2017

Mrub_1867, Mrub_1868, and Mrub_1869 genes are predicted orthologs of the b2279, b2280, and b2281 genes found in *Escherichia coli* coding for the NADH dehydrogenase subunits K, J, and I respectively

Wade Smith
Augustana College, Rock Island Illinois

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

Mrub_1867, Mrub_1868, and Mrub_1869 genes are predicted orthologs of the b2279, b2280, and b2281 genes found in *Escherichia coli* coding for the NADH: ubiquinone oxidoreductase subunits K, J, and I respectively

Wade M. Smith
Dr. Lori Scott Laboratory
Biology Department, Augustana College
639 38th Street, Rock Island, IL 61201

INTRODUCTION

What is *Meiothermus ruber*?

Meiothermus ruber (*M. ruber*) is a gram-negative, obligately aerobic and thermophilic bacterium that is red-pigmented (Tindall et al., 2010). *M. ruber* grows within a temperature range of 35-70°C which is similar to the temperature ranges of the hot springs in the Russian city of Kamchatka where bacteria from the *Meiothermus* genus were first isolated (Loginova et al., 1984). The complete genome of *M. ruber* was sequenced in 2010 as a part of the Genomic Encyclopedia of Bacteria and Archaea project and 71.79% of the genes sequenced were identified to have a predicted function (Tindall et al., 2010). This recent sequencing of the *M. ruber* genome allows investigation into its genes and what proteins they specifically encode for. Since *M.* ruber is a relatively poorly studied organism, there is quite a bit of information missing in regards to its genome and the roles that these genes provide within the organism. Investigation into *M. ruber* could provide new information for genes or variants of processes that are not present in other well studied organisms. Therefore, *M. ruber* could prove to be a valuable organism to study in regards to helping expand our knowledge of this species and possibly many others.

E. coli as a model organism

Escherichia coli (E. coli) is a gram-negative, rod shaped bacteria that is commonly found in the lower gastrointestinal tract of mammals and it is one of the most extensively studied organisms in the field of biology due to its ability to grow/reproduce rapidly and its ability to survive in various growth conditions (Keseler et al., 2013). *E.* coli can also be genetically manipulated very easily due to its well understood genome which makes it a valuable model organism to help fill in missing gaps in knowledge about other species. Specifically, the process of oxidative phosphorylation and the various genes that encode for proteins in this process can be studied in *E. coli* to help make inferences about other species such as *M. ruber*.

Oxidative Phophorylation

As shown in figure 1, the process of oxidative phosphorylation in *E. coli* serves to transfer electrons from NADH to O_2 through a series of protein complexes called NADH: ubiquinone oxidoreductase (I), succinate dehydrogenase (II), and cytochrome c oxidase (IV) (Kanehisa et al., 2016). The transfer of electrons through this chain serves to provide the reducing power to translocate H+ protons across the membrane to generate an electrochemical gradient which allows for the generation of ATP from ATP synthase (Berg et al., 2012). This process is extremely important to investigate in order to better understand how organisms generate ATP and other forms of energy. Specifically, our focus was on the NADH: ubiquinone oxidoreductase complex (complex I) in *M. ruber* which is composed of 46 subunits both in and outside of the cytoplasmic membrane (Berg et al., 2012). NADH: ubiquinone oxidoreductase transfers protons from NADH through a series of iron-sulfur clusters to ubiquinone (Q) which results in the pumping of 4 protons across the membrane (Berg et al., 2012). Numerous genes in *E. coli* are involved in encoding proteins for subunits of this complex but we focused on only three of these genes: *NuoI*, *NuoJ*, and *NuoK*. As shown in figure 2, all three of these genes appear to be part of an operon identified to participate in carbohydrate metabolism by the Joint Genome Institute (Markoqitz et al., 2012). BLAST searches conducted for these three genes indicated that genes with the locus tags Mrub_1869, Mrub_1868, and Mrub_1867 share similar protein sequences respectively to those in *E. coli*. This may be an indicator that these genes are orthologous to one another. Since the NADH ubiquinone oxidoreductase complex is the first step in this important process for energy production in organisms, it may be important to study in *M. ruber* to identify if there are any differences compared to *E. coli* that could have resulted from evolutionary adaptations to the stressful environment that *M. ruber* thrives in.

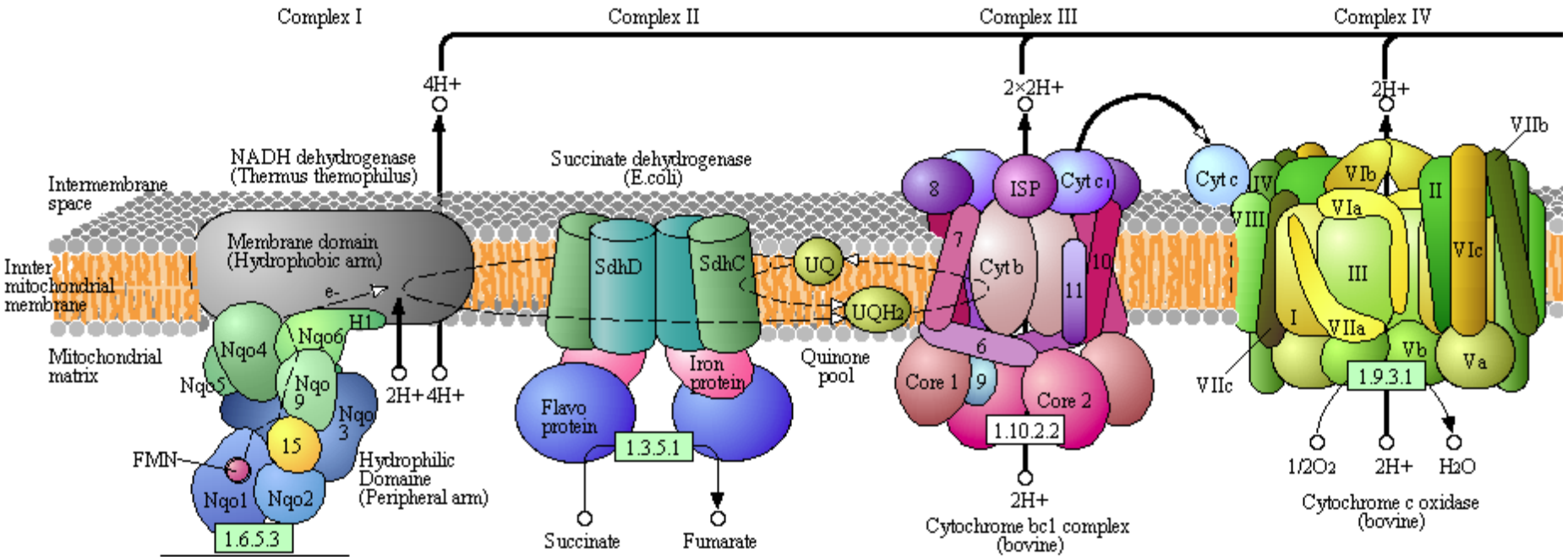


Figure 1: Protein complexes involved in the electron transport chain pathway in *E. coli* (highlighted in green). NADH: ubiquinone reductase in depicted on the far left and is the first step in the electron transport chain. Image taken from the KEGG Pathway Database (Kanehisa et al., 2016).

Figure 2: Genes *NuoI, NuoJ*, and *NuoK* are part of an operon identified to participate in carbohydrate metabolism in *E. coli*. Since these genes are predicted to have the same function and are transcribed together in the same direction, they are likely part of an operon. Image taken from Ecocyc website (Keseler et al., 2013).

Purpose/Hypothesis

The purpose of this project is to identify if the *E. coli* genes of *NuoI, NuoJ,* and *NuoK* with the respective locus tags of b2281, b2280, and b2279 are orthologous to the genes with respective locus tags of Mrub_1869, Mrub_1868, and Mrub_1867 in *M. ruber* by using various bioinformatics programs. We hypothesize that all of these three genes are orthologous to one another based on preliminary BLAST searches of the *E. coli* genes that yielded similar protein sequence matches to the genes listed in *M. ruber* with extremely low E-values. These low E-values indicate that these protein sequences are highly conserved between the two species and that these proteins likely have similar functions to one another in both species.

Methods:

The first step taken in this study was choosing a metabolic pathway. Numerous pathways were available on the KEGG database and oxidative phosphorylation was chosen since this pathway is known to exist in both *E. coli* and *M. ruber* due to its crucial function in energy production and due to the fact that *M. ruber* is an obligate aerobe (Tindall et al., 2010). Genes for this system were identified from the respective KEGG maps for *E. coli* and *M. ruber* after the maps showed that there were no significant differences between the two species (Kanehisa et al., 2016). Three genes from *E. coli* (b2279, b2280, b2281) were chosen and their respective predicted orthologs in *M. ruber* (Mrub_1867, Mrub_1868, and Mrub_1869) were chosen from KEGG since they were labelled with the same gene names within the listed subunits of the NADH dehydrogenase complex. These locus tags were then compared to a complete list of locus tags for each organism in GENI-ACT (http://www.geni-act.org/) to verify that they had not been previously annotated. Preliminary BLAST comparisons were conducted for each pair of predicted orthologs to verify that they had similar sequences and to serve as a basis of investigating these genes (Madden et al., 2002). After it was determined that the predicted ortholog pairs shared similarity, the following bioinformatics tools were used following instructions through the GENI-ACT platform (http://www.geni-act.org/): NCBI's BLAST (Madden et al., 2002), CDD (Marchler-Bauer et al., 2016), T-coffee (Notredame et al., 200), WebLogo (Crooks et al., 2004), TMHMM (Krogh et al., 2016), SignalP 4.0 (Petersen et al., 2011), LipoP (Juncker et al., 2003), PSORT-B v3.0 (Yu et al., 2010), Phobius (Kall et al., 2004), IMG (Markowitz et al., 2012), TIGRFAM (Haft et al., 2001), Pfam (Finn et al., 2016), PDB (Berman et al., 2000), MetaCyc (Caspi et al., 2014), and Phylogeny.fr (Dereeper et al., 2008). Annotation data was collected within the GENI-ACT online notebook.

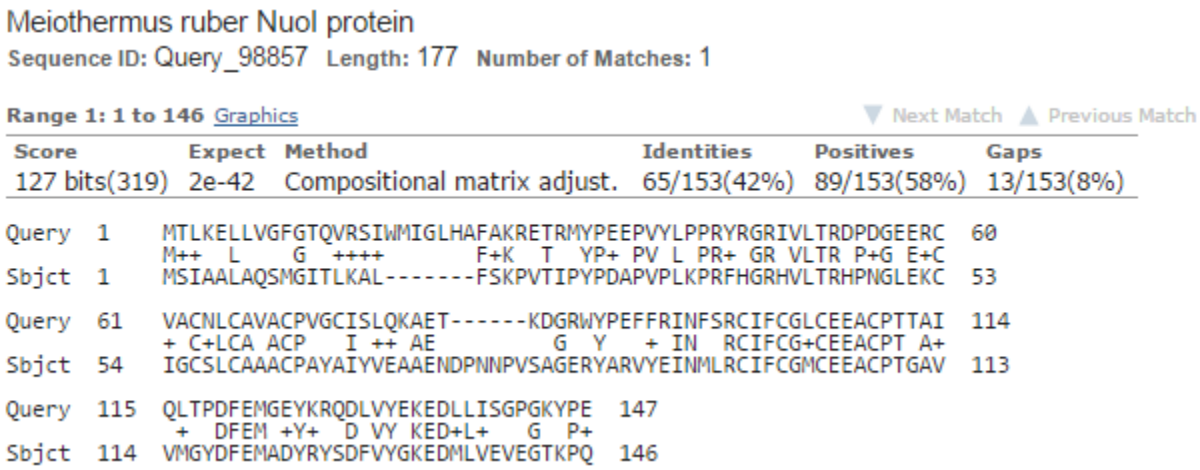
Results

Table 1 provides a summary of the various bioinformatics programs used to compare *E. coli* b2281 and Mrub_1869 genes. The first row indicates the BLAST search conducted to determine the similarities between *E. coli* b2281 and Mrub_1869 gene protein products. The proteins produced by the two genes are different in length so the bit score does not provide much information. The most important piece of information is the very small E-value (3e-38) which indicates the probability that these sequences of amino acids in each protein aligned strictly due to chance. Since it is so small, these two proteins share highly conserved sequences and this is a likely indicator of them having similar functions. Data from CDD yielded the same COG number (COG1143) and name which means that that these proteins likely belong to the same family and have similar functions. The significance of this is noted in the extremely small E-values for both proteins indicating that these did not match to the COG number and family by chance. Thus, these genes both code for the same type of subunit of the NADH: ubiquinone oxidoreductase enzyme. All of the tools used to determine the location of these proteins within the cell (TMHMM, SignalP, LipoP, and PSORTB) indicated that both of the proteins are located in the cytoplasm. This is consistent with the understanding that subunit I is located on the peripheral arm of the enzyme complex (Keseler et al., 2013). The similar location of both of these proteins is another piece of evidence indicating these genes may be orthologs. TIGRFAM searches yielded the same number (TIGR01971) and name (NuoI: NADH-quinone oxidoreductase, chain I) for both proteins along with extremely small E-values indicating that these proteins have the same function. Since TIGRFAMs are constructed using full-length protein sequences, this piece of evidence is particularly useful to argue that these two genes are homologous (Haft et al., 2001). Pfam searches yielded the same number (PF12838) and name (4Fe-4S dicluster domain) for both protein queries along with low E-values indicating the most highly conserved domain between these proteins. This is also consistent with expectations since the NADH: ubiquinone oxidoeductase complex is known to transport electrons via the use of Fe and cysteine residues that have sulfur groups (Berg et al., 2012). The protein database also pulled the same crystal structure for both protein queries along with low E-values for both proteins. These searches also yielded the same enzyme commission number. Since the protein database is a worldwide depository of information regarding 3-D structures of proteins and their sequences, these significant matches for both queries are strong evidence that these proteins have the same 3-D structure as well as the same function (Berman et al., 2000). Lastly, both genes were predicted to participate in the production of the same subunit of NADH: ubiquinone oxidoreductase in the oxidative phosphorylation pathway.

Table 1: *E. coli NuoI* and Mrub_1869 are predicted orthologs

Bioinformatics Tool Used	***E. coli* b2281 protein *(NuoI)***	**Mrub_1869 protein**
BLAST *E. coli* against *M. ruber*	Score: 127 bits E-value: 2e-42	
CDD Data (COG match)	COG Number: COG1143 Formate hydrogenlyase subunit 6/NADH:ubiquinone oxidoreductase 23 kD subunit (chain I)	
	E-value: 10e-172	E-value: 7.53e-55
Cellular Localization	Cytoplasm	
TIGRFAM Protein Family	TIGRFAM Number: TIGR01971 TIGRFAM Name: NuoI: NADH-quinone oxidoreductase, chain I	
	E-value: 4.8e-82	E-value: 5.3e-82
Pfam Match	Pfam Number: PF12838 Pfam Name: 4Fe-4S dicluster domain Clan Name: 4Fe-4S Ferredoxins	
	E-value: 1.3e-11	E-value: 2.1e-13
Protein Database	PDB Code: 2FUG PDB Name: Crystal structure of the hydrophilic domain of respiratory complex I from *Thermus thermophilus*	
	E-value: 1.995e-32	E-value: 2.886e-88
EC Number	E.C.1.6.5.3 - NADH:ubiquinone reductase (H^+-translocating)	
KEGG Pathway Map	Oxidative Phosphorylation Pathway	

Figure 3 shows the results of the protein BLAST search done using *E. coli* b2281 as the query sequence to match to *M. ruber*. The data indicates that 42% of the amino acids are an exact match between the two proteins and 58% of the amino acids were similar in character of the R-groups. The extremely low E-value for this match also indicates that these sequences had an extremely low probability of aligning strictly due to chance. This is the first piece of evidence to suggest that these two genes are related and share similar functions.



Meiothermus ruber NuoI protein

Sequence ID: Query_98857 **Length:** 177 **Number of Matches:** 1

Range 1: 1 to 146 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
127 bits(319)	2e-42	Compositional matrix adjust.	65/153(42%)	89/153(58%)	13/153(8%)

```
Query  1    MTLKELLVGFGTQVRSIWMIGLHAFAKRETRMYPEEPVYLPPRYRGRIVLTRDPDGEERC  60
            M++  L    G  ++++       F+K  T  YP+ PV L PR+ GR VLTR P+G E+C
Sbjct  1    MSIAALAQSMGITLKAL-------FSKPVTIPYPDAPVPLKPRFHGRHVLTRHPNGLEKC  53

Query  61   VACNLCAVACPVGCISLQKAET------KDGRWYPEFFRINFSRCIFCGLCEEACPTTAI  114
            + C+LCA ACP   I ++ AE         G  Y   + IN  RCIFCG+CEEACPT A+
Sbjct  54   IGCSLCAAACPAYAIYVEAAENDPNNPVSAGERYARVYEINMLRCIFCGMCEEACPTGAV  113

Query  115  QLTPDFEMGEYKRQDLVYEKEDLLISGPGKYPE  147
             +  DFEM +Y+  D VY KED+L+   G  P+
Sbjct  114  VMGYDFEMADYRYSDFVYGKEDMLVEVEGTKPQ  146
```

Figure 3: *E. coli NuoI* and Mrub_1869 have similar protein sequences. The query sequence is *E. coli* b2281 and the subject sequence is Mrub_1869. The search was conducted using the NCBI BLAST bioinformatics program (Madden et al., 2002).

Figure 4 indicates the TMH hydropathy plots for *E. coli NuoI* and Mrub_1869 proteins. The red peaks indicate the probabilities that transmembrane helices are located near those amino acid positions. While there are several very small red peaks, these are not significant enough to predict the presence of TMH areas. These hydropathy plots are consistent with one another indicating that this protein is likely located in the cytoplasm as opposed to the cellular membrane.

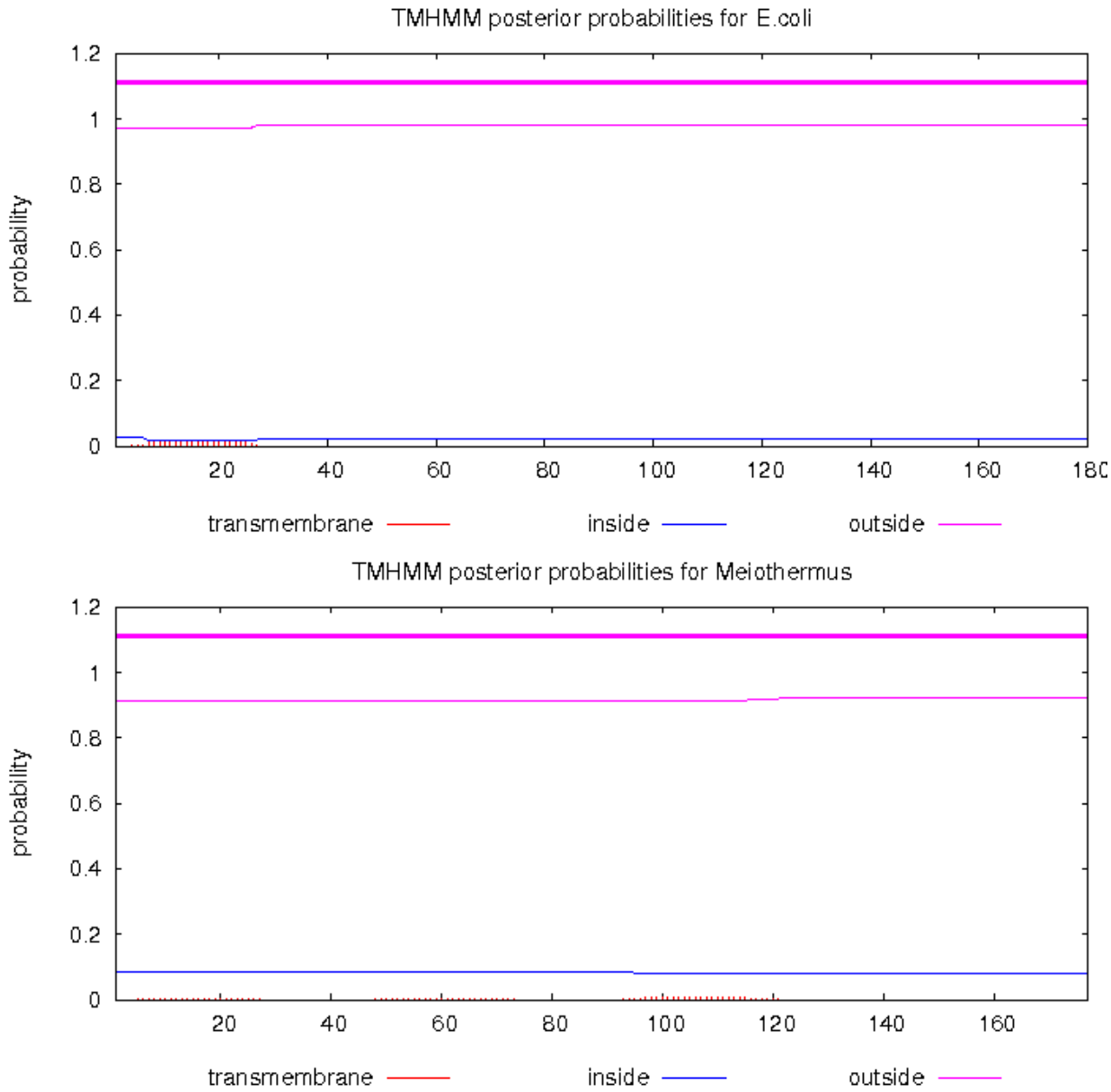


Figure 4: *E. coli NuoI* and Mrub_1869 do not contain any TMH regions indicating that both proteins are likely not associated with the cellular membrane. Top graph: probability plot for *E. coli* b2281 protein. Bottom graph: probability plot for Mrub_1869 protein. Graphs were generated using TMHMM server (Krogh 2016).

Investigation of possible cleavage sites in both *E. coli NuoI* and Mrub_1869 proteins using SignalP yielded no predictions for cleavage sites (Figure 5). This tool assigns each protein a D-value which is calculated using an S-score, Y-score, and a cutoff value to indicate the probability of the protein containing a cleavage site. *E. coli NuoI* had a D-value of 0.167 and Mrub_1869 had a D-value of 0.242, both of which are well below the cutoff value of 0.450 indicating that these proteins do not contain cleavage sites. This supports previous findings that this protein is not associated with the cellular membrane. Findings from LipoP and PSORT-B (cytoplasmic score = 9.12) also indicated that both proteins are located in the cytoplasm (Yu et al., 2010).

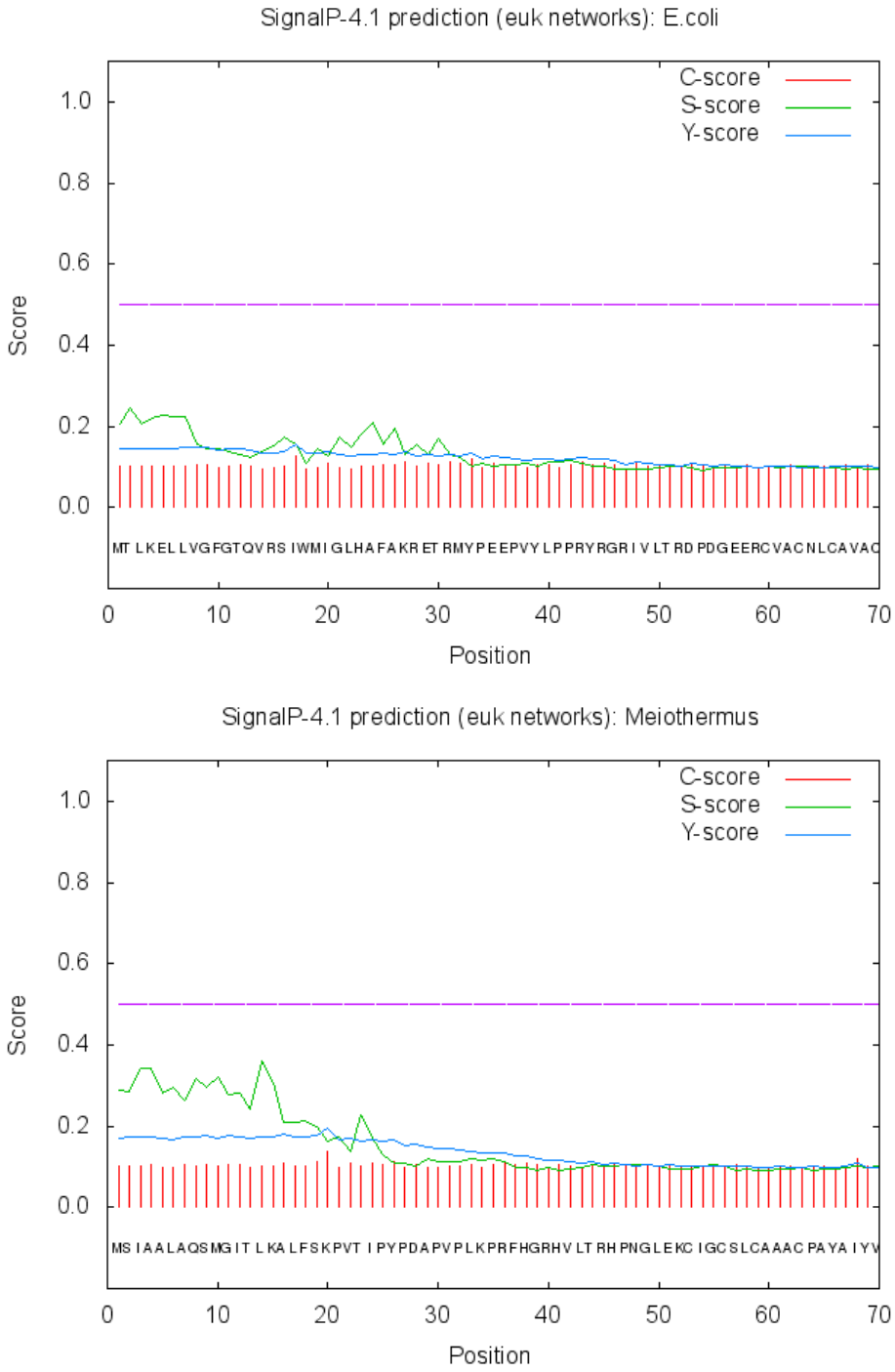


Figure 5: *E. coli NuoI* and Mrub_1869 do not contain cleavage sites. Top graph: SignalP plot for *E. coli NuoI*. Bottom graph: SignalP plot for Mrub_1869. D values for both plots were below the cutoff values. Singal P server v. 4.1 was used to generate the plots (Peterson et al., 2011).

Figure 6 shows the NADH: ubiquinone oxidoreductase complex and the subunits involved that are highlighted in green for both species from the KEGG database (Kanehisa et al., 2016).. *E. coli* and *M. ruber* yielded identical pathway maps for NADH dehydrogenase and both indicated the presence of *NuoI*, *NuoJ*, and *NuoK* genes to produce subunits for the complex. This provides further evidence that these genes are functionally related and orthologous.

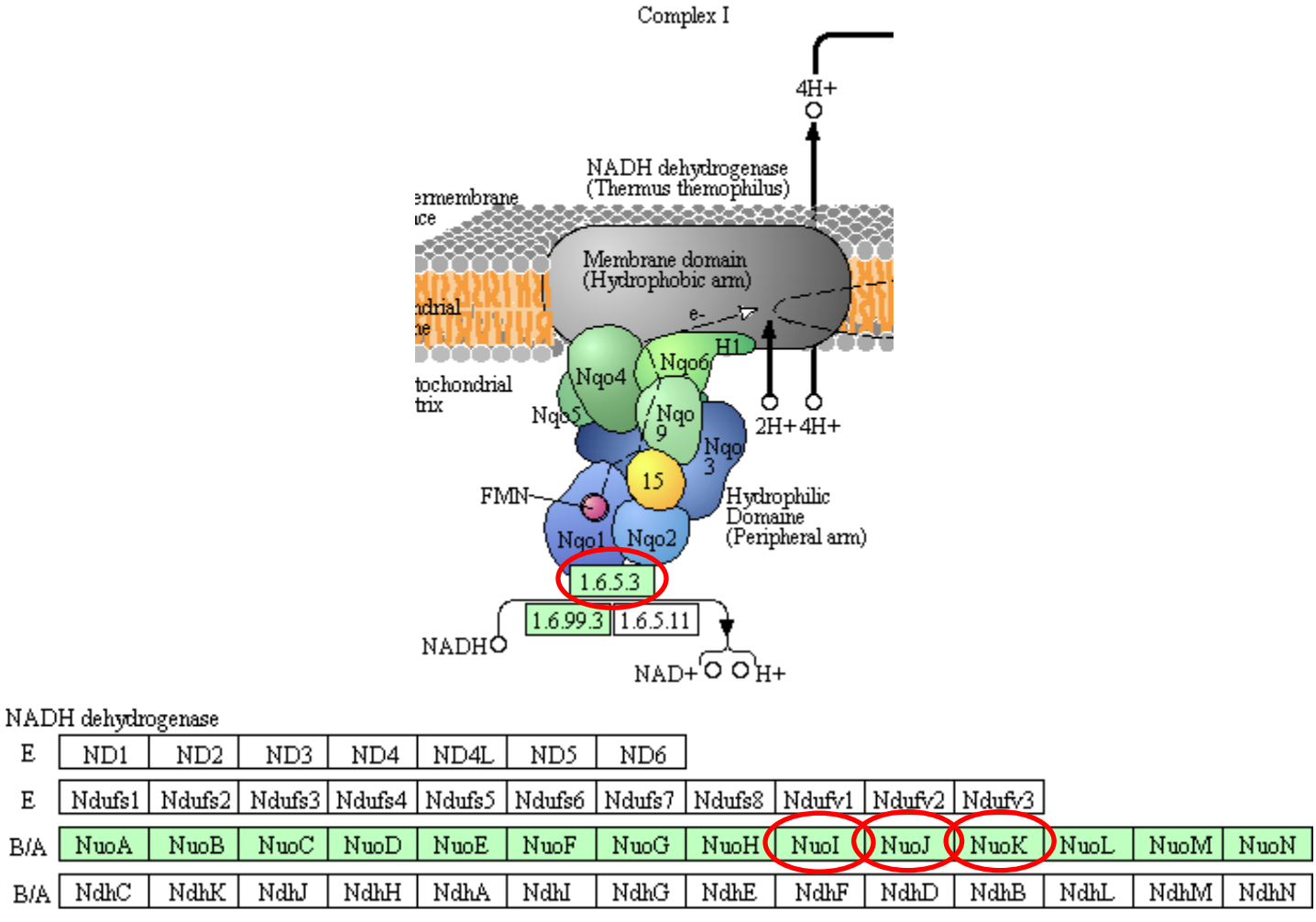


Figure 6: *E. coli* and *M. ruber* have identical KEGG pathway maps regarding NADH dehydrogenase and both have *NuoI*, *NuoJ*, and *NuoK* genes present. Top diagram: NADH: ubiquinone oxidoreductase complex indicating the E.C. number for both species. Bottom diagram: highlighted subunits that are present in the NADH: ubiquinone oxidoreductase complex for both species. Images for both organisms were not shown since both maps are identical. Images taken from the KEGG Pathway Database (Kanehisa et al., 2016).

Evidence for highly conserved amino acids in *E. coli NuoI* and Mrub_1869 proteins was found from the pairwise alignments obtained from the Pfam website (Finn et al., 2016). Figure 7 shows the pairwise alignments for both species and indicates conservation of the same 8 cysteine residues between *E. coli NuoI* and Mrub_1869. Since Pfam compares query sequences to a database of consensus sequences obtained from hundreds of proteins, this is very strong evidence that these genes are orthologous to one another (Finn et al., 2016). This is also strong evidence for these proteins sharing similar function since these highly conserved amino acids between distantly related species indicate their importance in function.

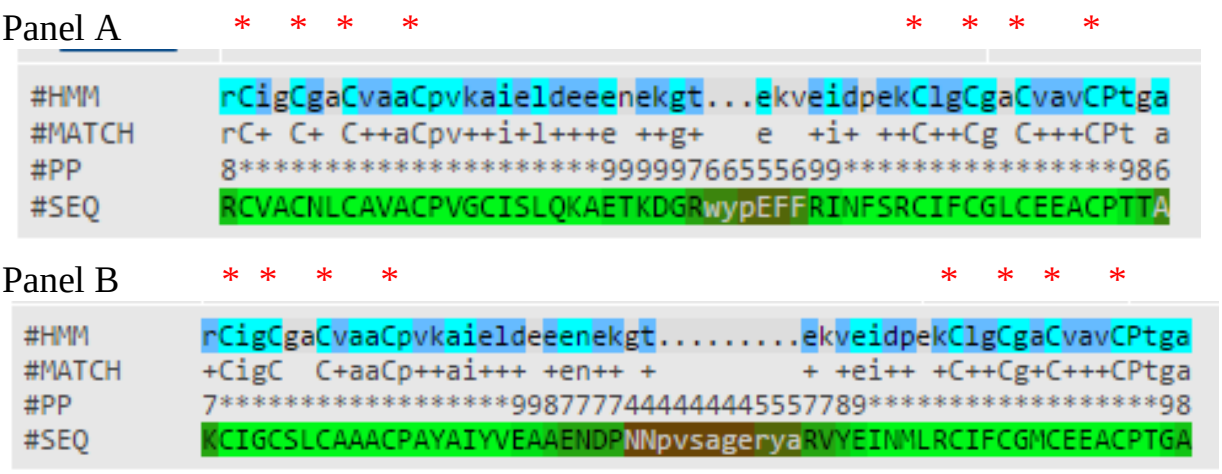
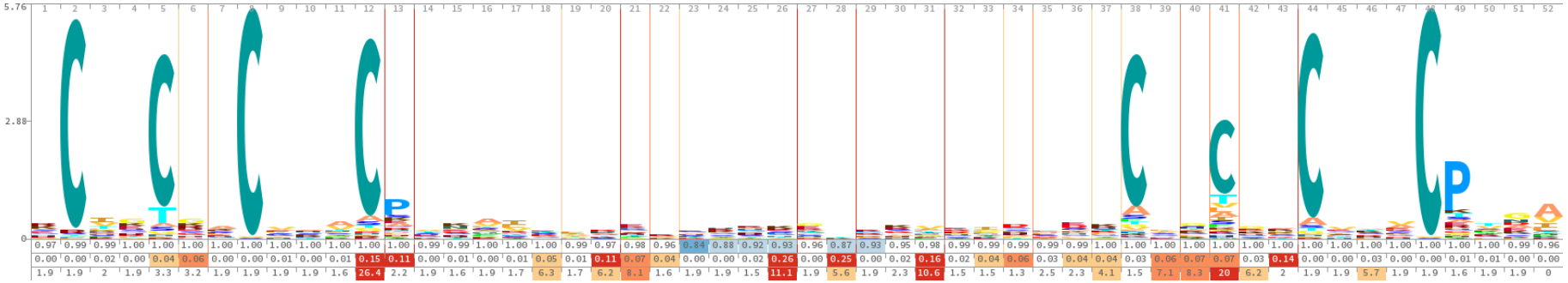


Figure 7: *E. coli NuoI* and Mrub_1869 have the same highly conserved amino acid residues for the same domain (4Fe-4S discluster domain). Panel A depicts the alignment for *E. coli NuoI* . Panel B depicts the alignment for Mrub_1869. Both panels indicate the same 8 highly conserved cysteine residues in the "#MATCH" rows (indicated by asterisks above each panel). Pairwise alignments were generated using the Pfam website (Finn et al., 2016).

Figure 8 shows the prevalence of the cysteine residues among the consensus sequence that was compared to the protein queries for *E. coli NuoI* and Mrub_1869. The large "C" figures indicate the cysteines depicted in figure 7 with asterisks.

Panel A



Panel B

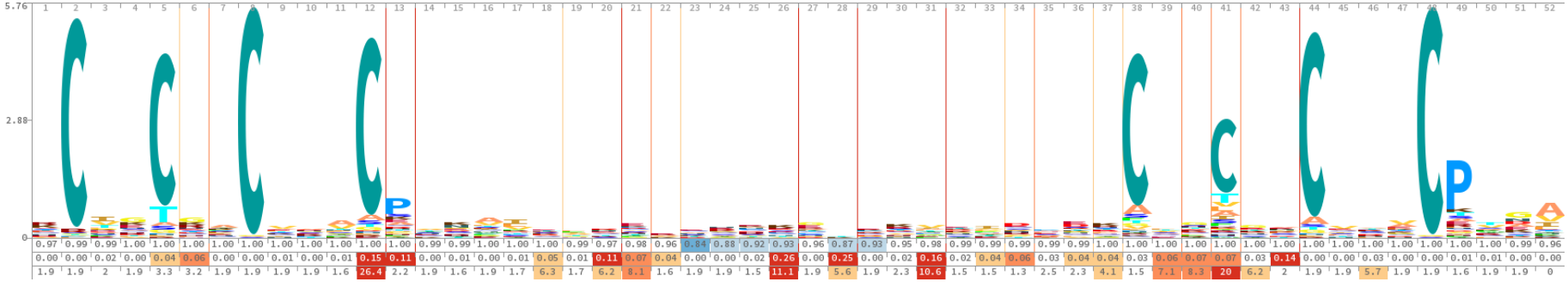
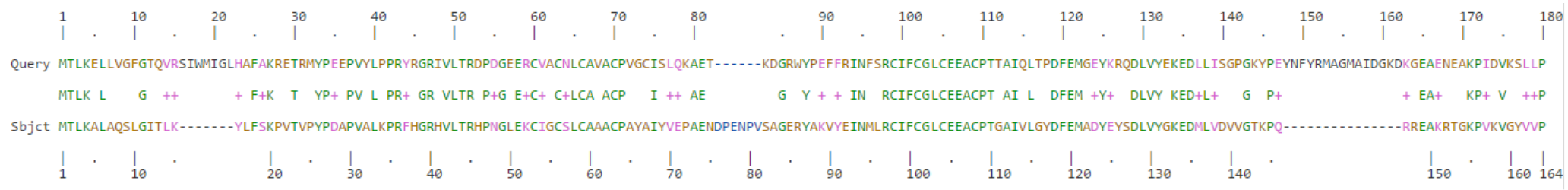


Figure 8: *E. coli NuoI* and Mrub_1869 have the same highly conserved amino acid residues for the same domain (4Fe-4S discluster domain). Panel A depicts the HMM logo for *E. coli NuoI* . Panel B depicts the HMM logo for Mrub_1869. Both panels indicate the same 8 highly conserved cysteine residues. Logos were generated using the Pfam website (Finn et al., 2016).

Figure 9 shows the pairwise alignment between the crystal structure of the hydrophilic domain of respiratory complex I from *Thermus thermophilus* and the *E. coli NuoI* and Mrub_1869 gene protein products. The alignments show that Mrub_1869 is more similar in sequence to the matched protein than *E. coli NuoI* is to the same protein. This is simply due to the close relatedness of *M. ruber* and *T. thermophilus* as species. This figure shows us that both *E. coli NuoI* and Mrub_1869 are similar in sequence to this respiratory complex chain which is evidence that these two genes may be orthologous.

Panel A



Panel B

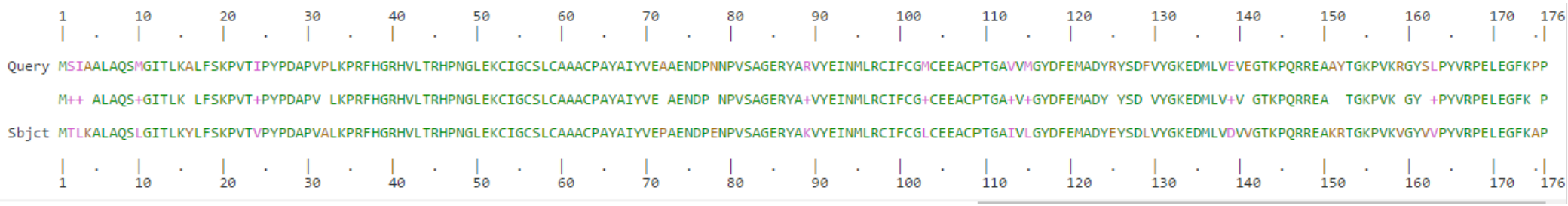


Figure 9: *E. coli NuoI* and Mrub_1869 are similar in sequence to the hydrophilic domain of respiratory complex I from *Thermus thermophilus.* Panel A: *E. coli NuoI* query sequence. Panel B: Mrub_1869 query sequence. This pairwise alignment was obtained from PDB (Berman et al., 2000).

Two cladograms were constructed in search of evidence of horizontal gene transfer within *E. coli* and *M. ruber* using Phylogeny.fr (Dereeper et al., 2008). Figure 10 shows the cladograms for both species that were constructed from the b2281 and Mrub_1869 genes. Since all of the species in the b2281 cladogram were from the same phylum (Proteobacteria) and all of the species from the Mrub_1869 cladogram were from the same phylum (Deinococcus-Thermus), there is no evidence of horizontal gene transfer within these cladograms.

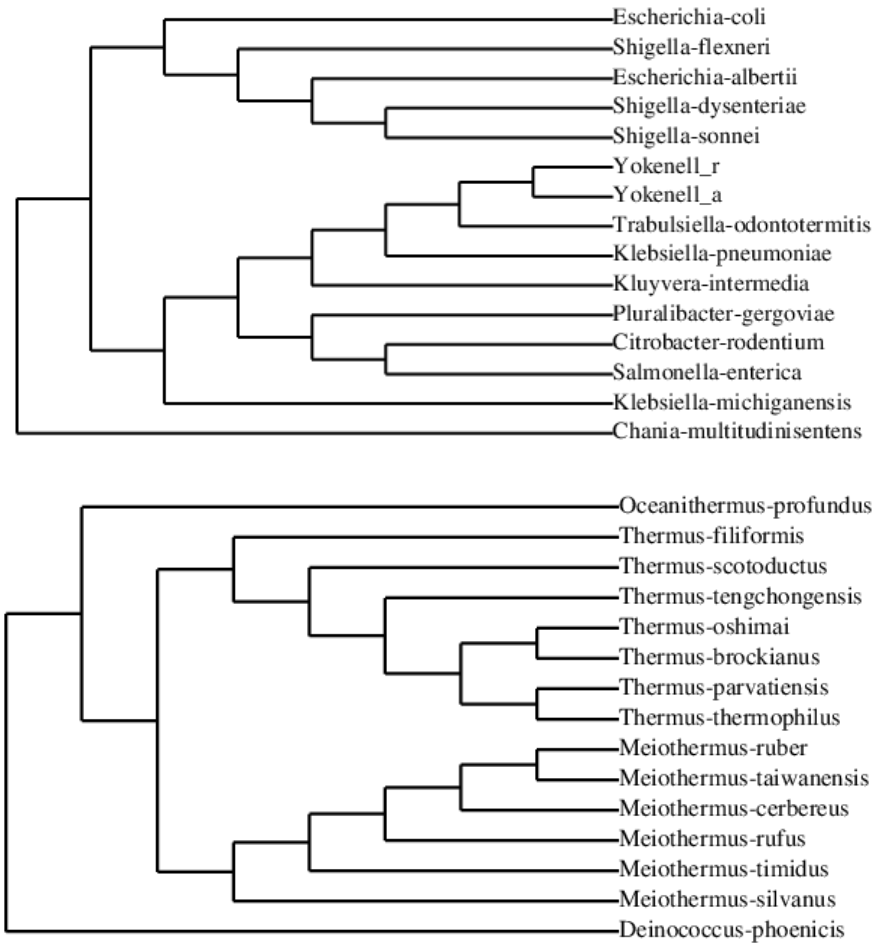
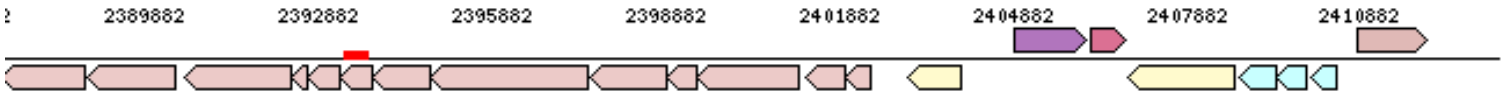


Figure 10: *E. coli NuoI* and Mrub_1869 display no evidence of horizontal gene transfer. Top cladogram: *E. coli NuoI* gene compared to other closely matched species. Bottom cladogram: Mrub_1869 gene compared to other closely matched species. Images generated using Phylogeny.fr (Dereeper et al., 2008).

Figure 11 shows the gene neighborhoods of *E. coli NuoI* and Mrub_1869 colored by KEGG to indicate predicted functions of the genes. Both genes appear to be part of an operon since they are all transcribed in the same direction and they are colored to indicate the same function within the species. The color of the *E. coli NuoI* gene along with its neighboring genes indicates a function of carbohydrate metabolism while the color of the Mrub_1869 gene in its operon indicates “Global and overview maps” which likely indicates a function hasn’t been assigned yet

to this operon. Regardless, the shared coloring of the neighboring genes indicates that *E. coli NuoI* and Mrub_1869 are located in operons. The two genes directly to the left of *NuoI* in panel A represent b2279 (two away) and b2280 (one away) for the *E. coli NuoK* and *NuoJ* genes respectively. The same also applies to panel B for Mrub_1867 (two away) and Mrub_1868 (one away) genes. Therefore, this figure provides evidence that all of the genes being investigated are part of an operon.

Panel A



Panel B

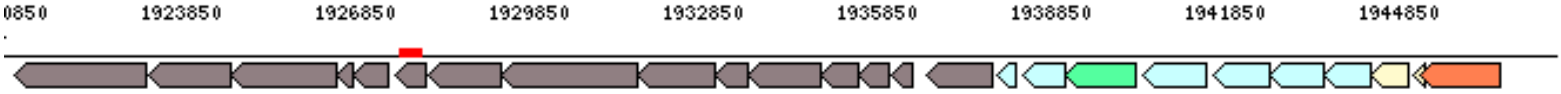


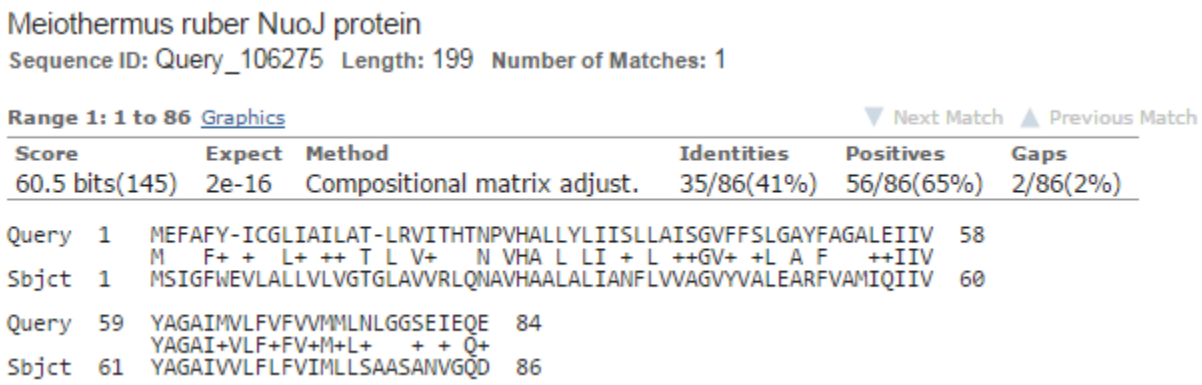
Figure 11: *E. coli NuoI* and Mrub_1869 are located within an operon. Panel A: *E. coli NuoI* gene neighborhood. Panel B: Mrub_1869 gene neighborhood. Respective genes are labeled with a red line above their location. Images generated using the respective genes chromosome viewer colored by KEGG on IMG (Markowitz et al. 2012).

Table 2 provides a summary of the various bioinformatics programs used to compare *E. coli* b2280 and Mrub_1868 genes. The first row indicates the BLAST search conducted to determine the similarities between *E. coli* b2280 and Mrub_1868 gene protein products. The most important piece of information is the small E-value (2e-16) which indicates the probability that these sequences of amino acids in each protein aligned strictly due to chance. While this number is relatively large for the field of bioinformatics, it is still small enough to hypothesize that these two proteins share highly conserved sequences and they have similar functions. Data from CDD yielded the same COG number (COG0839) and name which means that that these proteins likely belong to the same family and have similar functions. The significance of this is noted in the extremely small E-values for both proteins indicating that these did not match to the COG number and family by chance. Thus, these genes both code for the same type of subunit of the NADH: ubiquinone oxidoreductase enzyme. All of the tools used to determine the location of these proteins within the cell (TMHMM, SignalP, LipoP, and PSORTB) indicated that both of the proteins are located in the cytoplasmic membrane. This is consistent with the understanding that subunit J is located within the membrane domain enzyme complex (Keseler et al., 2013). The similar location of both of these proteins is another piece of evidence indicating these genes may be orthologs. TIGRFAM searches yielded no results for both proteins which may indicate that the protein family for this particular subunit is not within the TIGRFAM database and we need further evidence to predict the name and function of the gene products. Pfam searches yielded the same number (PF00499) and name (NADH-ubiquinone/plastoquinone oxidoreductase chain 6) for both protein queries along with low E-values indicating the most highly conserved domain between these proteins. This is also consistent with expectations since evidence from KEGG and Ecocyc predicted these proteins to be a part of the NADH: ubiquinone oxidoeductase complex. The protein database pulled the different PDB codes but indicated the same type of crystal structure for both protein queries, just from different species. This simply tells us that *E. coli* b2280 shares more sequence resemblance to its own structure of the respiratory complex I while Mrub_1868 is more similar in sequence to the respiratory complex from *Thermus thermophilus.* This makes sense since *M. ruber* is more closely related to *T. thermophilus* than *E. coli* and the amino acid sequence of the protein should match more closely. Both of these PDB matches yielded low E-values. These searches also yielded the same enzyme commission number. Since the protein database is a worldwide depository of information regarding 3-D structures of proteins and their sequences, these significant matches for both queries are strong evidence that these proteins have the same function, just in different organisms (Berman et al., 2000). Lastly, both genes were predicted to participate in the production of the same subunit of NADH: ubiquinone oxidoreductase in the oxidative phosphorylation pathway.

Table 2: *E. coli NuoJ* and Mrub_1868 are predicted orthologs

Bioinformatics Tool Used	***E. coli* b2280 protein *(NuoJ)***	**Mrub_1868 protein**
BLAST *E. coli* against *M. ruber*	Score: 60.5 bits E-value: 2e-16	
CDD Data (COG match)	COG Number: COG0839 NADH:ubiquinone oxidoreductase subunit 6 (chain J)	
	E-value: 2.54e-40	E-value: 3.19e-13
Cellular Localization	Cytoplasmic Membrane	
TIGRFAM Protein Family	No results from either queries	
Pfam Match	Pfam Number: PF00499 Pfam Name: NADH-ubiquinone/plastoquinone oxidoreductase chain 6	
	E-value: 3.1e-33	E-value: 1.2e-28
Protein Database	PDB Code: 3RKO PDB Name: Crystal structure of the membrane domain of respiratory complex I from *E. coli*	PDB Code: 4HE8 PDB Name: Crystal structure of the membrane domain of respiratory complex I from *Thermus thermophilus*
	E-value: 1.470e-86	E-value: 2.943e-12
EC Number	E.C.1.6.5.3 - NADH:ubiquinone reductase (H^+-translocating)	
KEGG Pathway Map	Oxidative Phosphorylation Pathway	

Figure 12 indicates the results of the protein BLAST search done using *E. coli* b2280 as the query sequence to match to *M. ruber*. The data indicates that 41% of the amino acids are an exact match between the two proteins and 65% of the amino acids were similar in character of the R-groups. The low E-value for this match also indicates that these sequences had a fairly small probability of aligning strictly due to chance. While the E-value is not as small as we would like it to be, this is the first piece of evidence to suggest that these two genes are related and share similar functions. Further evidence is needed to confidently predict that these two genes are orthologous.


Meiothermus ruber NuoJ protein
Sequence ID: Query_106275 Length: 199 Number of Matches: 1

Range 1: 1 to 86 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
60.5 bits(145)	2e-16	Compositional matrix adjust.	35/86(41%)	56/86(65%)	2/86(2%)

```
Query  1   MEFAFY-ICGLIAILAT-LRVITHTNPVHALLYLIISLLAISGVFFSLGAYFAGALEIIV  58
           M   F+ +  L+ ++ T L V+   N VHA L LI + L ++GV+ +L A F   ++IIV
Sbjct  1   MSIGFWEVLALLVLVGTGLAVVRLQNAVHAALALIANFLVVAGVYYVALEARFVAMIQIIV  60

Query  59  YAGAIMVLFVFVVMMLNLGGSEIEQE  84
           YAGAI+VLF+FV+M+L+   + + Q+
Sbjct  61  YAGAIVVLFLFVIMLLSAASANVGQD  86
```

Figure 12: *E. coli NuoJ* and Mrub_1868 have similar protein sequences. The query sequence is *E. coli* b2280 and the subject sequence is Mrub_1869. The search was conducted using the NCBI BLAST bioinformatics program (Madden et al., 2002).

Figure 13 indicates the TMH hydropathy plots for *E. coli NuoJ* and Mrub_1868 proteins. The red clusters of peaks indicate the probabilities that transmembrane helices are located near those amino acid positions. There are 5 areas in both *E. coli* and *M. ruber* with clusters of peaks that indicate areas where transmembrane helices are very likely located. These hydropathy plots are consistent with one another with respect to location of these helices at amino acid positions indicating that these proteins are likely associated with the cellular membrane and very likely orthologous.

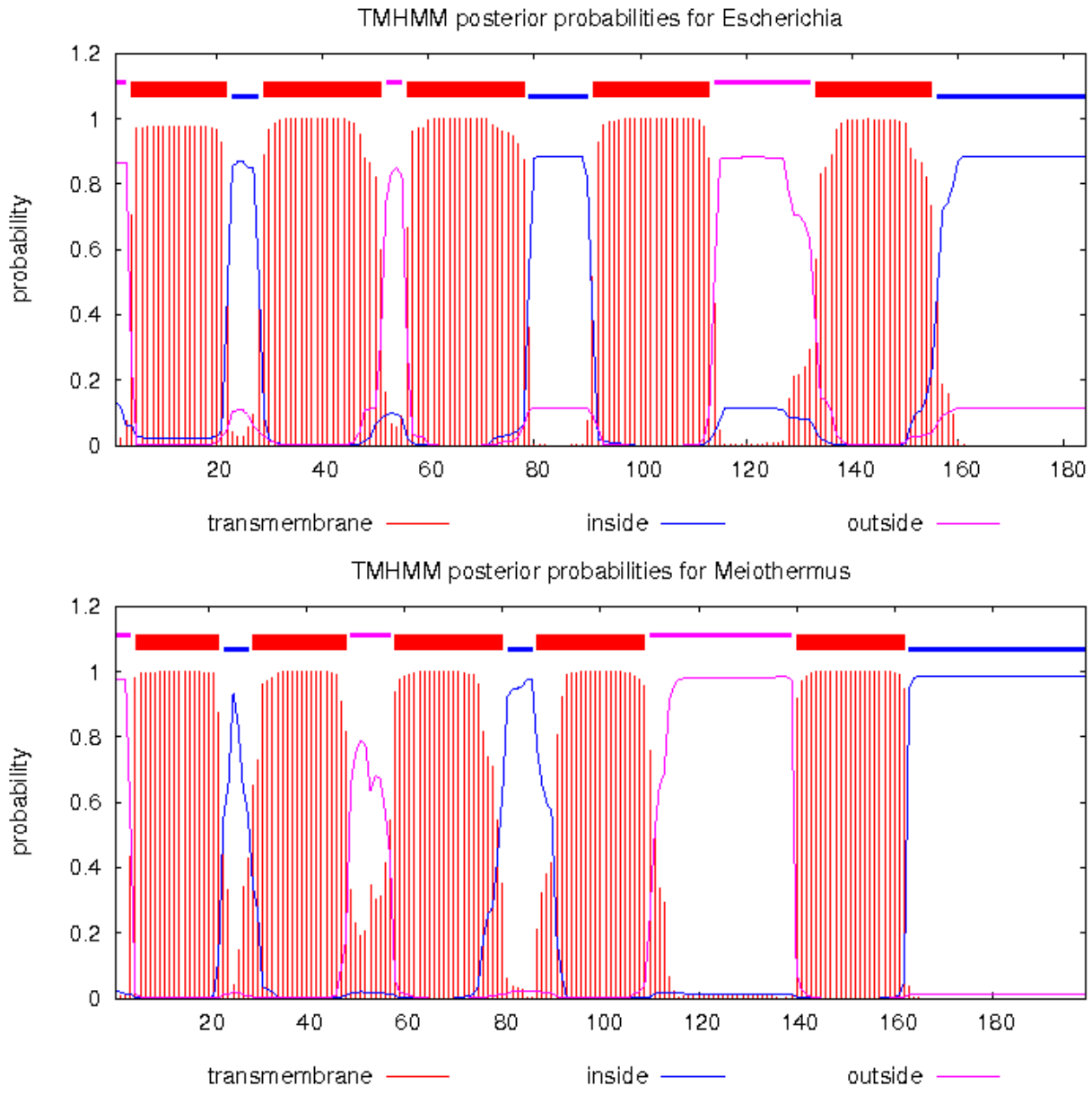


Figure 13: *E. coli NuoJ* and Mrub_1868 contain 5 TMH regions indicating that both proteins are likely associated with the cellular/cytoplasmic membrane. Top graph: probability plot for *E. coli* b2280 protein. Bottom graph: probability plot for Mrub_1868 protein. Graphs were generated using TMHMM server (Krogh 2016).

Investigation of possible cleavage sites in both *E. coli NuoJ* and Mrub_1868 proteins using SignalP yielded no predictions for cleavage sites (Figure 14). *E. coli NuoJ* had a D-value of 0.259 which at first glance would suggest that there are no cleavage sites present in this protein. However, Mrub_1868 yielded a D-value of 0.484 and figure 14 indicates that the C and Y-scores are very close to the cutoff value of 0.500. While the result did not yield a significant value, it may indicate the possibility that this protein does have a signal peptide sequence but further evidence would be needed to support this hypothesis. Findings from LipoP indicated both protein

locations to be “Transmembrane helix” and predicted no cleavage sites (Juncker et al., 2003).PSORT-B findings predicted both proteins to be located in the cytoplasmic membrane with scores of 10.00 for that category (Yu et al., 2010).

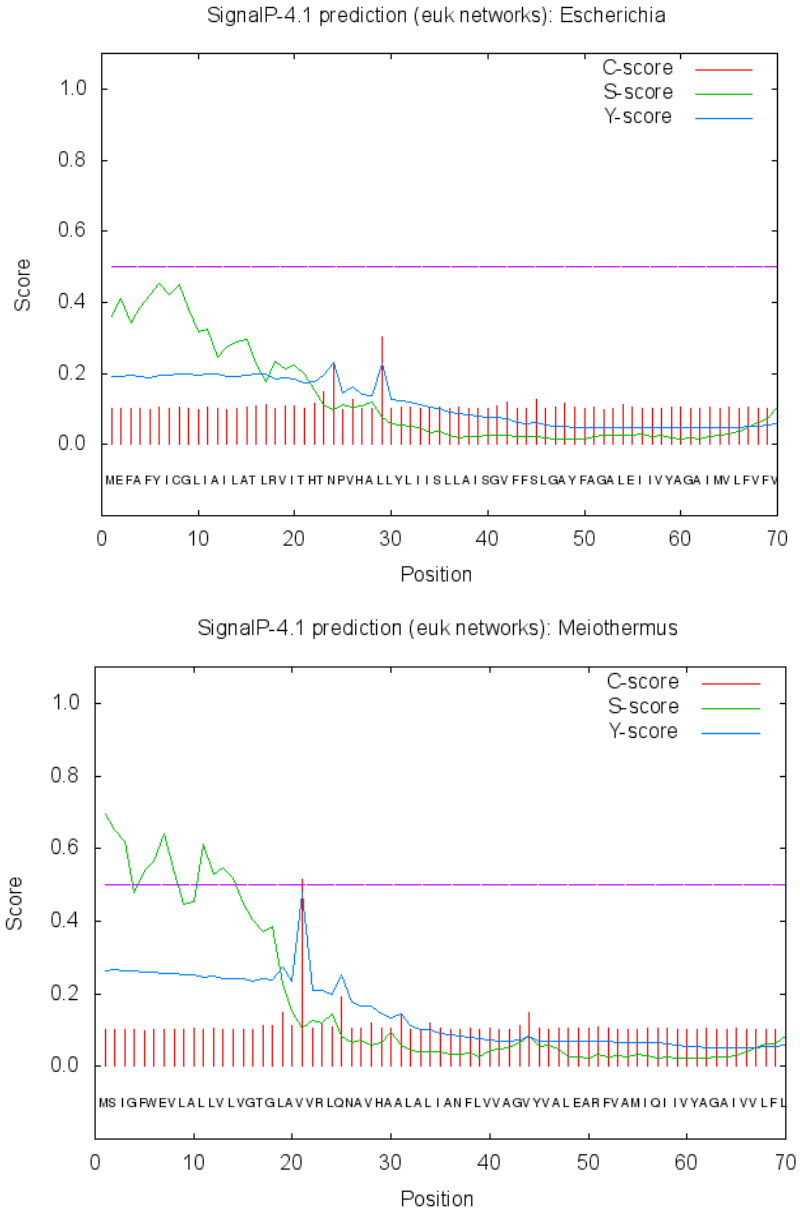
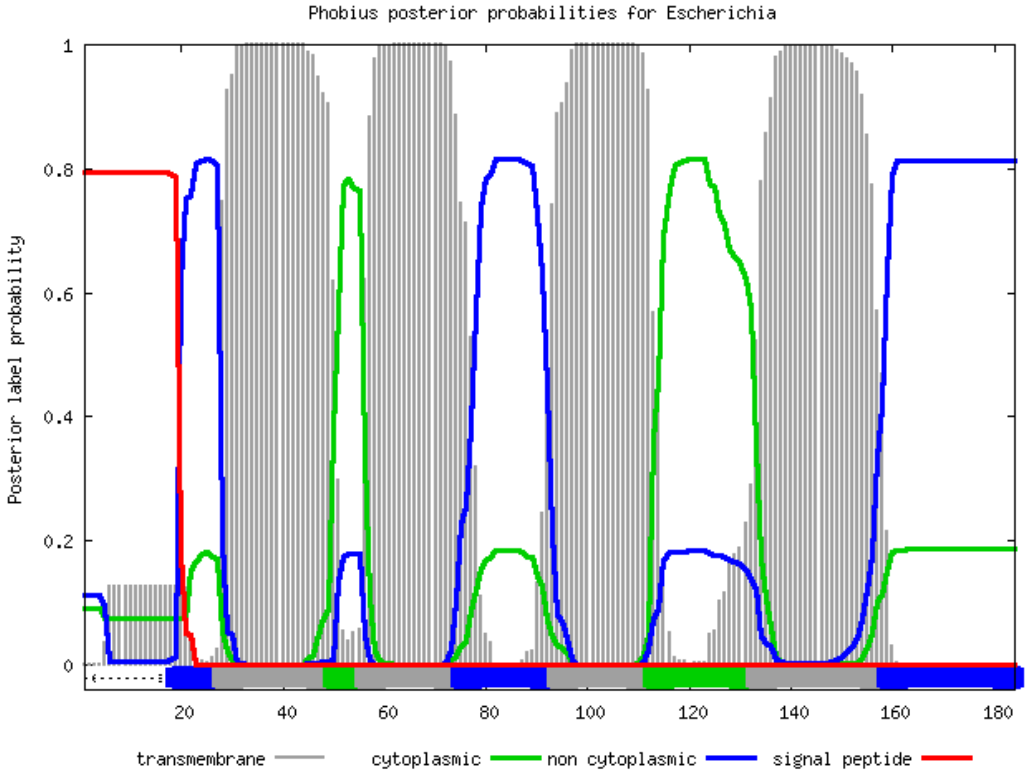


Figure 14: *E. coli NuoJ* and Mrub_1868 are not predicted to contain cleavage sites by SignalP. Top graph: SignalP plot for *E. coli NuoJ*. Bottom graph: SignalP plot for Mrub_1868. D values for both plots were below the cutoff values. Singal P server v. 4.1 was used to generate the plots (Peterson et al., 2011).

Figure 15 provides evidence that *E. coli NuoJ* and Mrub_1868 may have signal peptide sequences and cleavage sites, rather than a TMH, located near the N-terminus of the proteins. The Phobius probability graph for *E. coli* indicates a probability near 0.8 for a signal peptide

sequence between amino acids 1-20 and a low probability (~0.10) of a TMH in the same positions. The Phobius probability graph for Mrub_1868 indicates similar findings with a signal peptide probability near 0.6 within the first 20 amino acids and a low probability (~0.4) of TMH in the same positions. Phobius combines the methods used in TMHMM and SignalP to provide more accurate info regarding the presence of helices and signal peptides (Kall et al., 2004). This supports our earlier hypothesis that these proteins may actually have signal sequences and cleavage sites. Rather than 5 transmembrane helices for these proteins, there are likely 4 and the N-terminus of each protein contains a signal peptide sequence.



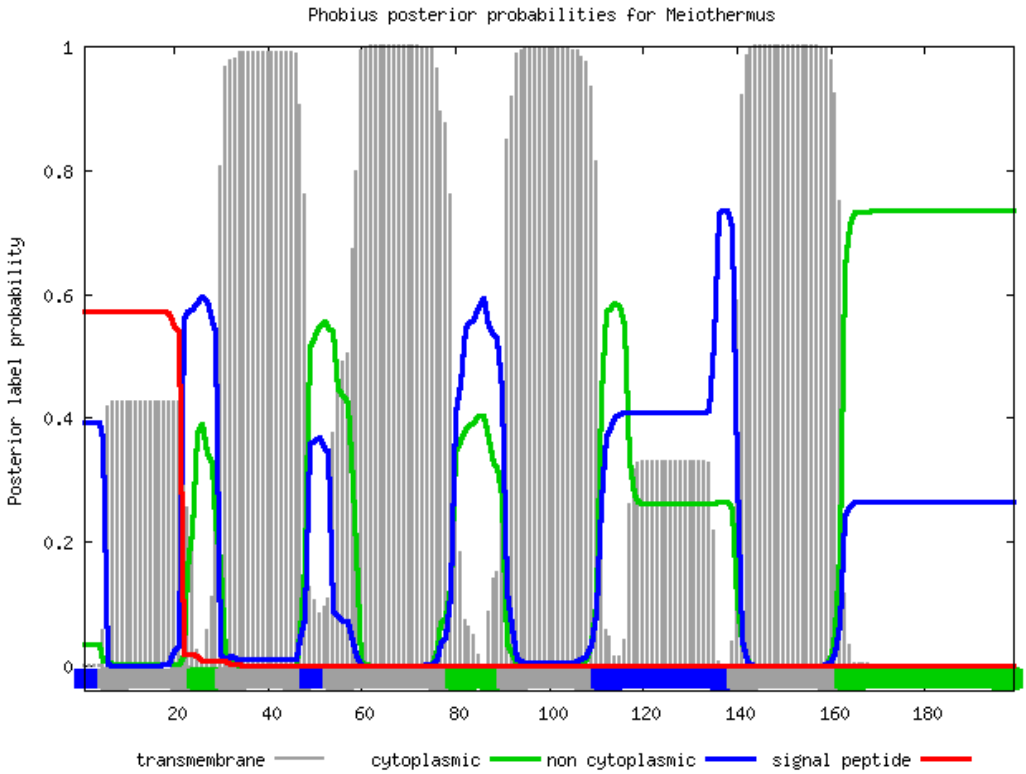
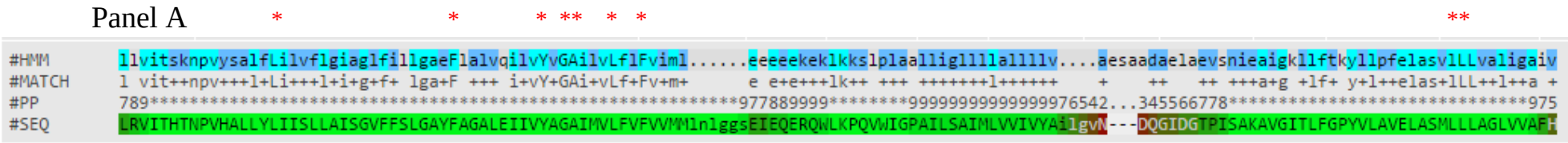


Figure 15: *E. coli NuoJ* and Mrub_1868 contain cleavage sites. Top graph: Phobius probability plot for *E. coli NuoJ*. Bottom graph: Phobius probability plot for Mrub_1868. Signal peptide probabilities for the first 20 amino acids in both proteins indicate the presence of a signal peptide sequence rather than a TMH. Plots were generated using the Phobius bioinformatic tool (Kall et al., 2004).

Evidence for highly conserved amino acids in *E. coli NuoJ* and Mrub_1868 proteins was found from the pairwise alignments obtained from the Pfam website (Finn et al., 2016). Figure 16 shows the pairwise alignments for both species and indicates conservation of the same 9 amino acid residues between *E. coli NuoJ* and Mrub_1868. Since Pfam compares query sequences to a database of consensus sequences obtained from hundreds of proteins, this is very strong evidence that these genes are orthologous to one another (Finn et al., 2016). This is also strong evidence for these proteins sharing similar function since these highly conserved amino acids between distantly related species indicate their importance in function. The HMM logos obtained from Pfam were too long to fit onto a page and include as a figure but both logos are consistent with the Pfam alignment conserved residues in figure 16.

Panel A



Panel B

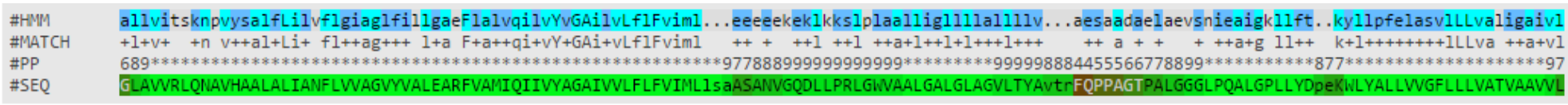
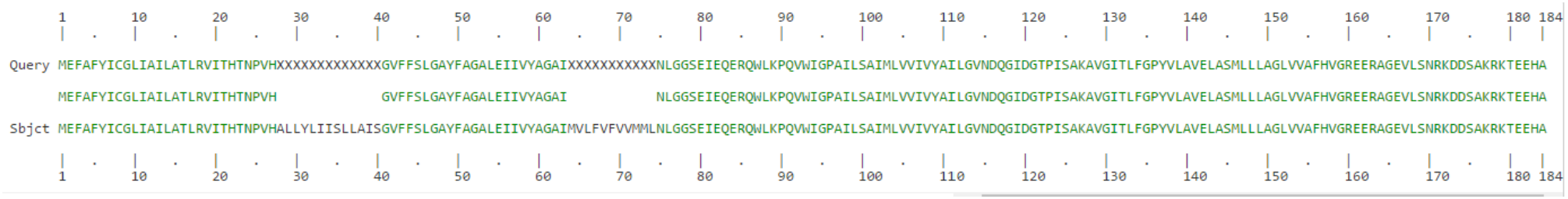


Figure 16: *E. coli NuoJ* and Mrub_1868 have the same highly conserved amino acid residues for the same domain (NADH-ubiquinone/plastoquinone oxidoreductase chain 6). Panel A depicts the alignment for *E. coli NuoJ* . Panel B depicts the alignment for Mrub_1868. Both panels indicate the same 9 highly conserved amino acid residues in the "#MATCH" rows (indicated by asterisks above each panel). Pairwise alignments were generated using the Pfam website (Finn et al., 2016).

Figure 17 shows the PDB alignment between *E. coli NuoJ* and the membrane domain of respiratory complex I from *E. coli* which confirms our previous knowledge that this protein is associated with the cellular membrane (Berman et al., 2000). Figure 17 also shows Mrub_1868 and its PDB alignment with the membrane domain of respiratory complex I from *Thermus thermophilus* (Berman et al., 2000). This sequence similarity provides further evidence that this protein is associated with the cellular membrane since its domain is aligning with the known structure of the membrane domain of the NADH: ubiquinone oxidoreductase complex in a closely related species. The difference in alignment length of the two genes is due to the fact that they matched with the proteins from different species (albeit indicating the same type of functional protein).

Panel A



Panel B

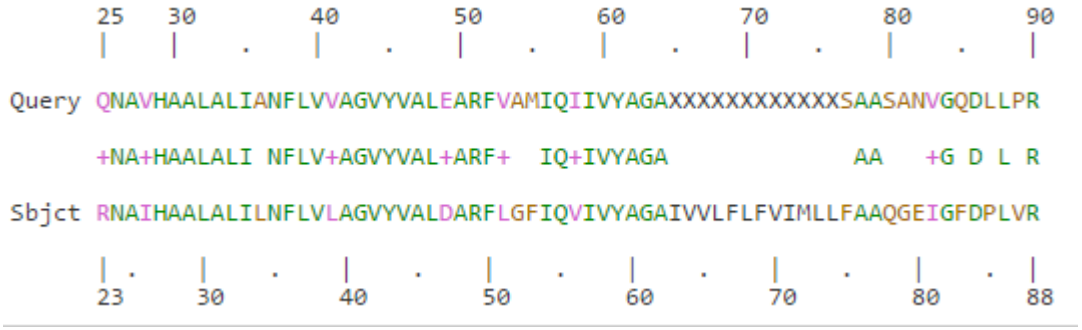


Figure 17: *E. coli NuoJ* and Mrub_1868 are parts of the membrane domain of NADH: ubiquinone oxidoreductase (complex I). Panel A: PDB alignment of *E. coli NuoJ* with the membrane domain of respiratory complex I from *E. coli.* Panel B: Mrub_1868 alignment with the membrane domain of respiratory complex I from *Thermus thermophilus*. Alignments were obtained from PDB website upon using query sequences (Berman et al., 2000).

Two cladograms were constructed in search of evidence of horizontal gene transfer within *E. coli NuoJ* and Mrub_1868 using Phylogeny.fr (Dereeper et al., 2008). Figure 18 shows the cladograms for both species that were constructed from the b2280 and Mrub_1868 genes. Since all of the species in the b2280 cladogram were from the same phylum (Proteobacteria) and all of the species from the Mrub_1868 cladogram were from the same phylum (Deinococcus-Thermus), there is no evidence of horizontal gene transfer within these cladograms.

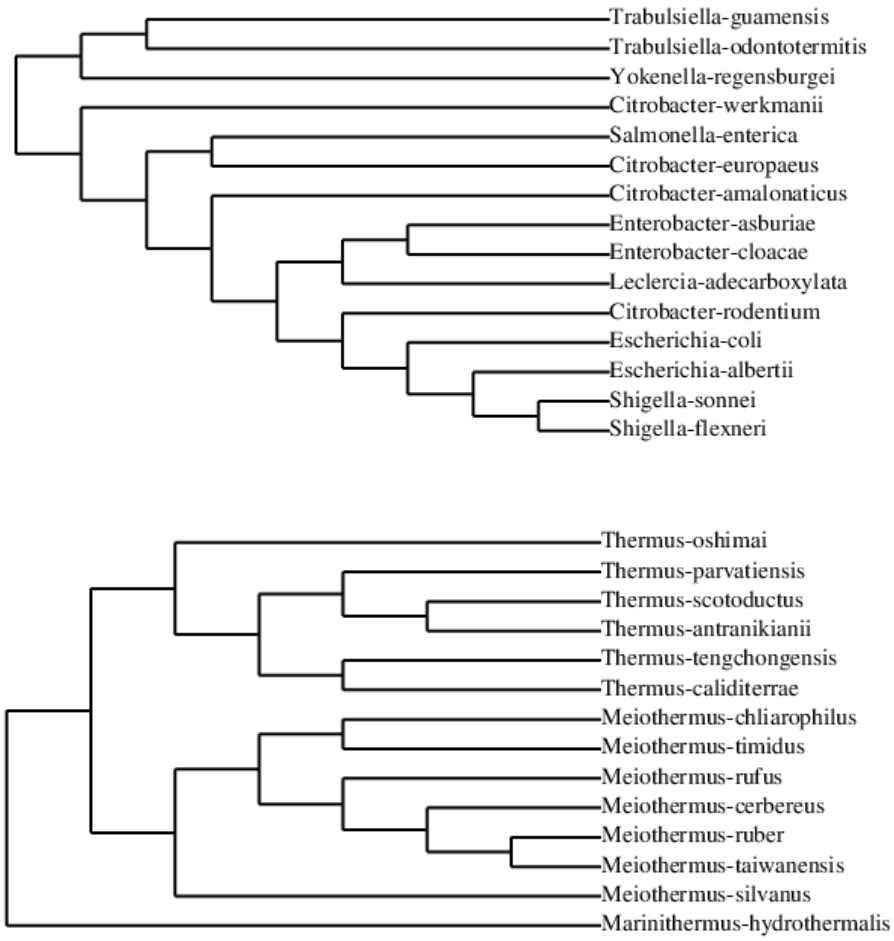


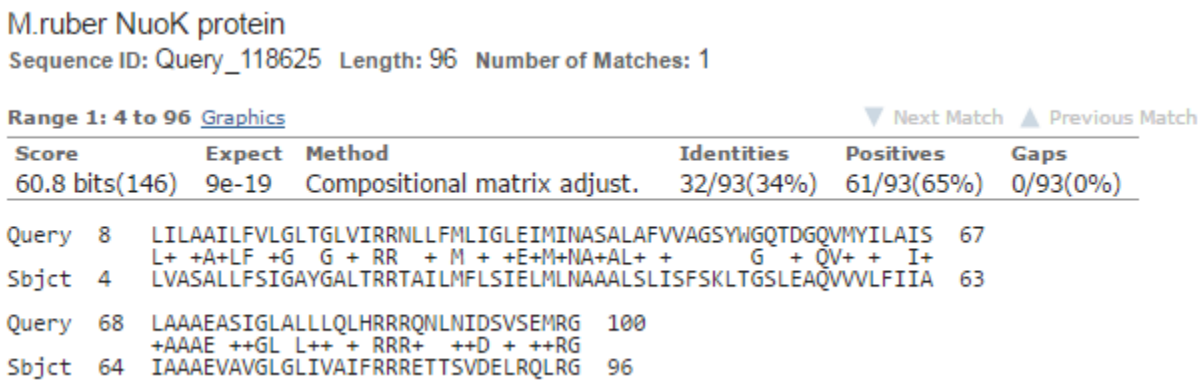
Figure 18: *E. coli NuoJ* and Mrub_1868 display no evidence of horizontal gene transfer. Top cladogram: *E. coli NuoJ* gene compared to other closely matched species. Bottom cladogram: Mrub_1868 gene compared to other closely matched species. Images generated using Phylogeny.fr (Dereeper et al., 2008).

Table 3 provides a summary of the various bioinformatics programs used to compare *E. coli* b2279 and Mrub_1867 genes. The first row indicates the BLAST search conducted to determine the similarities between *E. coli* b2279 and Mrub_1867 gene protein products. The most important piece of information is the small E-value (9e-19) which indicates the probability that these sequences of amino acids in each protein aligned strictly due to chance. While this number is relatively large for the field of bioinformatics, it is still small enough to hypothesize that these two proteins share highly conserved sequences and they have similar functions. Data from CDD yielded the same COG number (COG0713) and name which means that that these proteins likely belong to the same family and have similar functions. The significance of this is noted in the extremely small E-values for both proteins indicating that these did not match to the COG number and family by chance. Thus, these genes both code for the same type of subunit of the NADH: ubiquinone oxidoreductase enzyme. All of the tools used to determine the location of these proteins within the cell (TMHMM, SignalP, LipoP, and PSORTB) indicated that both of the proteins are located in the cytoplasmic membrane. This is consistent with the understanding that subunit K is located within the membrane domain enzyme complex (Keseler et al., 2013). The similar location of both of these proteins is another piece of evidence indicating these genes may be orthologs. TIGRFAM searches yielded no results for both proteins which may indicate that the protein family for this particular subunit is not within the TIGRFAM database and we need further evidence to predict the name and function of the gene products. Pfam searches yielded the same number (PF00420) and name (NADH-ubiquinone/plastoquinone oxidoreductase chain 4L) for both protein queries along with low E-values indicating the most highly conserved domain between these proteins. This is also consistent with expectations since evidence from KEGG and Ecocyc predicted these proteins to be a part of the NADH: ubiquinone oxidoeductase complex. The protein database yielded different PDB codes but indicated the same type of crystal structure for both protein queries, just from different species. This simply tells us that *E. coli* b2279 shares more sequence resemblance to its own structure of the respiratory complex I while Mrub_1867 is more similar in sequence to the respiratory complex from *Thermus thermophilus.* This makes sense since *M. ruber* is more closely related to *T. thermophilus* than *E. coli* and the amino acid sequence of the protein should match more closely. Both of these PDB matches yielded low E-values. These searches also yielded the same enzyme commission number. Since the protein database is a worldwide depository of information regarding 3-D structures of proteins and their sequences, these significant matches for both queries are strong evidence that these proteins have the same function, just in different organisms (Berman et al., 2000). Lastly, both genes were predicted to participate in the production of the same subunit of NADH: ubiquinone oxidoreductase in the oxidative phosphorylation pathway.

Table 3: *E. coli NuoK* and Mrub_1867 are predicted orthologs

Bioinformatics Tool Used	***E. coli* b2279 protein *(NuoJ)***	**Mrub_1867 protein**
BLAST *E. coli* against *M. ruber*	Score: 60.8 bits E-value: 9e-19	
CDD Data (COG match)	COG Number: COG0713 NADH:ubiquinone oxidoreductase subunit 11 or 4L (chain K)	
	E-value: 1.11e-35	E-value: 21e-113
Cellular Localization	Cytoplasmic Membrane	
TIGRFAM Protein Family	No results from either queries	
Pfam Match	Pfam Number: PF00420 Pfam Name: NADH-ubiquinone/plastoquinone oxidoreductase chain 4L	
	E-value: 7.9e-21	E-value: 5.3e-23
Protein Database	PDB Code: 3RKO PDB Name: Crystal structure of the membrane domain of respiratory complex I from *E. coli*	PDB Code: 4HE8 PDB Name: Crystal structure of the membrane domain of respiratory complex I from *Thermus thermophilus*
	E-value: 6.721e-31	E-value: 1.255e-17
EC Number	E.C.1.6.5.3 - NADH:ubiquinone reductase (H^+-translocating)	
KEGG Pathway Map	Oxidative Phosphorylation Pathway	

Figure 19 indicates the results of the protein BLAST search done using *E. coli* b2279 as the query sequence to match to *M. ruber*. The data indicates that 34% of the amino acids are an exact match between the two proteins and 65% of the amino acids were similar in character of the R-groups. The low E-value for this match also indicates that these sequences had a fairly small probability of aligning strictly due to chance. While the E-value is not as small as we would like it to be, this is the first piece of evidence to suggest that these two genes are related and share similar functions. Further evidence is needed to confidently predict that these two genes are orthologous.


M.ruber NuoK protein

Sequence ID: Query_118625 Length: 96 Number of Matches: 1

Range 1: 4 to 96 Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps
60.8 bits(146)	9e-19	Compositional matrix adjust.	32/93(34%)	61/93(65%)	0/93(0%)

```
Query  8   LILAAILFVLGLTGLVIRRNLLFMLIGLEIMINASALAFVVAGSYWGQTDGQVMYILAIS  67
           L+ +A+LF +G  G + RR  + M + +E+M+NA+AL+ +      G  + QV+ +  I+
Sbjct  4   LVASALLFSIGAYGALTRRTAILMFLSIELMLNAAALSLISFSKLTGSLEAQVVVLFIIA  63

Query  68  LAAAEASIGLALLLQLHRRRQNLNIDSVSEMRG  100
           +AAAE ++GL L++ + RRR+  ++D + ++RG
Sbjct  64  IAAAEVAVGLGLIVAIFRRRETTSVDELRQLRG  96
```

Figure 19: *E. coli NuoK* and Mrub_1867 have similar protein sequences. The query sequence is *E. coli* b2279 and the subject sequence is Mrub_1867. The search was conducted using the NCBI BLAST bioinformatics program (Madden et al., 2002).

Figure 20 indicates the TMH hydropathy plots for *E. coli NuoK* and Mrub_1867 proteins. The red clusters of peaks indicate the probabilities that transmembrane helices are located near those amino acid positions. There are 3 areas in both *E. coli* and *M. ruber* with clusters of peaks that indicate areas where transmembrane helices are very likely located. These hydropathy plots are consistent with one another with respect to location of these helices at amino acid positions indicating that these proteins are likely associated with the cellular membrane and very likely orthologous.

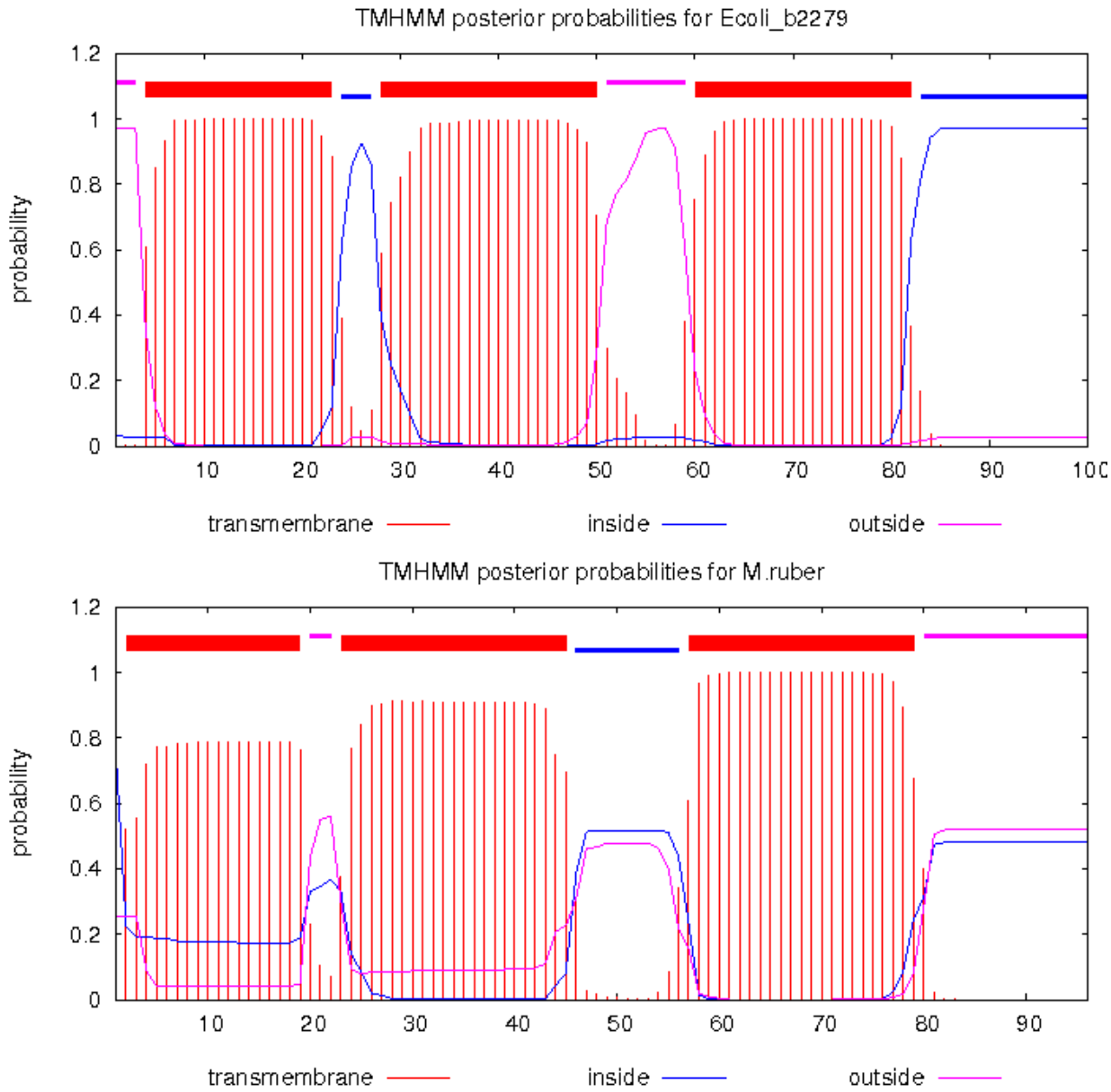


Figure 20: *E. coli NuoK* and Mrub_1867 contain 3 TMH regions indicating that both proteins are likely associated with the cellular/cytoplasmic membrane. Top graph: probability plot for *E. coli* b2279 protein. Bottom graph: probability plot for Mrub_1867 protein. Graphs were generated using TMHMM server (Krogh 2016).

Investigation of possible cleavage sites in both *E. coli NuoK* and Mrub_1867 proteins using SignalP yielded no predictions for cleavage sites (Figure 21). *E. coli NuoK* had a D-value of 0.286 suggesting that there are no cleavage sites present in this protein. However, Mrub_1868 yielded a D-value of 0.321 and figure 21 indicates that the C and Y-scores fluctuate near amino acid 18. While the result did not yield a significant value, it may indicate the possibility that this protein does have a signal peptide sequence but further evidence would be needed to support this hypothesis. Findings from LipoP indicated both protein locations to be “Transmembrane helix” and predicted no cleavage sites (Juncker et al., 2003). PSORT-B findings predicted both proteins

to be located in the cytoplasmic membrane with scores of 10.00 for that category (Yu et al., 2010).

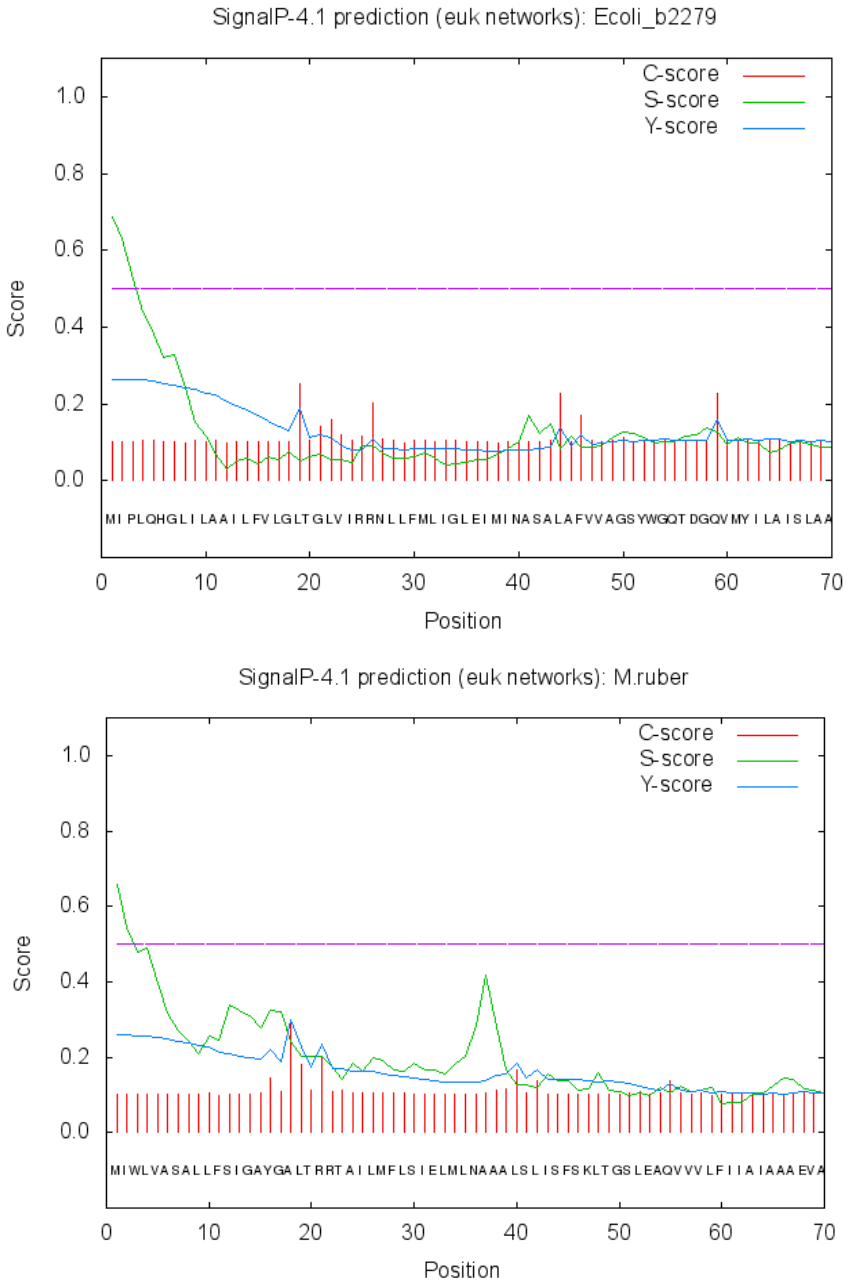
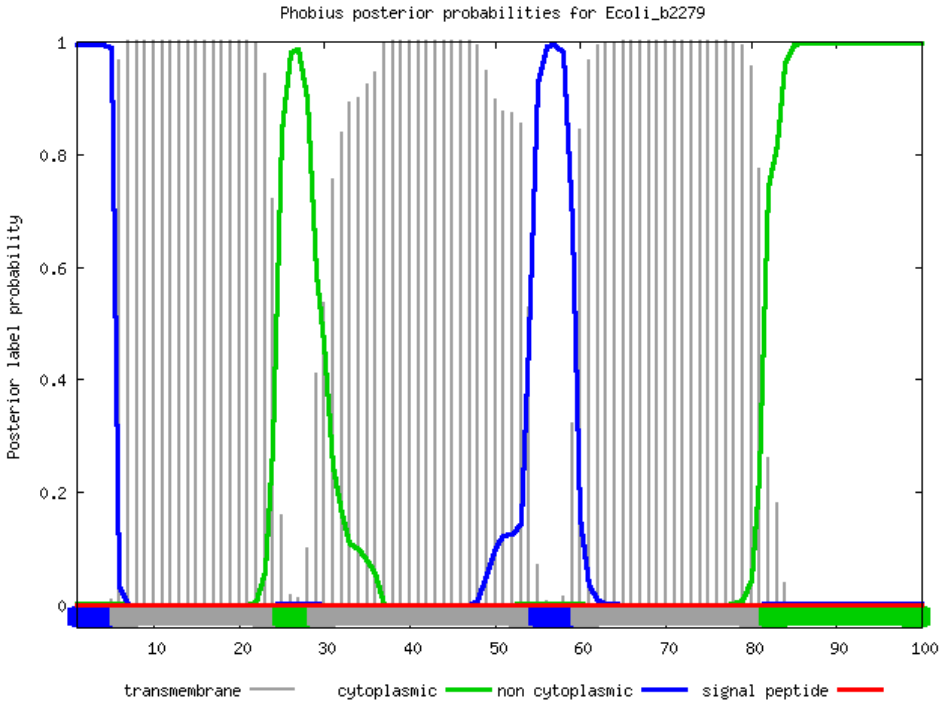


Figure 21: *E. coli NuoK* and Mrub_1867 are not predicted to contain cleavage sites by SignalP. Top graph: SignalP plot for *E. coli NuoK*. Bottom graph: SignalP plot for Mrub_1867. D values for both plots were below the cutoff values. Singal P server v. 4.1 was used to generate the plots (Peterson et al., 2011).

Figure 22 provides evidence that Mrub_1868 may have signal peptide sequences and cleavage sites, rather than a TMH, located near the N-terminus of the protein. The Phobius probability graph for *E. coli* b2279 indicates a probability of 0.0 for a signal peptide sequence for all amino

acids and a high probability (~01.00) of all transmembrane helices. The Phobius probability graph for Mrub_1867 indicates much different findings with a signal peptide probability of 1.00 within the first 20 amino acids and a low probability (~0.00) of TMH in the same positions. This indicates a stark contrast between the two proteins. *E. coli NuoK* likely has 3 transmembrane helices and no signal peptide sequence while Mrub_1867 appears to only have 2 transmembrane helices due to the presence of a signal peptide sequence near the N-terminus that replaces our previously predicted TMH for that protein. While this may appear to be refuting evidence, both proteins are still predicted to be associated with the cytoplasmic membrane and the presence of a signal peptide is not required for this feature.



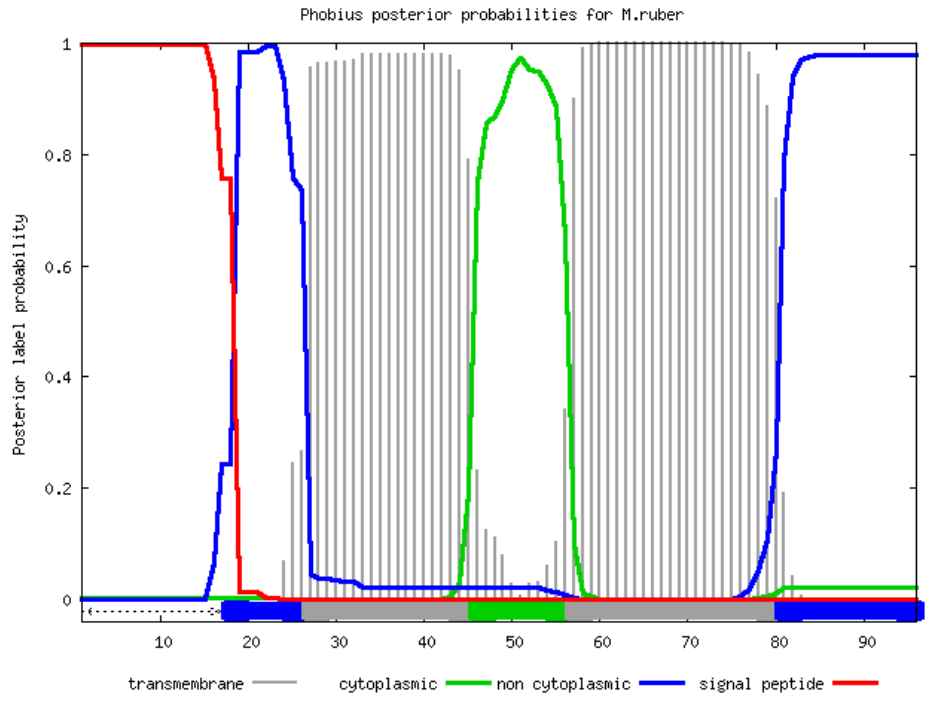


Figure 22: Mrub_1867 contains a signal peptide sequence but *E. coli NuoK* does not. Top graph: Phobius probability plot for *E. coli NuoK* protein. Bottom graph: Phobius probability plot for Mrub_1868 protein. Signal peptide probabilities for the first 20 amino acids in Mrub_1867 indicate the presence of a signal peptide sequence rather than a TMH. Plots were generated using the Phobius bioinformatic tool (Kall et al., 2004).

Evidence for highly conserved amino acids in *E. coli NuoK* and Mrub_1867 proteins was found from the pairwise alignments obtained from the Pfam website (Finn et al., 2016). Figure 23 shows the pairwise alignments for both species and indicates conservation of 2 amino acid residues between *E. coli NuoK* and Mrub_1867. Panel B of figure 23 indicates a third highly conserved amino acid (L) not conserved in *E. coli NuoK* but this is likely due to the alignment program used to adjust the placement of amino acids in sequence. A leucine residue is located in a very similar position in Panel A but it is not indicated as significant. This does not mean that amino acid is not conserved in *E. coli NuoK*, it simply indicates the program did not notice it to be significant from the alignment. Since Pfam compares query sequences to a database of consensus sequences obtained from hundreds of proteins, this is very strong evidence that these genes are orthologous to one another (Finn et al., 2016). This is also strong evidence for these proteins sharing similar function since these highly conserved amino acids between distantly related species indicate their importance in function. The HMM logos obtained from Pfam were too long to fit onto a page and include as a figure but both logos are consistent with the Pfam conserved residues in figure 23 and indicate the leucine residue missing in panel A as highly conserved.

Panel A *   *

```
#HMM     llliallfliGlygvllrrnllkvliglelisnavnlllvlagvftegtlvitnpatefadplpqalvLtaivigaaetavgLallvrlyreygtddvqel
#MATCH   l+l+a+lf++Gl g+++rrnll +ligle++ na  l++v+ag+++             +++ +q++ ++ai ++aae+++gLall++l+r +++++++++
#PP      6899******************************************.............9**************************************987
#SEQ     LILAAILFVLGLTGLVIRRNLLFMLIGLEIMINASALAFVVAGSYW-------------GQTDGQVMYILAISLAAAEASIGLALLLQLHRRRQNLNIDSV
```

Panel B *   *   *

```
#HMM     lllliallfliGlygvllrrnllkvliglelisnavnlllvlagvftegtlvitnpatefadplpqalvLtaivigaaetavgLallvrlyreygtddvqelr
#MATCH   +l+ +allf iG+yg l+rr ++ +++++el++na+ l l+ ++++t             +++ +q++vL++i+i+aae+avgL+l+v+++r ++t++v+elr
#PP      57899******************************************.............*****************************************96
#SEQ     WLVASALLFSIGAYGALTRRTAILMFLSIELMLNAAALSLISFSKLT-------------GSLEAQVVVLFIIAIAAAEVAVGLGLIVAIFRRRETTSVDELR
```

Figure 23: *E. coli NuoK* and Mrub_1867 have the same highly conserved amino acid residues for the same domain (NADH-ubiquinone/plastoquinone oxidoreductase chain 4L). Panel A depicts the alignment for *E. coli NuoK* . Panel B depicts the alignment for Mrub_1867. Both panels indicate the same 2 highly conserved amino acid residues in the "#MATCH" rows (indicated by asterisks above each panel) while panel B indicates a 3rd conserved residue missed by the alignment program in panel A. Pairwise alignments were generated using the Pfam website (Finn et al., 2016).

Figure 24 shows the PDB alignment between *E. coli NuoK* and the membrane domain of respiratory complex I from *E. coli* which confirms our previous knowledge that this protein is associated with the cellular membrane (Berman et al., 2000). Figure 24 also shows Mrub_1867

and its PDB alignment with the membrane domain of respiratory complex I from *Thermus thermophilus* (Berman et al., 2000). This sequence similarity provides further evidence that this protein is associated with the cellular membrane since its domain is aligning with the known structure of the membrane domain of the NADH: ubiquinone oxidoreductase complex in a closely related species. The difference in alignment length of the two genes is due to the fact that they matched with the proteins from different species as seen before in figure 17.

Panel A

Panel B

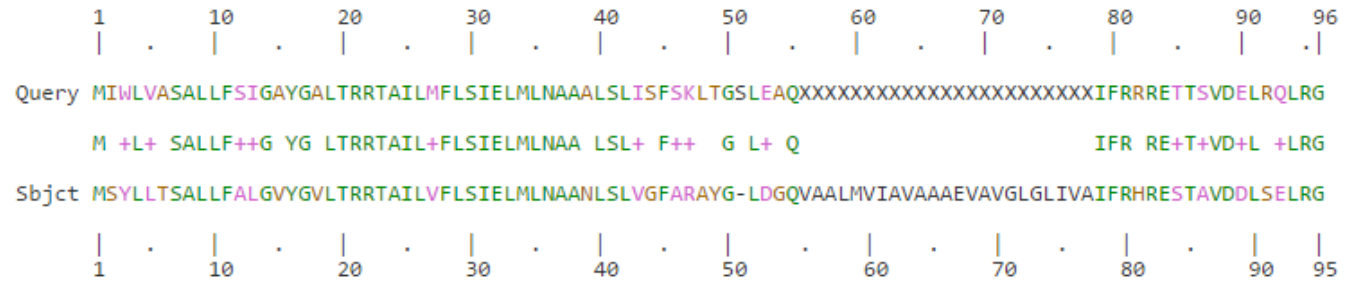


Figure 24: *E. coli NuoK* and Mrub_1867 are parts of the membrane domain of NADH: ubiquinone oxidoreductase (complex I). Panel A: PDB alignment of *E. coli NuoK* with the membrane domain of respiratory complex I from *E. coli.* Panel B: Mrub_1867 alignment with the membrane domain of respiratory complex I from *Thermus thermophilus*. Alignments were obtained from PDB website upon using query sequences (Berman et al., 2000).

Two cladograms were constructed in search of evidence of horizontal gene transfer within *E. coli NuoK* and Mrub_1867 using Phylogeny.fr (Dereeper et al., 2008). Figure 25 shows the cladograms for both species that were constructed from the b2279 and Mrub_1867 genes. Since all of the species in the b2279 cladogram were from the same phylum (Proteobacteria) and all of the species from the Mrub_1867 cladogram were from the same phylum (Deinococcus-Thermus), there is no evidence of horizontal gene transfer within these cladograms.

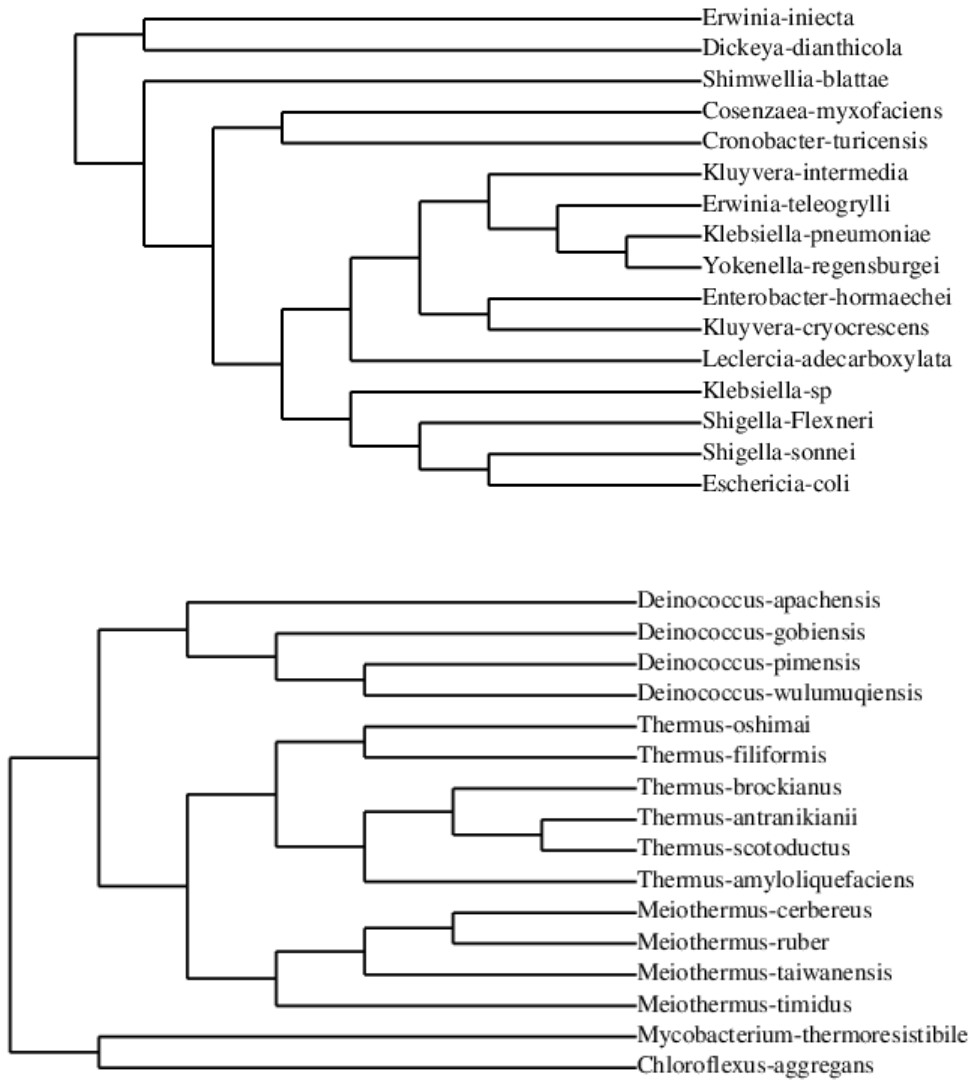


Figure 25: *E. coli NuoK* and Mrub_1867 display no evidence of horizontal gene transfer. Top cladogram: *E. coli NuoK* gene compared to other closely matched species. Bottom cladogram: Mrub_1867 gene compared to other closely matched species. Images generated using Phylogeny.fr (Dereeper et al., 2008).