Spring 2-2016

Valine Biosynthesis: Mrub_2994 is Orthologous to *E. coli* b3770 and Mrub_1844 is Orthologous to *E. coli* b3771

Bennett A. Hartmann
Augustana College, Rock Island Illinois

Dr. Lori Scott
Augustana College, Rock Island Illinois

Bennett Hartmann
Bio-375
Dr. Scott
2/4/16

Valine Biosynthesis: Mrub_2994 is Orthologous to E. coli b3770 and Mrub_1844 is Orthologous to E. coli b3771

Background

The purpose of this project is to better understand Meiothermus ruber and its pathway of valine synthesis. Valine is an amino acid that is synthesized in plants and microorganisms from pyruvic acid. Fowl and mammals cannot synthesize valine and it is, therefore an essential amino acid, so they depend on plants and microorganisms that synthesize valine as dietary sources(Valine 2015). Valine is an extremely hydrophobic amino acid. Therefore valine is usually found in the interior of globular proteins to build three dimensional structure (Valine 2015). Figure 1 shows the chemical structure of valine (Valine 2016). Valine is an important amino acid and for these reasons understanding valine biosynthesis is of great interest.

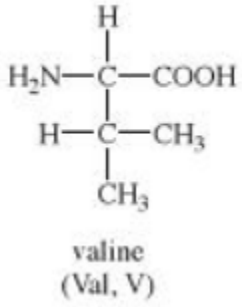


Figure 1. The chemical structure of valine.
Image obtained from Encylopedia Brittanica
(http://www.britannica.com/science/valine).

In order to do this we need to compare it to the model organism *Escherichia coli*. *E. coli* is a very well studied organism and a lot is known about the pathways which it uses(Cooper 2000). Since M. ruber is not a well studied organism it needs to be determined if it can be compared to *E. coli* (Geni-science 2015). In order to do this several bioinformatics programs will be used to analyze the protein product sequence of these genes. There is currently a large amount of bioinformatics information available on databases that has not been analyzed yet that can be used to compare these two genes. These programs and databases will compare thousands of sequences and determine the best matches, and provide E-values for each result. The E-values provided are very important statistics to consider when interpreting data. A low e-value means that it is very unlikely that the sequences are similar by chance. In general a lower e-value means it is more likely that there is an evolutionary relationship between two sequences, and that they may be orthologs. Knowledge in understanding and using bioinformatics tools in the future is very important. With advanced genome sequencing power that is available today the genetic sequence of everything can be easily studied. Bioinformatics will not just be limited to genetics research, but will be used in the clinical setting. Custom and personalized medical treatment will be available based on genetic information(Bayat 2002). So understanding the use of bioinformatics tools is important heading into the future.

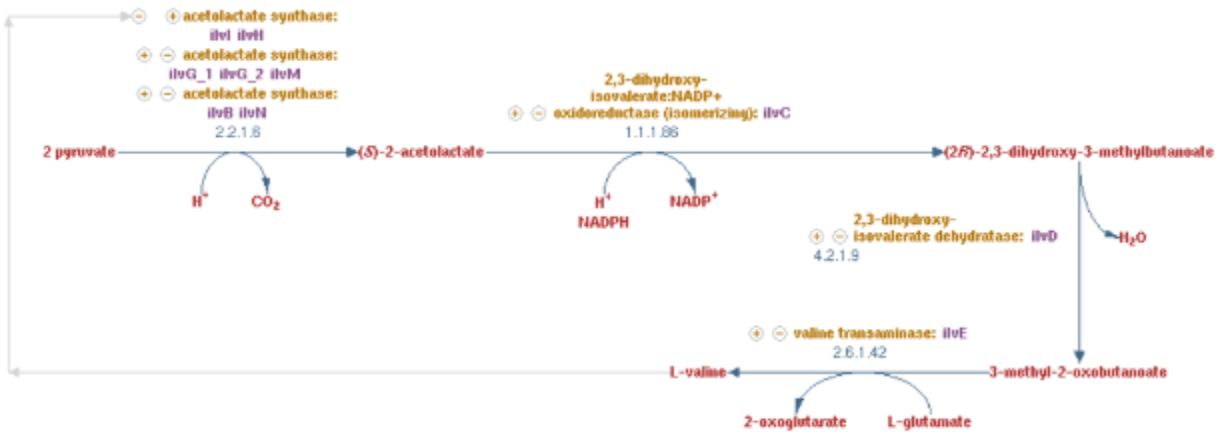


Figure 2. Pathway of L-valine biosynthesis in E. coli. Pathway map obtained from Ecocyc (http://ecocyc.org/ECOLI/NEW-IMAGE?type=PATHWAY&object=VALSYN-PWY).

The known pathway for valine synthesis in *E. coli* is shown in Figure 2. The first gene investigated in this study is Mrub_2994, which could possibly be an ortholog of *E. coli* b3770. The E coli b3770(E.C. 2.6.1.42) gene product ilvE has been purified to homogeneity(Lee-Peng 1979). The enzyme structure was determined to have a molecular weight of roughly 182,000 and to be a hexamer with identical subunits(Lee-Peng 1979). The gene product ilvE was found to be a branched-chain amino acid aminotransferase(Lee-Peng 1979). The other M. ruber gene that will be studied is Mrub_1844, which is a suspected ortholog of *E. coli* b3771. Research into *E. coli* b3771 found that the purified product is dihydroxy-acid dehydratase with a subunit molecular weight of 6,000. This enzyme, called ilvD in *E. coli*, was found to contain a [4Fe-4S]2+ cluster linked to the active site. Resonance studies on this cluster suggests that is directly involved in catalysis(Flint 1993).

One of the reasons that M. ruber was the organism chosen to be studied is because it comes from a poorly studied section of the Tree of Life. Since little is known

about the organisms in this part of the tree, any research will greatly increase the scientific community's knowledge.The Genomic Encyclopedia of Bacteria and Archaea is currently sequencing genomes from bacteria and archaea to try and fill in the Tree of Life. There is a major gap in the knowledge of microbial genomes and metabolisms and research into less studied organisms will begin to fill in this gap (Wu 2009). The question that this specific project is trying to answer is if Mrub_2994 and Mrub_1844 are the M. ruber version of *E. coli* genes. Since *E. coli* is well studied we can compare the M. ruber gene and protein product sequence to *E. coli*, to determine if it is an orthologs. My hypothesis is that Mrub_2994 is indeed an ortholog of *E. coli* b3770 and that Mrub_1844 is an ortholog of *E. coli* b3771.

Methods

The GENI-ACT lab notebook protocol was used in this study. A list of bioinformatics programs with descriptions is available at http://geni-act.org/education/main (GENI-ACT). Deviations from the standard protocol include using the bioinformatics program ecocyc was used instead of metacyc, these are similar programs but this deviated from the set protocol. The gene context maps generated by IMG/EDU were colored by Kegg. Additionally *E. coli* was BLAST'ed against the entire M. ruber genome to determine if any *E. coli* gene was related to M. ruber, which was not in the set protocol of the GENI-ACT notebook.

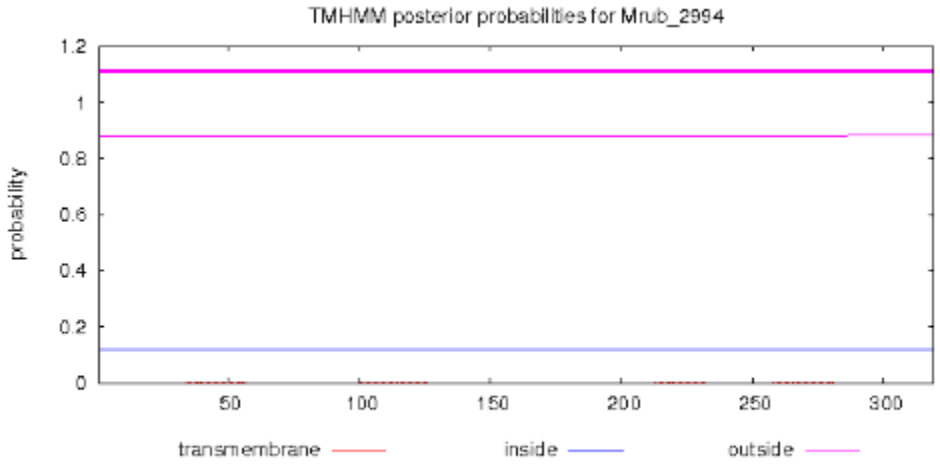
Results

Table 1. E. coli b3770 and Mrub_2994 are orthologs

Description of evidence collected	E. coli(b3770)	M. ruber(Mrub_2994)
Cellular localization	Cytoplasm	
Blast *E. coli* against *M. ruber*	Score: 232 bits ; E-value: 1e-93	
Pfam-protein family	PF01063(Amino-transferase class IV)	
Pfam E-values	E=3.9e-51	E=4.4e-48
CDD (COG category)	COG0115 (Branched-chain amino acid aminotransferase)	
CDD E-values	E=4.63e-108	E=3.93e-96
TIGRfam-protein family	TIGR01122(ilvE_I: Branched-chain amino acid aminotransferase	
TIGRfam E-values	E=7.1e-237	E=1.7e-169
E.C. number	E.C. 2.6.1.42 ; Branched-chain-amino-acid-transaminase	
PDB	1A3G (Branched-chain amino acid aminotransferase from Escherichia coli)	
PDB E-values	E=0.0	E=3.68427e-76

The overview of results tabulated from the GENI-ACT protocol are shown in Table 1. The results of a protein Blast comparing Mrub_2994 and E. coli b3770 gave a bit score of 232 and an E-value of 1e-93. This is a small E-value which represents that it is very likely that these two sequences are evolutionarily related. Table 1 shows information compiled from several bioinformatics programs. The first important piece of information regarding the protein products of these genes is cellular location. TMHMM was used to predict the number of transmembrane helices in the protein products for Mrub_2994 and E. coli b3770. The TMHMM hydropathy plot results are shown in Figure 3.

Panel A



Panel B

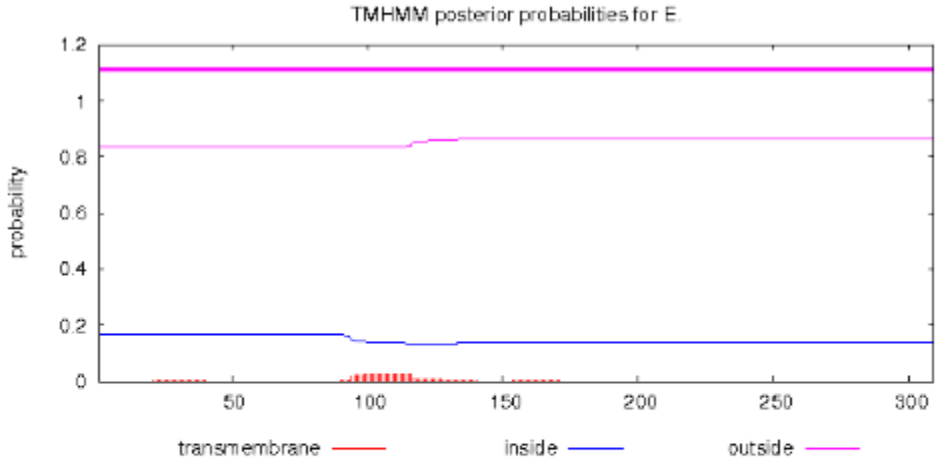
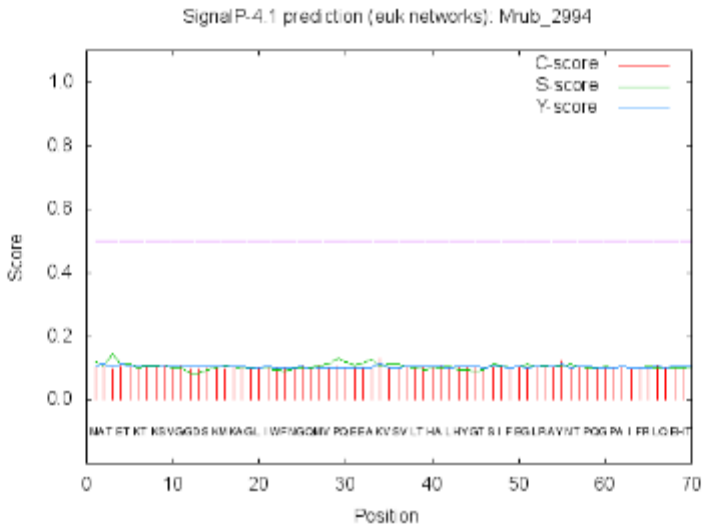


Figure 3. Mrub_2994 and E. coli b3770 do not contain TMH regions; therefore it is predicted that the protein products are located in the cytoplasm. Panel A= Mrub_2994 ; Panel B=E. coli b3770. TMHMM Server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) created this hydropathy plot.

Since there are no predicted transmembrane helices, on the protein products, it signifies that the protein likely resides entirely within the cytoplasm of the cell. Additionally Figure 4. from SignalP shows that neither Mrub_2994 or *E. coli* b3770 contain any cleavage sites.To further verify this cellular location, PSORTb was used on the protein sequences. PSORTb gave a cytoplasmic score of 9.97 to Mrub_2994 and a

cytoplasmic score of 9.97 for *E. coli* b3770. The PSORTb scores have a maximum of 10, so from these results both proteins are expected to be found in the cytoplasm.

Panel A



Panel B

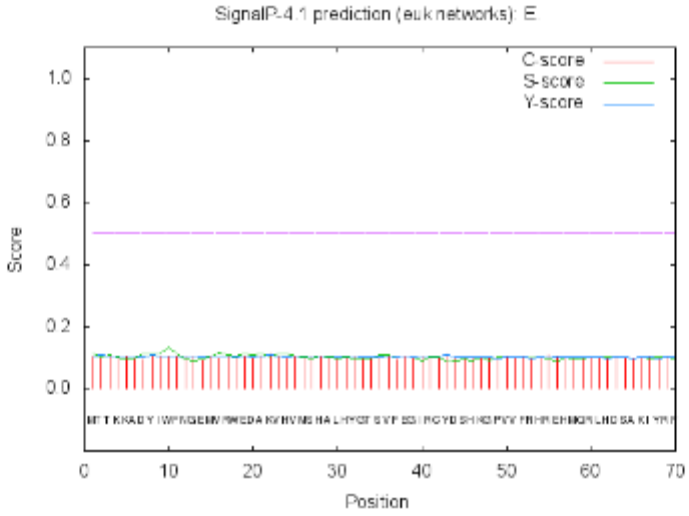
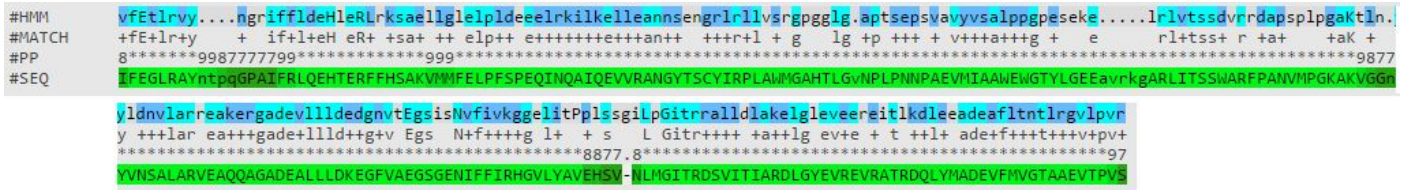


Figure 4. Mrub_2994 and E. coli b3770 do not contain any cleavage sites. Panel A: Mrub_2994 ; Panel B: E. coli b3770. Figure generated by SignalP 4.1 Server (http://www.cbs.dtu.dk/services/SignalP/).

The bioinformatics program Pfam is used to locate domains in the proteins. Using Pfam, it was determined both Mrub_2994 and *E. coli* b3770 have the same

protein domain, PF01063. PF01063 is an amino-transferase class IV protein family. This result was expected and verified with the E-values(4.4e-48 for Mrub_2994 and 3.9e-51 for *E. coli* b3770). The pairwise alignments with PF01063 for both Mrub_2994 and *E. coli* b3770 are shown in Figure 5. The results show that both are AA_Kinase and share highly conserved amino acid residues such as E3 and H22. This supports the hypothesis because sharing highly conserved residues is indicative that they may be orthologs. Additionally CDD was used to identify both Mrub_2994 and *E. coli* b3770 as part

Panel A



Panel B

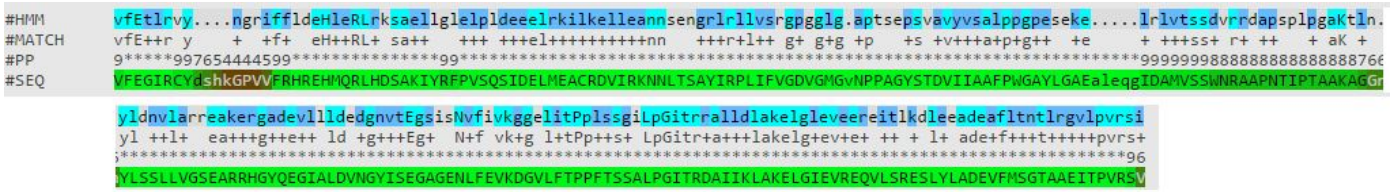


Figure 5. Pfam pairwise alignment shows Mrub_2994 and E. coli b3770 are both Amino-transferase and share highly conserved amino acid residues such as E3 and H22. Panel A: Mrub_2994 ; Panel B: E. coli b3770. Pairwise alignment generated by Pfam(http://pfam.xfam.org/search).

of the same COG group. COG groups are clusters of orthologous groups. A significant hit represents that the subject sequence is part of an ortholog set. Both genes had a significant hit with COG0115, branched-chain amino acid aminotransferase(E-values of 3.93e-96 for Mrub_2994 and 4.63e-108 for *E. coli* b3770).

The low e-values of represent that there is almost no possibility that the sequences lined up by chance, and are indeed evolutionarily related. This result gives strong evidence that Mrub_2994 and *E. coli* b3770 are orthologs. Another bioinformatics program that was used to compare these two genes was TIGRFAM. TIGRFAM uses a library of protein families and sequences to predict the name and function of a gene product. TIGRFAM results showed both Mrub_2994 and *E. coli* b3770 belonged to TIGR01122. Both genes had a significant match to ilvE_I: Branched-chain amino acid aminotransferase. (E-values of 1.7e-169 and 7.1e-237 respectively). The enzyme commision number was determined using ExPASy ENZYME. Both Mrub_2994 and *E. coli* b3770 returned the same Enzyme Commision (E.C.) number of 2.6.1.42. The Protein Data Bank (PDB) results concluded that both proteins returned the same PDB code of 1A3G. This corresponds to branched-chain-amino acid aminotransferase from Escherichia coli. Mrub_2994 and *E. coli* b3770 had E-values of 3.68427e-79 and 0.0 respectively. The e-value of 0.0 was expected for *E. coli* b3770 because the protein sequence is from *E. coli*. The e-value from Mrub_2994 represents a close relationship between the proteins. The small e-value corresponds to a high likelihood of an evolutionary relationship between the two proteins.

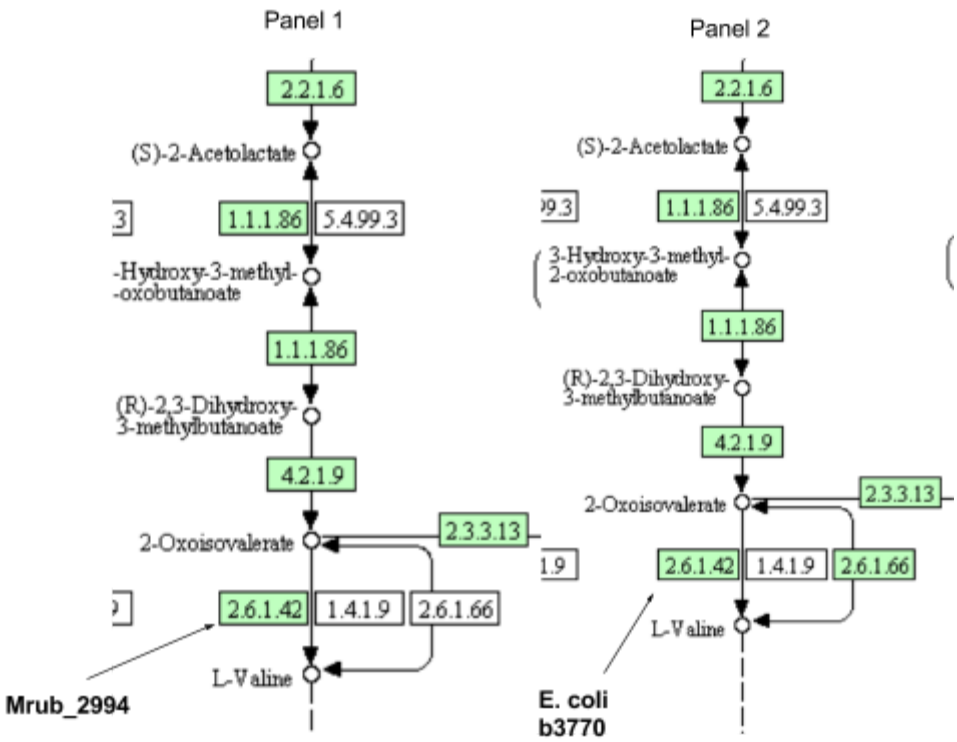


Figure 6. Kegg pathways show that Mrub_2994 and E. coli b3770 are both used in the same pathway step of valine synthesis. Panel 1: Mrub_2994 ; Panel 2: E. coli b3770. Pathways generated by Kegg Pathway Database(http://www.genome.jp/kegg/pathway.html).

Figure 6 shows the Kegg pathway maps for M. ruber and *E. coli* valine synthesis. The enzymes shown in green are the enzymes used by the respective organisms. The enzyme commision number 2.6.1.42 can be used to identify which enzymes are Mrub_2994 and *E. coli* b3770. Both are involved in the same step of the valine biosynthesis pathway. The similar pathway of valine synthesis and the use of the respective enzymes in the same step represents a strong likelihood that the enzymes are related evolutionarily.

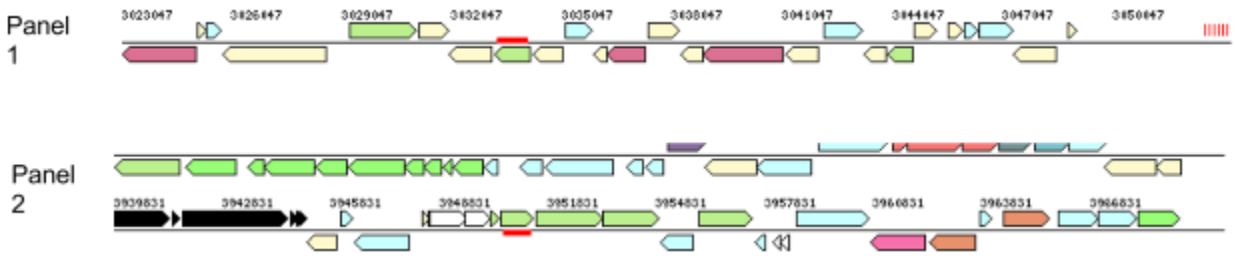


Figure 7. Mrub_2994 does not appear to be involved in an operon. E. coli b3770 appears to be involved with an operon but the context map for Mrub_2994 does not suggest any operon involvement. Panel 1=Mrub_2994 ; Panel 2= E. coli b3770. Gene context map generated by IMG/EDU (https://img.jgi.doe.gov/cgi-bin/edu/main.cgi).

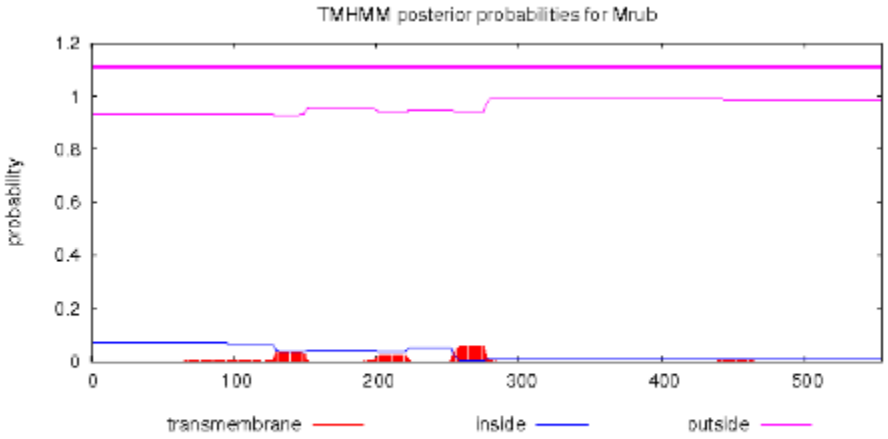
Figure 7 shows the gene context maps for Mrub_2994 and *E. coli* b3770. The gene marked with the red line indicates the particular gene of interest. Mrub_2994 has two nearby genes facing the same direction, however they are a different color which means that are involved with a different category of cellular function. *E. coli* b3770 has several genes facing the same direction that are of the same color, which suggests that it may be involved in an operon with these genes. However this operon information alone does not refute the hypothesis that these genes are orthologs.

Table 2. E. coli b3771 and Mrub_1844 are orthologs

Description of evidence collected	E. coli(b3771)	M. Ruber(Mrub_1844)
Cellular localization	Cytoplasm	
Blast *E. coli* against *M. ruber*	Score: 372 bits ; E-value 3e-121	
Pfam-protein family	PF00290 (Dehydratase Family)	
Pfam E-values	E=9.6e-219	E=3.1e-199
CDD (COG category)	COG0129(Dihydroxyacid dehydratase/phosphogluconate dehydratase)	
CDD E-values	E=0.0	E=0.0
TIGRfam-protein family	TIGR00110(ilvD: dihydroxy-acid dehydratase)	
TIGRfam E-values	E=0.0	E=7.5e-258
E.C. number	4.2.1.9	
PDB	2GP4 (6-phosphogulconate dehydratase from shewanella ondeidensis)	
PDB E-values	E=8.45877e-47	E=1.23489e-56

The overview of results tabulated from the GENI-ACT protocol are shown in Table 2. The results of a protein blast comparing Mrub_1844 and *E. coli* b3771 gave a bit score of 373 and an E-value of 3e-121. This is a small E-value which represents that it is very likely that these two sequences are evolutionarily related. Table 2 shows information compiled from several bioinformatics programs. The first important piece of information regarding the protein products of these genes is cellular location. TMHMM was used to predict the number of transmembrane helices in the protein products for Mrub_1844 and *E. coli* b3771. The TMHMM hydropathy plot results are shown in Figure 8.

Panel A



Panel B

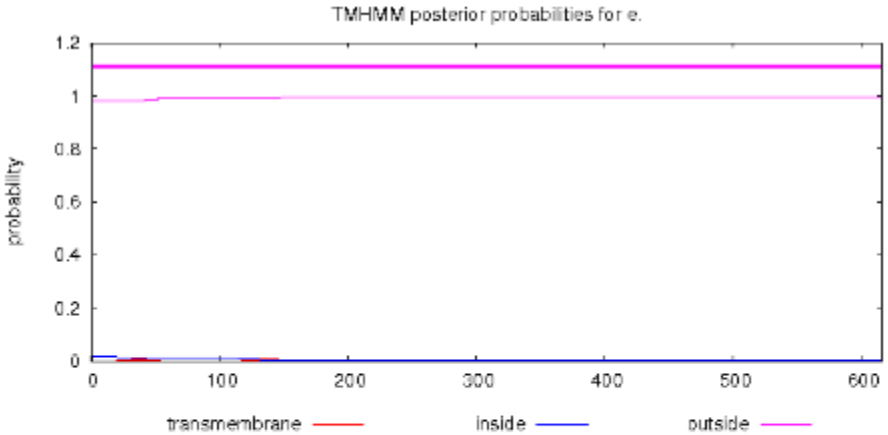


Figure 8. Mrub_1844 and E. coli b3771 do not contain TMH regions; therefore it is predicted that the protein products are located in the cytoplasm. Panel A= Mrub_1844 ; Panel B=E. coli b3771. TMHMM Server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) created this hydropathy plot.

Since there are no predicted transmembrane helices, on the protein products, it signifies that the protein likely resides entirely within the cytoplasm of the cell. Additionally Figure 9 from SignalP shows that neither Mrub_1844 or *E. coli* b3771 contain any cleavage sites. To further verify this cellular location, PSORTb was used on the protein sequences. PSORTb gave a cytoplasmic score of 9.97 to Mrub_1844 and a cytoplasmic score of 9.97 for *E. coli* b3771. The PSORTb scores have a maximum of 10, so from these results both proteins are expected to be found in the cytoplasm.

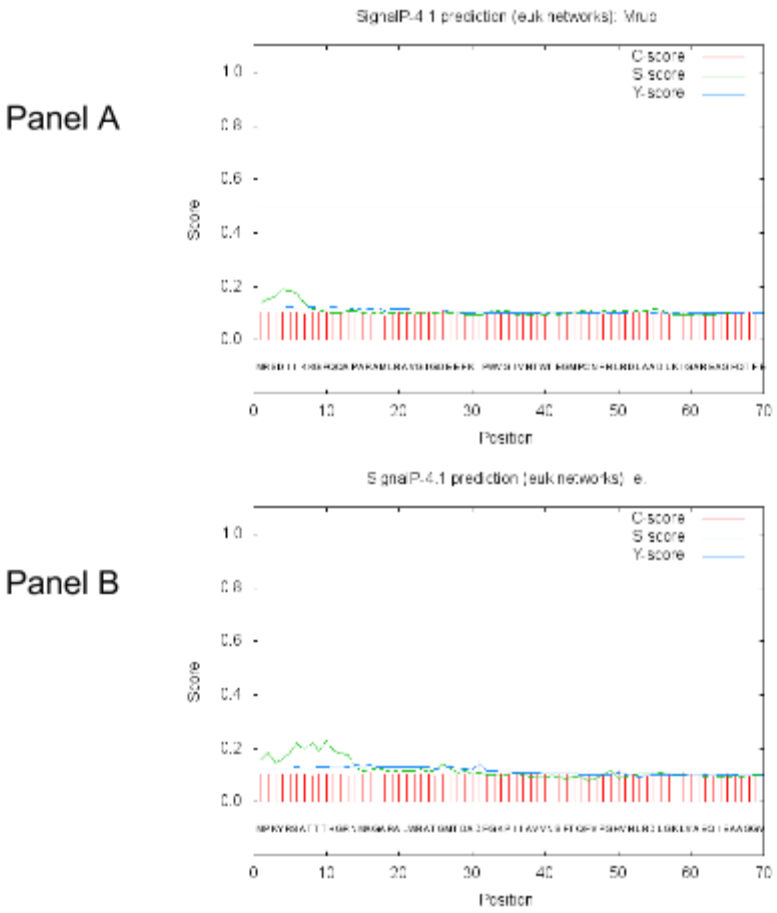


Figure 9. Mrub_1844 and E. coli b3771 do not contain any cleavage sites. Panel A: Mrub_1844 ; Panel B: E. coli b3771. Figure generated by SignalP 4.1 Server (http://www.cbs.dtu.dk/services/SignalP/).

The bioinformatics program Pfam is used to locate domains in the proteins. Using Pfam, it was determined both Mrub_1844 and *E. coli* b3771 have the same protein domain, PF00290. PF00290 is in the dehydratase protein family. This result was expected and verified with the E-values(3.1e-199 for Mrub_1844 and 9.6e--219 for *E. coli* b3771). The pairwise alignments with PF00290 for both Mrub_1844 and *E. coli* b3771 are shown in Figure 10. The results show that both are dehydratase and share highly conserved amino acid residues.

Panel A

Panel B

Figure 10. Pfam pairwise alignment shows Mrub_1844 and E. coli b3771 are both Amino-transferase and share highly conserved amino acid residues. Panel A: Mrub_1844 ; Panel B: E. coli b3771. Pairwise alignment generated by Pfam(http://pfam.xfam.org/search).

Additionally CDD was used to identify both Mrub_1844 and *E. coli* b3771 as part of the same COG group. COG groups are clusters of orthologous groups. A significant hit represents that the subject sequence is part of an ortholog set. Both genes had a significant hit with COG0129,dihydroxyacid dehydratase/phosphogluconate dehydratase(E-values of 0.0 for Mrub_1844 and 0.0 for *E. coli* b3771). The extremely low e-values represent that there is almost no possibility that the sequences lined up by chance, and are indeed evolutionarily related. This result gives strong evidence that Mrub_1844 and *E. coli* b3771 are orthologs. Another bioinformatics program that was used to compare these two genes was TIGRFAM. TIGRFAM uses a library of protein families and sequences to predict the name and function of a gene product. TIGRFAM

results showed both Mrub_1844 and *E. coli* b3771 belonged to TIGR00110. Both genes had a significant match to ilvD: dihydroxy-acid dehydratase. (E-values of 7.5e-258 and 0.0 respectively). The enzyme commision number was determined using ExPASy ENZYME. Both Mrub_1844 and *E. coli* b3771 returned the same Enzyme Commision (E.C.) number of 4.2.1.9. The Protein Data Bank (PDB) results concluded that both proteins returned the same PDB code of 2GP4. This corresponds to 6-phosphogluconate dehydratase from shewanella ondeidensis. Mrub_1844 and *E. coli* b3771 had E-values of 1.23489e-56 and 8.45877e-47 respectively. The e-value from Mrub_1844 represents a close relationship between the proteins. The small e-values correspond to a high likelihood of an evolutionary relationship between the two proteins.

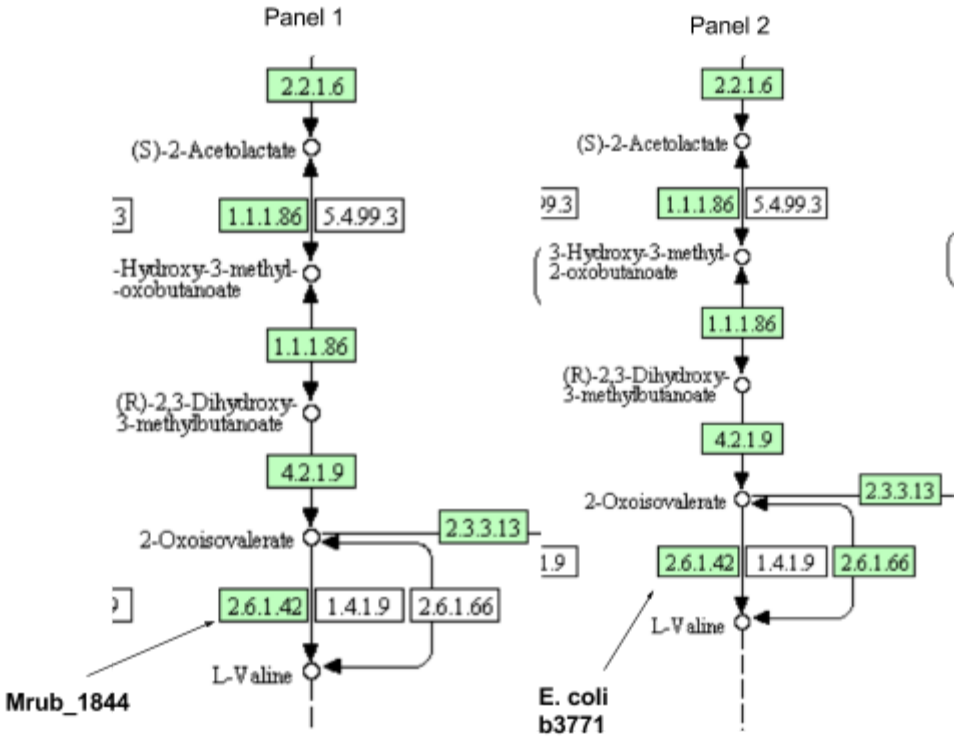


Figure 11. Kegg pathways show that Mrub_1844 and E. coli b3771 are both used in the same pathway step of valine synthesis. Panel 1: Mrub_1844 ; Panel 2: E. coli b3771. Pathways generated by Kegg Pathway Database(http://www.genome.jp/kegg/pathway.html).

Figure 11 shows the Kegg pathway maps for M. ruber and *E. coli* valine synthesis. The enzymes shown in green are the enzymes used by the respective organisms. The enzyme commision number 4.2.1.9 can be used to identify which enzymes are Mrub_1844 and *E. coli* b3771. Both are involved in the same step of the valine biosynthesis pathway. The similar pathway of valine synthesis and the use of the respective enzymes in the same step represents a strong likelihood that the enzymes are related evolutionarily.

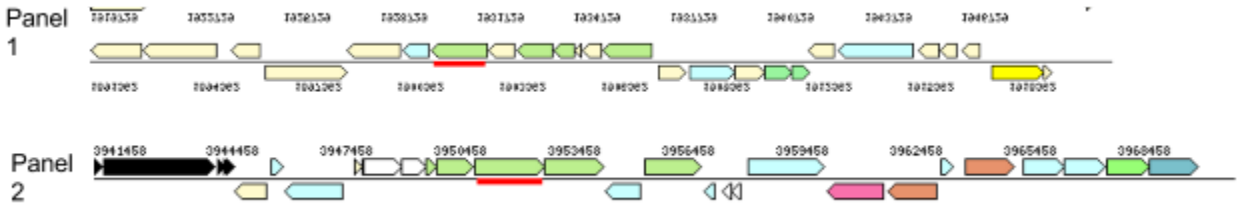


Figure 12. Mrub_1844 and E. coli b3771 are not involved with operons of similar function. E. coli b3771 appears to be involved with an operon and the context map for Mrub_1844 suggests there also may be operon involvement but with different genes. Panel 1=Mrub_1844 ; Panel 2= E. coli b3771. Gene context map generated by IMG/EDU (https://img.jgi.doe.gov/cgi-bin/edu/main.cgi).

Figure 12 shows the gene context maps for Mrub_1844 and *E. coli* b3771. The gene marked with the red line indicates the particular gene of interest. Mrub_1844 has many nearby genes facing the same direction, some of which are the same color. This suggests that Mrub_1844 may be involved in an operon with these genes. *E. coli* b3771 has several directly neighboring genes that are facing the same direction and are the same color. The same color represents that all the genes are involved with the same type of cellular process. This is highly indicative of operon involvement for *E. coli* b3771.

Conclusion

Based on the results obtained in this study it can be concluded that Mrub_2994 is indeed an ortholog of *E. coli* b3770 and that Mrub_1844 is an ortholog of *E. coli* b3771. Analysis of both genes and protein products produced similar results and strong evidence that there is an evolutionary relationship between them. Based on TMHMM results both Mrub_2994 and *E. coli* b3770 were predicted to be located in the cytoplasm. Both Mrub_1844 and *E. coli* b3771 were also determined to be located in the cytoplasm. Both pairs of proteins were given the same E.C. numbers of 2.6.1.42 and 4.2.1.9 and have the same role in valine synthesis. This was determined from the CDD and TIGRFAM results, which both identified the sequence as branched-chain amino acid transferase for Mrub_2994 and *E. coli* b3770, and dihydroxy-acid dehydratase for Mrub_1844 and *E. coli* b3771. The PFAM results showed highly conserved amino acid residues between the two gene pairs. Additionally the gene context of Mrub_1844 and *E. coli* b3771 was not very similar. Different genes flank both of these genes as shown in Figure 12. The only results that conflicted are the gene context maps for Mrub_2994 and *E. coli* b3770 which were shown in Figure 7. In these results it appeared that b3770 was involved in an operon but the gene context for Mrub_2994 did not suggest operon involvement. Despite the gene context result, the results of every other analysis suggested that the two genes are orthologs. All of these results together paint a clear picture that Mrub_2994 and *E. coli* b3770 as well as Mrub_1844 and *E. coli* b3771 are evolutionarily related. Mrub_2994 is in fact the M. ruber b3771 gene and Mrub_1844 is in fact the M. ruber b3771 gene.

Sources

Bayat A. Bioinformatics. BMJ : British Medical Journal. 2002;324(7344):1018-1022.

Cooper GM. The Cell: A Molecular Approach. 2nd edition. Sunderland (MA): Sinauer Associates; 2000. Cells As Experimental Models. Available from: http://www.ncbi.nlm.nih.gov/books/NBK9917/

Flint DH, EMptage MH, et al. The role and properties of the iron-sulfer cluster in Escherichia coli dihydroxy-acid dehydratase. Journal of Biological Chemistry. 1993;268(20):14732-42.

Geni-science.[Internet] 2015. Meiothermus ruber genome Analysis Project. Available from: http://www.geni-science.org/secure/projects/view/

Lee-Peng F-C, Hermodson MA, Kohlhaw GB. Transaminase B from *Escherichia coli*: Quaternary Structure, Amino-Terminal Sequence, Substrate Specificity, and Absence of a Separate Valine-α-Ketoglutarate Activity. *Journal of Bacteriology.* 1979;139(2):339-345.

U.S. Department of Energy Joint Genome Institute. [2016 Feb 5]. Available from: http://jgi.doe.gov/our-science/science-programs/microbial-genomics/phylogenetic-diversity/ [accessed 7-29-14]

Valine. [Internet]. 2016. *Encyclopædia Britannica*; [2016 Feb 5]. Available from: http://www.britannica.com/science/valine

Valine. [Internet]. 2015. National Cancer Institute; [2016 Feb 5]. Available from: https://ncit.nci.nih.gov/ncitbrowser/ConceptReport.jsp?dictionary=NCI_Thesaurus&ns=NCI_Thesaurus&code=C29604

Wu D, Hugenholtz P, Mavromatis K, et al. A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature.* 2009;462(7276):1056-1060.