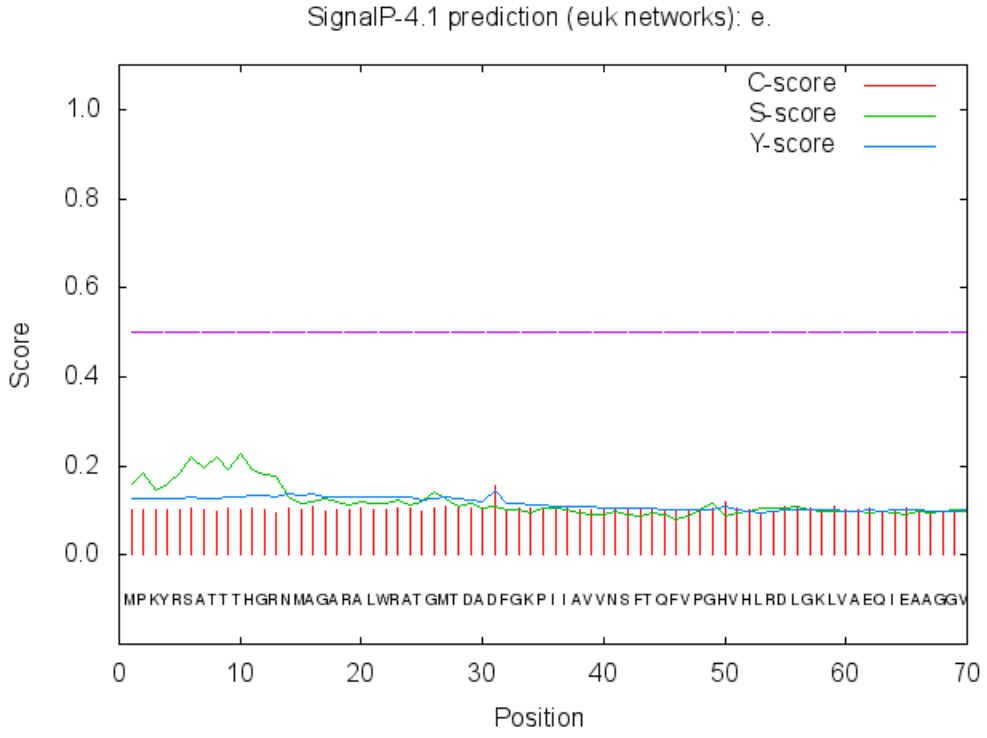
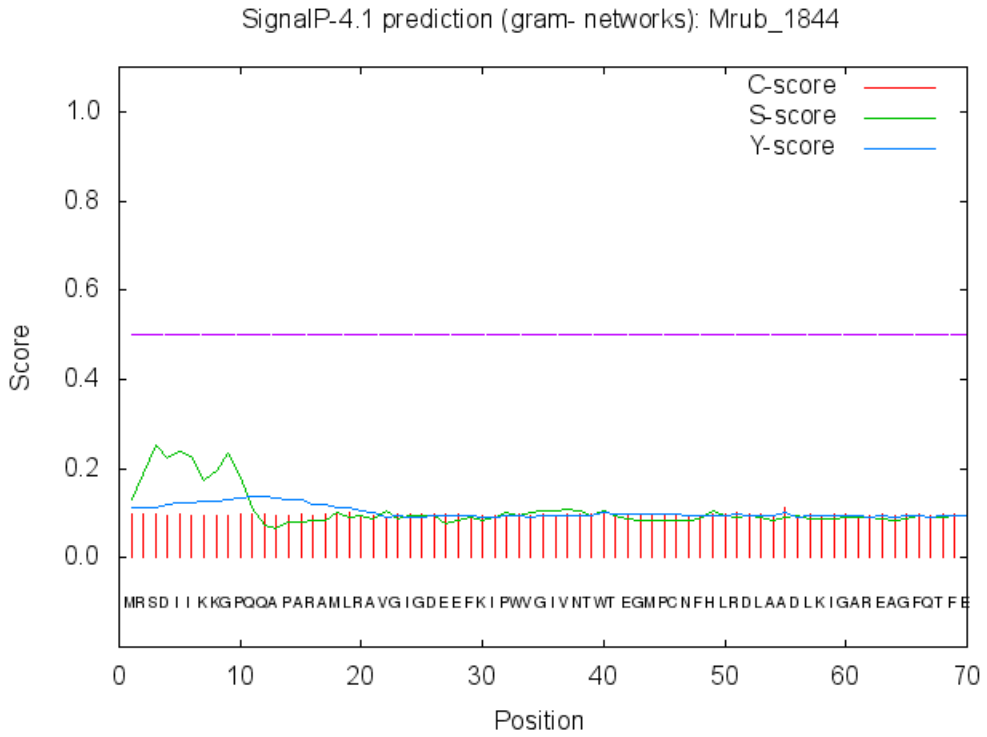


Figure 7. E. coli ilvD and Mrub_1844 do not contain TMH regions. Both enzymes are predicted to be in the cytoplasm. (A) = E. coli ilvD; (B) = Mrub_1844. Plots created by TMHMM Server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/).


SignalP-4.1 prediction (euk networks): e.
C-score
S-score
Y-score
Score
1.0
0.8
0.6
0.4
0.2
0.0
MPKYRSATTTHGRNMAGARALWRATGMTDADFGKPIIAVVNSFTQFVPGHVHLRDLGKLVAEQIEAAGGV
0
10
20
30
40
50
60
70
Position

A


SignalP-4.1 prediction (gram- networks): Mrub_1844
C-score
S-score
Y-score
Score
1.0
0.8
0.6
0.4
0.2
0.0
MRSDIIKKGPQQAPARAMLRAVGIGDEEFKIPWVGIVNTWTEGMPCNFHLRDLAADLKIGAREAGFQTFE
0
10
20
30
40
50
60
70
Position

B

Figure 8. E. coli ilvD and Mrub_1844 do not contain signaling sequences (signal proteins). (A) = E. coli ilvC; (B) = Mrub_1844. E. coli ilvD D-score = 0.144; Mrub_1907 D-score = 0.218. Plots created by SignalP 4.1 Server (http://www.cbs.dtu.dk/services/SignalP).

Conclusion:

Based off of the results of the comparative genome analysis, our hypothesis is confirmed that *M.ruber* and *E.coli* are orthologs. Starting at the cellular location, we can start to see the similarities between the two coded proteins. TMH plots of both proteins showed that they are both found outside of the membrane of the cell and in the cytoplasm (according to PSORT-B data). We can also infer a similar function by comparison of the similar sequences they show through Pfam bioinformatics. In comparing the sequences with the consensus sequence of the protein family, we see multiple key amino acids vital for the function of both ketol-acid reductoisomerase and dihydroxy acid dehydratase being conserved. Seeing these key amino acids conserved shows strong evidence that the function of both enzymes contain similar function .This argument can be further supported when looking at the pathway the two enzymes according to KEGG. The KEGG pathway for the enzyme *E. coli* and *M. ruber* code for are shown to be the same according to Figure 3. Table 1 and 2 are able to summarize all of the bioinformatics relative to the argument of *Mrub*_1907/ilvC and pMrub_1844/ilvD code for the same protein. The E-values determined in the BLAST results show that the genes from *M. ruber* and *E. coli* are related not do to chance. With evidence of the same pathway, conserved amino acids, cellular location and low E-value, can be said the *Mrub*_1080 genes and *E. coli* b0242 genes are orthologues.

References:

Betts M.J, Russell R.B. Amino acid properties and consequences of substitutions. Bioinformatics for Geneticists, M.R. Barnes, I.C. Gray eds, Wiley, 2003.

Freundlich M., Bums R. O., Umbarger H. E.. 1962. Control of isoleucine, valine and leucine biosynthesis. I. Multivalent repression. Proc. Natl. Acad. Sci. U.S.A. 48:1804-1808.

Tindall B, Sikorski J, Lucas S, Goltsman E, Copeland A, Del Rio TG, Nolan M, Tice H, Cheng J, 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 Y, Jeffries CD, Rohde M, Göker M, Woyke T, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk H, and Lapidus A. 2010. Complete genome sequence of meiothermus ruber type strain. Standards in Genomic Sciences. 3(26).

Umbarger HE, Brown B. 1957. Threonine deamination in Escherichia coli: II. Evidence for two L-threonine deamination. J. Bacteriol. 73(1):105-112