

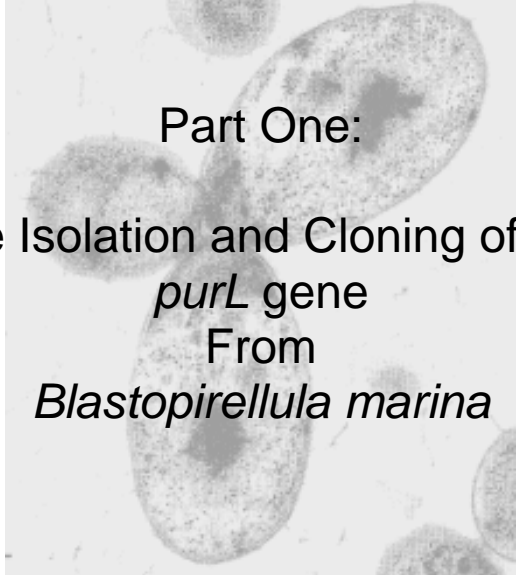
The Isolation, Cloning, and Phylogenetic Analysis of Formylglycinamide Ribonucleotide Amidotransferase

Presented to the faculty of Lycoming College in partial fulfillment of the requirements for
Departmental Honors in Biology

By:
Jessica L. Bennett
Lycoming College
April 2007

Table of Contents

1. Part One: The Isolation and Cloning of <i>purL</i> from <i>B. marina</i>	3-21
a. Background.....	4-7
b. Methods/Procedures.....	7-13
c. Results/Discussion.....	13-20
d. Conclusion.....	20-21
2. Part Two: The Phylogenetic Analysis of FGARAT.....	22-52
a. Background.....	23-27
b. Methods/Procedures.....	27-30
c. Results/Discussion.....	30-40
d. Conclusion.....	40-41
3. Appendix.....	42-47
4. References.....	48-51



Part One:
The Isolation and Cloning of the
purL gene
From
Blastopirellula marina

BACKGROUND

The *purL* gene encodes an enzyme called formylglycinamide ribonucleotide amidotransferase (FGARAT), which acts in the fourth step of the purine biosynthetic pathway (Hoskins et al 2004). In this pathway FGARAT is responsible for transforming formylglycinamide ribonucleotide (FGAR), ATP (energy source), and glutamine into formylglycinamide ribonucleotide (FGAM), ADP, inorganic phosphate (P_i), and glutamate (Anand et al 2004). Purines are the nitrogenous bases which make up DNA and RNA. The purines that compose DNA are adenine, abbreviated A, and guanine, abbreviated G; they are referred to as nitrogenous bases (Klug et al 2006).

This study was conducted with the organism known as *Blastopirellula marina*. *Blastopirellula marina* is a member of the Planctomycetes phylum (Wagner and Horn 2006). Studies performed by Schlesner et al illustrated *B. marina*'s characteristic gram negative cell wall (2004). Furthermore, *Blastopirellula marina* was observed in a rosette formation with the pole caps coming together into a central pole (Lindsay et al 1997). *Blastopiruella marina* was first discovered and isolated from giant tiger prawn tissue (Wagner et al 2006). Giant tiger prawns (*Penaeus monodon*) are marine invertebrates that thrive in brackish water (Fuerst et al 1991). Its habitat is mainly the waters of Japan and Taiwan, Tahiti, Australia, and Africa (Braak 2002).

Published research only discusses the presence of a monomeric protein form (denoted IgpurL) and a heterotetrameric protein form (denoted smpurL); but the *purL* gene from *Blastopirellula marina* is of an intermediate size of FGARAT (Sehi and Newman, unpublished results).

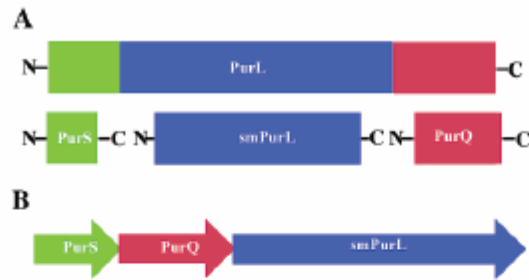


Figure 1. Diagram of lgpurL and smpurL. Figure A is the comparison between lgpurL and smpurL. The top diagram shows lgpurL organization. The green region is the N-terminal domain. The blue region is the FGAM synthetase domain. The red region is the C-terminal glutaminase domain. The diagram below shows smpurL. The green region is purS which is a dimer (meaning there are two purS's in reality). This corresponds to the N-terminal domain in lgpurL. The blue region is smpurL and corresponds to the FGAM synthetase domain in lgpurL. The red region is purQ and corresponds to the C-terminal glutaminase domain in lgpurL. smpurL is made of a 2:1:1 ratio of 2 purS:1smpurL:1purQ. Figure B shows the structure of *B. subtilis* purine operon. (Anand et al 2004)

The *lgpurL* gene produces FGARAT protein of roughly 140 kilodaltons and is found in gamma (γ) and beta (β) proteobacteria and eukaryotes (Anand et al 2004). The monomeric protein form was first purified and extensively studied in *Salmonella typhimurium* (Schendel et al 1988). In contrast, other prokaryotes are believed to contain the *smpurL* gene encoding FGARAT of 80 kilodaltons in size (Anand et al 2004). This PurL variant requires two additional subunits to function properly: PurQ and PurS (Anand et al 2004). *Figure 1* illustrates each enzyme structure in a simplified form. The monomeric protein form contains approximately 1300 amino acids and the heterotetrameric protein form contains approximately 750 amino acids (Anand et al 2004). *Figure 1B* shows the organization of FGARAT genes in *Bacillus subtilis* which has a characteristic heterotetrameric form (Ebbole and Zalkin 1986). The National Center for Biotechnology Information (NCBI) contains the sequence of the *purL* gene

from *Blastopirellula marina*. This intermediate size gene is of interest because there are no publications classifying it as either *IgpurL* or *smpurL*. Even though the sequences are in the database, one must consider the extent of the data represented for each organism. Each organism has its entire genome sequenced, and thus not every protein encoding gene has been studied and published. Therefore, if PurL from *Blastopirellula marina* is of an intermediate size, one would be interested in its 3D structure and how it compares to the monomeric or heterotetrameric form. The monomeric form has all four subunits (2PurS:1PurL:1PurQ) fused together whereas, the heterotetrameric form has each subunit as a separate polypeptide. The *purL* sequence of *Blastopirellula marina* contains 974 amino acids. This number does not correspond to either the 1300 amino acid or the 800 amino acid sequence of *IgpurL* and *smpurL*, respectively.

If *Blastopirellula marina*'s *purL* were isolated in, studies would be able to determine the 3D structure of the FGARAT it encodes. When the protein structure of the heterodimeric form is available, comparisons can be made to those of *IgpurL* and *smpurL*. The heterodimeric form of PurL is formed so that the PurL and the PurS subunits have fused to create one large "PurL" subunit. Furthermore, the heterodimeric FGARAT protein has a separate polypeptide subunit for PurQ which differentiates it from the monomeric FGARAT structure which has one continuous protein. This, in turn, allows scientists to determine the domains essential for PurL function. This information could advance understanding evolutionary trends in the formation of the different forms of FGARAT.

PurL may be a member of an, "ATP-requiring enzyme superfamily," (Hoskins et al. 2004). This superfamily is capable of using ATP to phosphorylate amide oxygens to

produce an iminophosphate intermediate (Hoskins et al. 2004). Secondly, *purL* may serve as a scaffold for the enzymes from the purine biosynthesis pathway (Anand et al 2004). Furthermore, *purL* may be responsible for helping to regulate toxin production in bacteria such as *Clostridium difficile* (Maegawa et al. 2002). Maegawa et al. proposed a linkage between toxin B production and the *lgpurL* gene in *Clostridium difficile* (2002). The toxin B produced by this bacteria is responsible for *C. difficile*-associated diarrhea (CDAD) and pseudomembranous colitis (PMC) in humans (Maegawa et al. 2002). Their experiment was able to compare the PurL amino acid sequence from *C. difficile* to *E. coli*, and *B. subtilis*'s. They observed three conserved motifs in all of the PurL sequences (Maegawa et al 2002). A conserved motif is a specific region in the amino acid sequence that is identical in all of the bacteria strains listed above (*C. difficile*, *E. coli*, and *B. subtilis*). This may spur further investigations and comparisons to see if, and how, PurL from *Blastopirellula marina* compares with the above bacterial strains. Maegawa et al demonstrated how inhibiting the purine biosynthesis pathway by introducing different reagents leads to deregulation of *purL* (2002). They discovered that sulfonamides, which produce a shortage of FGAM in the pathway, may be a way to treat *C. difficile*-associated diarrhea (CDAD) and pseudomembranous colitis (PMC).

The purpose of this experiment is to amplify and clone *Blastopiruella marinas*' *purL* gene, and in subsequent experiments to purify the protein, and eventually determine the protein's 3D structure to aid in comparisons to the large and small PurL protein forms.

METHODS/PROCEDURES

Bacteria Growth

The media recipes were obtained from the American Type Culture Collections's home website (ATCC 2006). The ATCC medium: 1657 M-14 medium contained yeast extract, glucose, modified hutner's basal salts, trizma, artificial seawater, and distilled water in amounts as shown in *Table 1* below.

Table 1: ATCC growth medium for <i>Blastopirellula marina</i>	
Media Component	Amount Added
Yeast Extract	1.0 g
Glucose	1.0 g
Modified Hutner's Basal Salts	20.0 mL
Trizma, pH 7.5	0.753 g
Artificial Seawater	250.0 mL
Distilled Water	730.0 mL

The artificial seawater and modified hutner's basal salts were prepared separately as stated on ATCC's website.

The artificial seawater was prepared to mimic *Blastopirellula marina*'s natural

brackish water habitat. Agar was added to a portion of the medium in order to allow plates to be poured and to solidify. The *Blastopirellula marina* was obtained from ATCC (catalog #49069). The bacteria were resuspended in test tubes which contained the ATCC growth medium. The bacteria were grown in an incubator placed at 30.0°C. After two days, liquid medium was pipetted onto the ATCC agar and spread. These cultures were allowed to incubate at 30.0°C for two days.

Gram Staining

The procedures followed were taken from the 2004 Biology 110 Laboratory Manual (Newman 2004). A glass slide was cleaned with distilled water and bacterial colonies grown on the ATCC medium were placed in a drop of liquid on the slide. The slide was allowed to dry at room temperature. The bacterial cells were then heat fixed by passing the slide through the blue flame quickly three or four times. Crystal violet

stain obtained in the microbiology lab was applied for one minute and rinsed off; followed by gram's iodine stain obtained in the microbiology lab for one minute (Newman 2004). Gram's iodine stain was subsequently washed off and 95% Ethanol decolorizer was used (Newman 2004). Then the ethanol was immediately washed off with distilled water. In the final step, safranin, obtained from the microbiology lab, was applied for one minute and washed off (Newman 2004). The slide was then put between bibulous papers and blotted dry. The cells were examined under oil immersion using the 100X objective of a brightfield light microscope. Observations of cell shape, and color were written in the laboratory notebook. No pictures were obtained.

Primer Design

Primers were designed following standard rules: approximately 18-22 bases of complementarity sequence; must have a 50% G-C content; must have a G or a C at the 3' end of the primer sequence. The primer sequences were designed following the DNA sequences found at the National Center for Biotechnology Information (NCBI) for the *purL* gene (Protein ID #: EAQ77872.1). The start primer used was as follows: Bm *purL* start NdeI→5' GGGCATATGACGCTGTGGGAAATTGAC. The stop primer used was as follows: Bm *purL* stop HindIII→5' AAAAAGCTTACCAGTCAAGCGGCGCGAG. The desired sequences were sent out to Sigma-Genosys for production.

PCR and Gel Electrophoresis

DNA was extracted from *Blastopirellula marina* from liquid culture. Cells were freeze thawed as follows: Liquid medium containing the bacteria was pipetted into a microcentrifuge tube and centrifuged at 13,000rpm for one min. The supernatant was removed by pipetting, leaving the bacterial cells in the bottom of the tube. The cells

were re-suspended in distilled water and transferred to a small microcentrifuge tube. One heating block was heated to 70.0°C and another was cooled to -70.0°C. The cell culture in the microcentrifuge tube was placed in the cool heating block until the solution was frozen. The tube was then instantly put into the hot heating block until completely thawed. This was repeated for two freeze thaw cycles. One freeze thaw cycle included one freeze and one thaw.

PCR primers from Sigma-Genosys were reconstituted in distilled water to a 100 micromolar concentration. For PCR, a 5 micromolar dilution of primers was used. The PCR reactions were conducted according to procedures in the 2005 Genetics lab manual, "PCR Amplification and Cloning of the Human Clotting Factor IX Gene." The PCR mix was placed into a 0.5 mL microcentrifuge tube with a 2X Premix (contained:

Taq Polymerase, buffer, dNTP's), start primer, stop primer, distilled water, freeze thawed *B. marina*, and mineral oil (to prevent evaporation). The microcentrifuge tube was placed in the thermocycler located in the Microbiology Research Lab at Lycoming College and placed on a preset program as outlined in Table 2. Primer

Table 2. First Cycle PCR		
Stage 1 - 1 cycle		
Process	Time (min)	Temperature (°C)
denature	3	94
anneal	1	55
primer extension	1	72
Stage 2 - 35 cycles		
denature	1	94
anneal	1	55
primer extension	3.5	72
Stage 3 - 1 cycle		
denature	1	94
anneal	1	55
primer extension	10	72

extension times, however, were extended to three minutes instead of the one minute outlined in the Genetics Laboratory Manual (Newman 2005). PCR products were run on a 1% agarose gel. For subsequent PCR's the annealing temperature was raised to 65°C to try to obtain a more specific product.

PCR Insertion into pCR2.1 TOPO cloning vector and *E. coli* transformation

The gel fragment located between lambda BstEII marker fragments 2,323 bp and 3,675 bp was cut out of the gel and chopped with a razor blade to fine pieces. The PCR fragment was ligated into the pCR2.1 TOPO cloning vector following the procedure outlined in Invitrogen's manual "TOPO TA Cloning – Five minute cloning of Taq Polymerase – amplified PCR product," (2006). The PCR/vector was transformed into supercompetent TOP10F' *E. coli* cells from Invitrogen (catalog #C3030-03) and spread onto an agar plate containing the antibiotic kanamycin (0.5 mg/mL) (Invitrogen 2006). The plate was incubated at 37.0°C for 16-20 hrs. Blue/white screening was then performed to locate supercompetent *E. coli* that were transformed. White cells were the cells of interest. White cells indicate the *E. coli* took in the pCR2.1 TOPO cloning vector with a gene insert in the multiple cloning site incorporated in the vector. If the vector took in a PCR product, the gene separated the lacZ gene from its corresponding operon which did not allow it to transcribe the β -galactosidase protein, resulting in a white colony. Otherwise lacZ would be able to transcribe β -galactosidase protein which would react with Xgal impregnated in the agar plate to turn a colony blue. The cells which remained white (ten were picked) were isolated and re-inoculated into LB medium and grown overnight at 32°C. To extract the plasmids from the *E. coli*, procedures outlined in Qiagen's Miniprep Handbook were followed on pages 22-23: "QIAprep Spin Miniprep Kit Protocol (2002). Once the TOPO vectors were re-isolated, they were subjected to a restriction digest with EcoRI for an hour and then run on a 1% agarose gel at 100 volts (V) for fragment length confirmation. Bands were expected in all lanes at 3,900 bp and 2,900 bp in length.

DNA sequencing

All sequences were sent to GeneWay Research for sequencing. The sequences to be analyzed were the Miniprep plasmids. GeneWay Research used T7 primers which bound upstream to the multiple cloning site (see region in appendix *Figure 3*) in the pCR2.1 TOPO cloning plasmid. GeneWay used the dideoxytermination method for DNA sequencing. Into the sequencing mixture they added the T7 primer, buffer, Taq polymerase, and dideoxynucleotides (ddNTP's) which will halt the addition of subsequent nucleotides to the DNA strand. These ddNTP's are fluorescently tagged so adenine (A), guanine (G), cytosine (C), and thymine (T) fluoresce a unique color depending which nucleotide was added. Once the reaction was completed a computer compiled the sequence overlaps to make one long continuous sequence. The ddNTP's were altered so their 3' -OH was substituted to have an H so no nucleotide could be added to extend the DNA sequence.

BLAST search for the Miniprep gene insert sequences

BLAST searches were conducted using the National Center for Biotechnology Information's (NCBI) database with the DNA sequences returned from GeneWay Research. BLAST searches were performed to see what proteins showed homology to the gene insert sequences. To conduct a BLAST search of the entire database one must go to the National Center for Biotechnology Information's website (NCBI 2007). On the top of the webpage the "BLAST" option was selected. On the subsequent pages, "blastx" was selected. A "blastx" search compares a nucleotide sequence to a protein sequence (NCBI 2007). The gene insert sequence was then copied and pasted into the query box and format was selected to perform the database search. Before the

sequences could be copied into the query, they had to be altered to remove plasmid sequences. To accomplish this, in EditSeq the “find” option was selected and “gaattc” was typed in. This sequence was used as a result it is the EcoRI restriction site sequence. EcoRI is the restriction enzyme which would be used to cut the gene insert from the plasmid since the site is present on either side of the PCR insert.

RESULTS/DISCUSSION

PCR to obtain the purL gene

The primers used to obtain the PCR product were designed to incorporate cleavage sites for the restriction enzymes NdeI and HindIII.

The start primer containing NdeI bound to the beginning of the *purL* sequence whereas, the stop primer containing the HindIII bound to the end of the *purL* sequence. Figure 2 shows the PCR products from DNA isolated from the freeze-thawed

Blastopirellula marina cells. The top band in lane one appeared between

the marker bands at: 3,675bp, and 2,323bp, and was estimated to be about 3,000bp in length. The *purL* gene from *Blastopirellula marina* was 2,925bp in length as reported by NCBI.

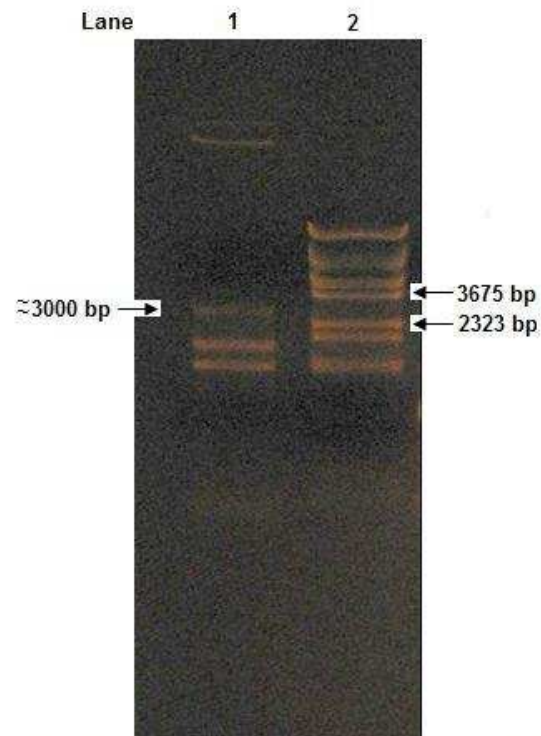


Figure 2. Gel photograph of *Blastopirellula marina* *purL* PCR product in a 1% agarose gel. Lane 1 contains the PCR product from *Blastopirellula marina* *purL* and Lane 2 contains DNA marker lambda BstEII. The left arrow points to the location of the *purL* band on the gel.

See *Figure 2* (Appendix) for protein sizes and subunit structure. Appendix *Figure 2* depicts the heterodimeric PurL, and compares its protein structure to monomeric PurL and heterotetrameric PurL (Newman 2000). One can notice the monomeric form has all the subunits fused. The heterotetrameric form has all the subunits as separate polypeptides. The heterodimeric form has PurS and PurL fused with a separate polypeptide subunit for PurQ.

Plasmid map with purL insert into pCR2.1 TOPO vector

When cloning into an expression vector such as pET28-a, one must first use an intermediate vector such as the one used in this study: pCR2.1 TOPO vector. The pET-28a expression vector does not have a β -galactosidase gene to encode the β -galactosidase protein which reacts with the Xgal impregnated on agar plates. Therefore, without the use of an intermediate vector such as pCR2.1 TOPO vector, blue/white screening would not be possible in order to determine which vector had picked up a gene insert. Using the intermediate vector such as pCR2.1, one would be able to select for vectors which contained the desired gene insert.

Figure 3 is a plasmid map of BmpurL ligated into the pCR2.1 TOPO cloning vector's multiple cloning site (MCS). The pCR2.1 TOPO cloning vector is 3,931bp in length. Inserting *purL* from *Blastopirellula marina*, 2925bp in length, increased the plasmid size to 6856bp. The map diagramed restriction enzyme digestion sites for EcoRI, HindIII, and NdeI. EcoRI restriction sites from the pCR2.1 TOPO cloning vector should have cut around the gene insert. HindIII and NdeI were incorporated into the PCR primers and are thus diagramed into *Figure 3* in addition the EcoRI present in the multiple cloning site of pCR2.1 TOPO vector.

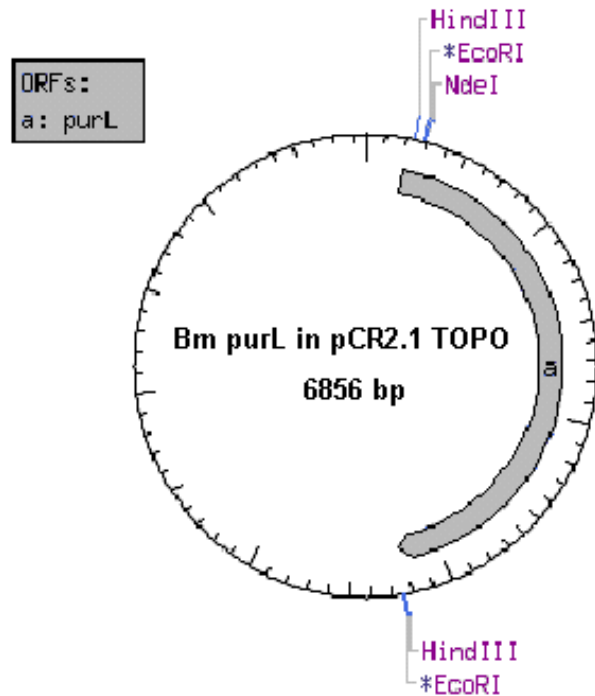


Figure 3. *Blastopirellula marina* ligated into pCR2.1 TOPO clone vector. pCR2.1 TOPO comes linearized with sticky ends with a T overhang. Taq Polymerase attaches an A on the end of the PCR fragment allowing ligation into the vector. Bm purL will be cut out of pCR2.1 TOPO with the restriction enzymes: EcoRI, HindIII, NdeI as diagramed above.

The vector was transformed into TOP10F' Supercompetent *E.coli* and grown on agar plates containing Xgal, kanamycin, and IPTG. Blue/white screening was performed following an overnight incubation period at 37.0°C. White colonies were observed on the plate that had 100 microliters of the TOP10F' transformed *E. coli* spread on it. There were seventeen white colonies on the agar plate. The bacteria that grew were only the ones which were transformed with the pCR2.1 TOPO vector because this

plasmid contained a gene coding for kanamycin resistance. There were numerous white colonies which contained blue centers.

Restriction digest with EcoRI

Plasmid DNA isolated from the white colonies were digested with restriction enzyme EcoRI and run on a 1% agarose gel. Many different bands were observed; indicating that the pCR2.1 vector took up inserts other than *purL*. The pCR2.1 vector was distinguished from the insert because it could be found at ~3900bp on the gel each time since EcoRI cut the insert out leaving the now linearized vector intact. One colony had a band on the 1% agarose gel at roughly 3000 bp indicating it could possibly have been *purL* isolated from *B. marina* cultures.

Once the plasmid preps were cut with EcoRI, 2 bands appeared on a 1%

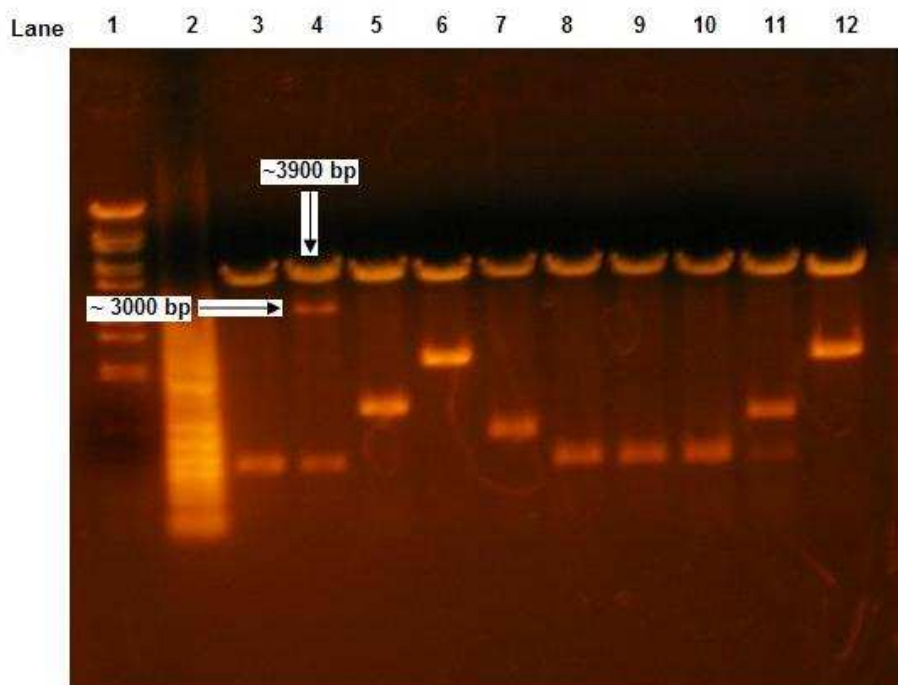


Figure 4. Mini preps cut with EcoRI and PCR product of gel product. Lane 1 contains lambda BstEII marker. Lane 2 contains the PCR product. Lanes 3-12 contain the mini preps in the order labeled MP1, MP10, MP2, MP3, MP4, MP5, MP6, MP7, MP8, MP9, MP10 where MP = mini prep. Products were run on a 1% agarose gel. The arrow pointing to a band located in lane 4 at ~3000bp is the *purL* suspect. The top band on all lanes 3-12 is the pCR2.1 TOPO cloning vector.

agarose gel roughly 3,900 bp and 3,000 bp in length. Many other bands were also seen in lanes 3-13. *Figure 4* depicted the plasmid preps cut with EcoRI. The pCR2.1 TOPO vector was completely digested in every lane. As *Figure 4* depicts, the vectors had many different size gene inserts. DNA cut from the gel lanes 3-7 were sent for sequence analysis. Lane 1 contained λ BstEII marker. Lane 2 contained the PCR product run with a higher annealing temperature (65°C) in an attempt to get a more defined PCR product. The PCR product had no distinct band pattern. Lanes 3-13 contained the mini preps done from the plasmids isolated from the supercompetent TOP10F' *E. coli* cells.

Sequence results of minipreps

GeneWay Research completed the sequences and sent them back November 19, 2007. None of the sequences from a BLAST search resulted in any significant similarity to any protein in the database. *Figure 5* is the DNA sequence sent back for the gene insert from lane 3 in the EcoRI gel (*Figure 4*).

```
GCCCTTAAAAAGCTTACCAGTCAAGCGGCGAGCCACGGGTATACAGGTGAGCAGCGCGGAGGCCACCGGCGCGGTGCG
CCACAGAATGTGCTGGATTGACGCAGACAACCTCGGCATGCGCTTCGCCCCGTGTTTGCGAGTCAGCATAACGAAAGGCA
CGCCGCAGAAAATCTCGTAATAGGCCGGCAGGGCCGAGAGTTTATGGCGTTCTATTCCCGCTGCAGGCAGGATTCAGGCG
CAGCTGCTTCTCAGCGCCGAAATCTGCGGCGACACTGCCAACCCCGTACGCGCGGAAATAGCCATCCGTCGCGAGAG
ATCTCCGATGCAACTCGACGCAATCGACTTACGCATTCTCAACCTGGTGCAGAAGAAGCAGTCAATTTCCACAGCGTCAT
ATGCCAAGGGC
```

Figure 5. DNA sequence from lane 3 on the miniprep gel (*figure 4*) restricted with EcoRI. The sequence was altered to remove the first 59 nucleotides and the last 413-845 nucleotides. This cut off the plasmid sequence that was incorporated into the sequence of the gene insert. The gene insert restricted out of the pCR2.1 TOPO vector was 412 bp in length which is approximately 137 amino acids in length.

The DNA sequence had to be altered so as to obtain the gene insert. As seen in *Figure 3* above, there are EcoRI cut regions on both ends of the insert. In EditSeq (part of the Lasergene computer package), “search”, then “find” were selected followed by the typing of “gaattc” and hit “find”. The program found the two EcoRI sites. The sequences

before and after these restriction sites were deleted to obtain the gene insert sequence. Nucleotides 1-59 and 413-845 were removed to leave a gene segment 412 bp in length. This corresponded to 137 amino acids. This would support its placement in the gel being below the 702 band in the marker lane (lane 1 *Figure 4*).

Figure 6, was the DNA sequence from the gene insert found on *Figure 4* from lane 4. Nucleotides from the 471st position ranging to the 887th were removed to obtain the actual gene insert DNA sequence. The resulting gene insert was 470 bp in length which corresponded to 157 amino acids.

```
TGGGCCTCTAGATGCATGCTCGAGCGGCCCGCCAGTGTGATGGATATCTGCAGNAATTCGCCCTTAAAAAGCTTA
CCAGTCAAGCGGCGCGAGCCACGGGTATACAGGTGAGCAGCGCGGAGGCCACCGGCGCGGTGCGCCACAGAA
TGTGCTGGATTGACGCAGACAACTTCGGCATGCGCTTCGCCCCCGTGTTTGCGAGTCAGCATAACGAAAGGCACG
CCGCAGAAAATCTCGTAATAGGCCGGCAGGGCCGAGAGTTTATGGCGTTCTATTCCCGCTGCAGGCAGGATTCA
GGCGCAGCTGTCTTCTCAGCGCCGAAATTCTGCGGCACACTGCCAACCCCGTACGCGCGGAAGTAGCCATCC
GTCGCGAGAGATCTCCGATGCAACTCGACGCAATCGACTTACGCATTCTCAACCTGGTGCAGAAGAACAGTCAA
TTCCACAGCGTCATATGCCCAAGGGC
```

Figure 6. Gene insert DNA sequence from figure 4 lane 4 miniprep gel cut with EcoRI. Nucleotides ranging from the 471-887 position were removed to obtain the gene insert DNA sequence. This leaves a gene insert with a length of 470 nucleotides which corresponds to 157 amino acids in length.

Figure 7, was the DNA sequence from the gene insert found on *Figure 4* from lane 5. Nucleotides ranging from the 1st to the 58th positions were removed to obtain the gene insert DNA sequence. This resulted in leaving a gene segment with a length of 850 bp which corresponded to 283 amino acids. The length of this gene would correspond to where it migrates in the gel (*Figure 4*) in that it should be larger than both the genes from lanes 3 and 4 because it did not migrate as far through the gel. The smaller the gene size the farther it should be able to migrate through the pores in the gel.

```
GCCCTTAAAAAGCTTACCAGTCAACGGCGGAGAGATGATTCGGATTTACGGGGCGCTGCGGAATAATCCAGCTGTAGGC
CCCTTTATCTGCAAATTAGTGGACTTCCAGGCTTTCACGTCCGCCATGCTCTTACTGCTCAACCTCTGTGGCTATTACA
GCAACACCGCGGTAGTAACTCTCAACCACCGGACCTCGAGCAAGACCAGAGAGATTCTGAACTCATTGACCAGACGATAG
GCCTGCTCAAGGATGCCACCCAAGAAGGTGGAGGAGTTGTTGCTGCACAAAGCGCTCAAGCCCTTGAGATGCTTGCTCGT
GCACGACATTGCTCAGAATCAGACATCAAAGATCGTCCTGCTGGAACCTGTCAAGTTTCCATTCCGTATTTCCGGACAAT
ATCAATCGGCATGGGCAAGCAGTTCATTCCCATCAAGCCTGGAACGTATGTTCAACGTTTCATCTAATTCTACTGTAACCTG
TTCCGACGAGGGGAGCCCCCAACACAGGGTTACCCACTCCTCCATCCATCACTTCGAGCAGCACGCAGCCATCTCCAT
CACCCTACCATCGACAATCCTCACCTGCAGCCCCCTCAAGCGCACAGNACTACCAGTATACGAGTCCACAGACTTCATA
GCCCCGGGCTTGACAAAATTGGCCTCAAGGCGACGACCCTTTGGTTACATTGATAGCTTCATGTCTTTCNCACCTCA
NAATATGCAAGACTATCCGTCTGCTGGTGGCACTGCAACCANCANTTCAGGTTTCACTCCAACAANATCTCAAGTCANT
TCCCANGCGTCATATGCCCAAGGGCGANTCNCACANNNGGCGNGTTACN
```

Figure 7. Gene insert from DNA sequence taken from the figure 4 EcoRI cut gel lane 5. Nucleotides 1-58 were removed to yield a gene insert with a length of 850 bp. This corresponds to 283 amino acids.

Figure 8 was the DNA sequence from the gene insert found in Figure 4 from lane

6. No sequence alterations had to be performed. The final gene insert was a length of 856 bp which corresponded to 285 amino acids. However, one would expect the gene to be much larger compared when referring to its location in lane 6 on the gel (Figure 4). During sequencing it was possible the polymerase experienced a premature stop codon and thus the sequence was not continued and the gene length was truncated.

```
CNATGGGCCCTCTAGNATGCATGCTCGAAGCGGCCGCCAGTGTGATGGATATCTGCAGNAATTCGCCCTTAAAAAGCTTA
CCAGTCAAGCGGCGCGAGATCTCAACTGGGTCGTGGTTGCCAGCGATTCCCCGTCTGGGTGCGTCTCGATAAACCCCCG
CTGGATGCCATGCATGTGGGCATGACCGCTCGGTCAAAGTGCATCATGACATCCTCCGCTGAGCCGGATCGAGGGCTGC
TCTCGTTTTCCCGTTCTCGCGCGAGAGCTTACGCCGCGGGAGGGGAGAGGTGCGGCCGTAGCCCGCATCAGTATCGGA
TGCGCCGTACGGTAGTCAATTGCAATGCTATTTGCGATCCCCGAGCCGGCTACATGGCCTACGTGGTTTTTCTCATCAG
CAAGGACGAGCAATCCGCCACCATCAAGTCGGCGTTGGGCGGCTTCGTTGCCGTGACATTGGCCGTGATCCTTTGCCTGC
TGCTTTTCATGCTCGATGCGGCGGAGCCTGCATTGAGGCTTCTGCCATGGCCCTGATGACTTTCGGTGCGATGTACACC
TCGAGAATTTTTGCGCTGGGTCNATTTCTTTCTCGCGGGGATTCGTGCTGGTGCTGGTGCAATCCCTGATCGANNAATC
CCTATTTGNANCGCTGANANGCACACTATGGCTATGGTCGTCNTACTCGTACCGGGNTCATCGCCNCGGGACAACCTT
TNGCGGGCANCCNCCCTGGTTACCAANGTANGNCTTCGGCANNAANCCGTCAANCAGNTNANGNTCACCNACNCAAANT
TGNCAAACCGNGGNAATATGGGGTTTCTTNACGANNTGGGTCCANCAAAAAGNAC
```

Figure 8. Gene insert from DNA sequence taken from figure 4 EcoRI cut gel lane 6. No sequences alterations were needed because no EcoRI restriction sites were found in EditSeq. The gene length was 856 bp long which corresponds to 285 amino acids.

Figure 9 was the DNA sequence from the gene insert found in Figure 4 from lane

7. By looking at the gene inserts location on Figure 4, one would hypothesize the gene length to be slightly smaller than the one found in lane 5, but larger than both the sequences observed in lanes 3 and 4. Thus one would hypothesis the gene length to be between 412-850 bp in length. The gene insert sequence compiled by GeneWay Research confirmed this hypothesis. The gene insert was a length of 621 bp which

corresponded to 207 amino acids. This gene insert fell between the hypothesized values. Nucleotides ranging from positions 1-57 and 622-843 were removed to yield the 621 bp long gene insert.

```
GCCCTTAAAAAGCTTACCAGTCAAGCGGCGGAGGAGATTAAAGCAATGGATAGATCGTCCATTAATTCATCAACAACAA  
ATAAACGCGCGTTTAGATACGGTTGAACAATTTATTACGCATTTTCATCGAACGTGATACTTTGCGTGGATATTTAAACCA  
AGTGTATGATATTGAACGTTTAGTTGGACGTGTGAGTTACGGTATGTAATGCACGTGATTTAATCCAACCTAAACATTC  
TATTATGGAAATCCCTAATATCAAAGCTTTGCTTAATGATTTTCGATGAGAAAATGCCTGCCCATTTTGAGGCATTAGAAC  
CGTTGGATGATTTATTGACTGTATTAGAAAATAGTATTGTGGAAGAACCACCAATTTCTGTAAAAGATGGTGGCTTATTC  
AAAAAAGGATTTAACGAGCAACTCGATGAATATTTAGAAGCATCACAAAATGGGAAGTCTTGGTTGGCTGAGTTACAAAG  
TAAAGAACGACAACGTACCGGGATAAAATCACTTAAAATAAGTTTCAATAAAGTATTTGGTTACTTTATTGAAATTACTA  
GGGCCAATTTACAGGGGATTTGAGCCGAGTCAATTTCCACAGCGTCATATGCCCAAGGGC
```

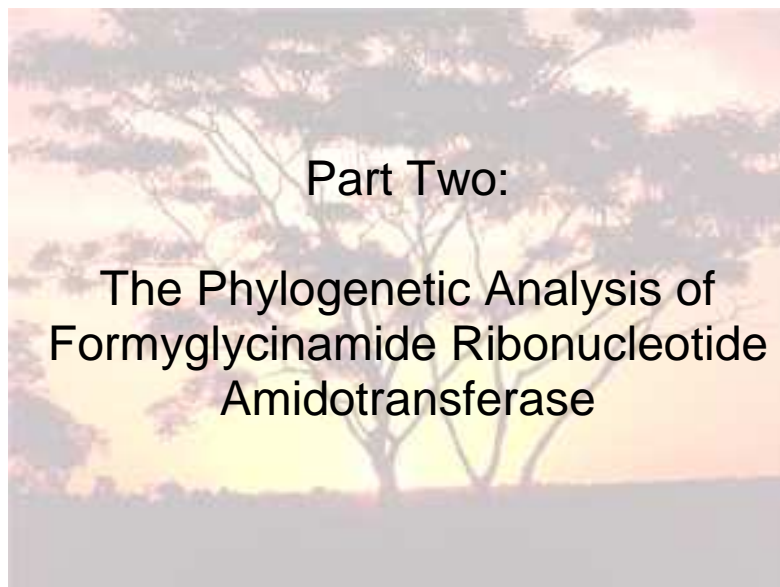
Figure 9. Gene insert from the DNA sequence taken from figure 4 lane 7. The sequence had to be altered in 2 locations. Nucleotides 1-57 were removed and nucleotides 622-843 were removed. Thus the resulting gene insert was a length of 621 nucleotides which corresponds to 207 amino acids.

CONCLUSIONS

The *BmpurL* gene was not successfully cloned. The PCR's did not give a specific band on a 1% agarose gel. This may be a result that *B. marina* was never cultured. The *B. marina* which was used was a freeze dried sample from the American Type Culture Collection (ATCC). The vial of cells was over a year old as the project was originally designed to be carried out a year earlier. Thus the cells may not have been viable. To support this conclusion, during initial wet mount observation of the cultured cells, the characteristic rosette formation was not observed. All the cells were free and not clustering under the 40X objective. As an alternative theory, the primers may have not bound specifically to the *purL* gene. To try to obtain a specific PCR, the annealing temperature was raised up to a maximum of 65°C and still the PCR product was still not producing specific bands. BLAST (blastn) searches performed within the *B. marina* genome revealed the primers had the potential to bind in other places along the genome. A BLAST search of *BmpurL* start NdeI showed it matched up 100% to the

purL gene sequence. The primer sequence (twenty-seven nucleotides) identified with twenty nucleotides along the *purL* gene. The next closest homology of 100% and fourteen identities was to a 60 kilodalton outer membrane protein. However, this gene was only 2258 bp long and does not correlate to the PCR fragment seen in *Figure 2*. The BLAST search did bring up a gene with 3044 bp in its sequence and it had a 100% sequence homology with twelve identities out of the twenty-seven from the primer sequence. This gene was the chemotaxis protein CheA. The BLAST search for Bmp*purL* stop HindIII generated more hits on possible primer binding sites within the genome. The first BLAST result showed a 100% sequence homology to the *purL* gene. Twenty three out of twenty-eight nucleotides bound to the *purL* gene. None of the other hits had nearly as many identities along its sequence.

In the future one may be more effective in amplifying *purL* from *B. marina* if one were to wait for the genome sequence to be completed so one would know the sequence published was as correct as it was going to be. *B. marinas'* genome is still in progress of being sequenced by the J. Craig Venter Institute. In addition, it may be beneficial to work with new cells rather than a vial of freeze dried cells which were sitting around for over a year. Future studies would also include the ligation of *purL* digested from pCR2.1 TOPO cloning vector into pET28-a. pET vectors are an extremely effective system to clone and express proteins into *E. coli* (Novagen 2006). The use of pET28-a will allow PurL protein to be expressed in high quantities in order to ease the isolation and study of FGARAT protein.



Part Two:

The Phylogenetic Analysis of
Formylglycinamide Ribonucleotide
Amidotransferase

BACKGROUND

Phylogenetic analysis is a powerful tool implemented to study the evolutionary relationship among organisms. Freeman et al (2007) defines a phylogenetic tree as

“A diagram (typically and estimate) of the relationships of ancestry and descent among a group of species or populations; in paleontological studies the ancestors may be known from fossils, whereas in studies of extant species the ancestors may be hypothetical constructs.”

Phylogenetic trees can be used to study the evolution of many aspects of biology including, but not limited to the evolution of a specific gene, species, or trait. A phylogenetic tree is organized based on grouping similar characteristics, in this studies case: nucleotide sequences, together into branches. The branches are connected by nodes which are indicative of a common ancestor where the branches split into different lineages in the evolutionary history. Synapomorphy is a term used to describe homologous traits (Freeman 2007). However, not all shared traits are synapomorphic even if they appear to be so. Many organisms are able to evolve certain characteristics as a result of being exposed to a common environment. For example, crocodiles and hippopotamuses both evolved a trait for their eyes to develop on the top of their heads (Freeman et al 2007). These two animals are completely unrelated taxonomically. However, as a result of spending the majority of their life submerged in water, they were both forced to adapt to their aquatic lifestyle. This results in what is termed convergent evolution. Freeman et al defines convergent evolution specifically as a, “Similarity between species that is caused by a similar, but evolutionarily independent, response to a common environmental problem,” (2007). In the context of genetics, two or more

organisms may evolve a gene to encode the same protein. These proteins appear homologous however there are distinct differences in the nucleotide sequence and slight differences in their protein structures the genes encode. Further investigation of the proteins would allow the investigator to examine whether the proteins evolved independently to solve a similar problem in their environment.

One of the largest problems that arise with the use of phylogenetic analysis is the tree constructed might not be the correct picture of evolution. There are thousands if not millions of ways to group organisms on a tree depending on the sample size of what is being compared. In addition, a scientist trying to construct an evolutionary tree must take into consideration reversals and convergent evolution. A reversal occurs in DNA when a nucleotide reverts back to an ancestral form (Freeman et al 2007). Reversals can lead a researcher to think two traits are more homologous than they actually are. However, to aid researchers in narrowing down the possible phylogenetic tree, one would utilize the Law of Parsimony. The Law of Parsimony allows a researcher to select the tree which has the least number of changes so as to lower its complexity (Freeman et al 2007). Today, the data is uploaded into programs such as Q-Align and MegAlign (from the Lasergene package). These computer systems compile all the possible phylogenetic trees using a series of algorithms and then it picks out the most parsimonious tree. The computer programs are able to perform bootstrapping as well to help the researcher evaluate the probability a certain node actually exists.

Variation in genes is a result of many different phenomena including, but not limited to chromosomal inversions, gene duplication events, transposons, addition of nucleotides and gene segments via unequal cross over, and horizontal (lateral) gene

transfer (Freeman et al 2007). Horizontal gene transfer is by far the most significant factor in prokaryotic gene diversity (Freeman et al 2007). Horizontal gene transfer is the transfer of one or more genes from an organism to another usually not related to the first. However, it is extremely difficult to prove because there is no record of it occurring other than what has been examined through genomic studies (Koonin et al 2001). Freeman et al proposes four mechanisms responsible for horizontal gene transfer (2007). Their research suggests viruses are capable of moving genes between prokaryotic species via transduction, plasmids transfer genes via conjugation, transformation results in gene transfer, and endosymbiosis can be responsible for the transfer of genes (Freeman et al 2007). Transduction is the process by which viruses are able to enter a bacterium and parasitize it while transferring new genes into its genome (Freeman et al 2007). Plasmids are circular pieces of DNA often found in bacteria that house a few genes some of which include antibiotic resistance genes. A plasmid can be transmitted from an F+ bacterium (has a fertility plasmid) to an F- bacterium (has no fertility plasmid) via a sex pilus. Freeman et al state bacteria may conjugate between species (2007). If the DNA (gene) is picked up directly from the environment and inserted into the host genome, this is termed transformation (Freeman et al 2007). This only occurs if the host cell does not metabolize the incoming DNA to use it as a nutrient source. In eukaryotes, endosymbiosis is responsible for horizontal gene transfer (Freeman et al 2007). Chloroplasts (in plants) and mitochondrion are a prime example of symbiotic events where introduce new genes different organisms.

Bacteria are a prime type of organism capable of horizontally transferring genes between species. Most often bacteria in nature are found in biofilms. These are

microenvironments where different bacterial species live in competition on a surface and fight for nutrients. *Figure 1*, is an example of a biofilm. As one can notice the

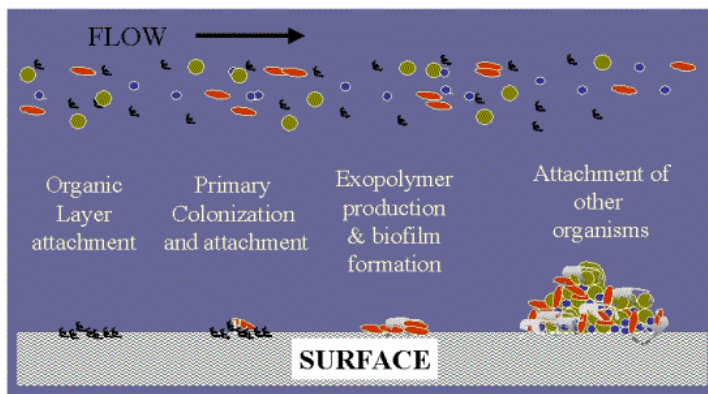


Figure 1. Biofilm formation resulting from an initial colonization of a bacterium. This bacterium then begins to grow and secretes substances which attract further microbes which attach. Competition will result in the formation of pores where bacteria were killed as a result of toxins (antibiotics) produced by other microbes in the film. (Saieb et al)

bacteria are formed in layers (result of competition) on the surface where they are constantly in contact with one another providing the perfect situation for horizontal gene transfer. The normal flora in the intestine of humans is another

example where a biofilm was formed and creates another opportunity for horizontal gene transfer to occur from eukaryotes to prokaryotes and vice versa (Koonin et al 2001). However, Koonin et al point out that prokaryotes may have trouble splicing out eukaryotic introns since they do not process the machinery necessary to perform this function (2001).

The analysis of horizontal gene transfer can be difficult. Freeman et al state, “Anomalous placement of particular genes on a phylogenetic tree can furnish strong support for LGT,” (2007). Phylogenetic analysis of *Thermotoga maritima* and *Escherichia coli* have shown 25% and more than 15% of their genomes respectively are a result of horizontal gene transfer (Gogarten et al 2002). Gogarten et al state horizontal gene transfer may be evident when gene sequences from distant and unrelated taxa group together on a phylogenetic tree (2002). Furthermore, Gogarten et al state, “Genes encoding core metabolic functions, conserved biosynthetic pathways, components of the transcription and translation machinery, and even ribosomal RNA

have been subject to HGT,” (2002). This would include the purine biosynthetic pathway.

One would expect from a phylogenetic analysis of FGARAT to observe all the monomeric, heterodimeric, and heterotetrameric gene forms to cluster together on the phylogenetic tree. The purpose of this experiment is to study the evolutionary relationships between the three forms of FGARAT.

METHODS/PROCEDURES

Compiling of Completed Microbe Genome chart: PurL, PurS, and PurQ protein sizes

Before any analysis could have been completed, the PurL, PurS, and PurQ protein subunit sizes had to be compiled into one list. The list was compiled in Excel with column headings, “Organism Name,” “Taxonomy,” “Subunit,” “PurL,” “PurQ,” “PurS,” “Order,” “Source,” and “Notes.” The list was started by Elizabeth Sehi when she originally started the phylogenetic analysis of FGARAT in 2006. However, since Sehi’s graduation, the completed genome list in NCBI’s database was almost doubled in size. The goal was to go through each organismal name and add the name to the excel spreadsheet if it was not already there from Sehi. To accomplish this task, the NCBI’s website was accessed (NCBI 2007). On the left margin of the home page, “Genomic Analysis” was selected. Then on the right of the following page under “Genome Resources” “Microbial” was selected. The page which popped up was the list of all the completed genomes to date. The list was searched through organism by organism. If an organism not on Sehi’s list was found then it was added to the list. To find the organism’s taxonomic classification, number of FGARAT subunits, PurL, PurS, PurQ protein sizes, and protein subunit order the “RefSeq” link was selected. The taxonomy

was recorded all the way to the organism taxonomic classification: order (-ales). If a COG table was present this link was selected followed by “Nucleotide transport and metabolism.” PurL (COG0046), PurQ (COG0047), and PurS (COG1828) protein size, subunit order, and the type of subunit the perspective organism contained could be identified from this link. However, if a COG table had not been compiled then a BLAST search had to be performed within the perspective organism’s genome. From the “RefSeq” link BLAST was selected. Starting with the PurL sequence from *Sinorhizobium meliloti* (organisms PurL, PurQ, and PurS was used for all organism BLAST searches), the sequence was pasted into the query box. Based on the PurL size found from this search one could infer on the number of subunits composing FGARAT. If PurL came out to be approximately 750 amino acids then most likely the organism had a heterotetrameric subunit form so BLAST searches for PurS and PurQ had to be conducted. If PurL came out to be on the order of 1200 amino acids then the organism had the monomeric form and PurS and PurQ searches did not have to be conducted. If PurL came out to be nearly 900 amino acids then a search for PurQ had to be conducted. To find the gene sizes on the BLAST searches, one clicked on the reference sequence link to the left of the protein name if it was PurL (phosphoribosylformylglycinamide synthase I), PurQ (phosphoribosylformylglycinamide synthase II), or PurS (phosphoribosylformylglycinamide synthetase). Gene order was assessed by copying the “GeneID” number and pasting it into the “RefSeq” GeneID box and then clicking on “Find Gene”. The location in the genome could be recorded. The following procedure was completed for all new microbes on the completed genome list.

Selecting Organisms for Comparison and obtaining their gene sequences

Representative organisms were selected from each taxonomic order present on the whole microbial list. However, if that order had a representative organism with an abnormal gene size (one that deviated from the rest in the order) then it was selected as well. The representative organisms had to have their sequences of PurL, PurQ, and PurS saved. The gene sequences used for analysis were saved via the “send to” option at the top of the sequence screen after the BLAST searches were completed in part one of the methods above. Click the down arrow and click “Send to File.” Download and save the file into a folder labeled “PurL” on a computer hard drive (save to PurQ or PurS if it falls under that gene). To turn the sequence to a .pro (used for analysis in MegAlign), the sequence was opened in EditSeq then closed and when asked to save the changes “ok,” was selected.

Combining SSLQs

In order to be able to perform a multiple sequence alignment of all the organisms genes had to be ordered into: SSLQ. This was accomplished for each organism in the EditSeq program unless they were monomeric (no alterations had to be performed). The sequences were then copied and pasted into a new window in EditSeq with 2PurS:1PurL:1PurQ. The new sequence was saved on a hard drive under a folder titled “Jess’s SSLQ for analysis.”

Multiple Sequence Alignment and Phylogenetic Analysis

To perform a multiple sequence alignment with all the SSLQ’s for the representative organisms, the program MegAlign from the Lasergene packet was

implemented. Once MegAlign was opened, “Enter Sequences” under the file menu was selected. All the sequences from the “Jess’s SSLQ for analysis” were copied, using the “Add All feature” and then selecting “Done”. Under the Align menu, ClustalW was selected. Once the multiple sequence alignment was completed to observe the phylogenetic tree one selected: “View” → “Phylogenetic Tree”. This tree was printed for reference. This window was closed so the main screen now showed the representative microbes which were subsequently ordered as they were arranged from top to bottom on the phylogenetic tree. The names were edited to remove the .pro at the end on the multiple sequence alignment and the phylogenetic tree for aesthetic purposes. To add a decoration to the multiple sequence alignment under the “Options” menu, “Decorations,” then “Decoration Manager” was selected. “New” was selected along with “shade” for the decoration. The sequences were compared to *E. coli*. To view the multiple sequence alignment, “Alignment Report” was selected under the “View” menu. Further analyses were performed using the BLAST program on NCBI’s website as outlined in part one of the methods section, and using MegAlign.

RESULTS/DISCUSSIONS

From all the completed genomes, one hundred organisms were selected for analysis (see Appendix *Figure 3*). Phylogenetic analysis revealed four different forms of the gene responsible for encoding FGARAT. There was a heterotetrameric form (four subunits), a heterodimeric form (two subunits), and two monomeric forms (one subunit). *Figure 2* was the phylogenetic tree constructed from the multiple sequence alignment computed from all one hundred sequences. There were several organisms which were found to cluster among unrelated organisms of different taxa. For example,

Symbiobacterium thermophilum which is an actinobacteria was seen clustering with *Morella thermoacetica* which is a



Figure 2. Phylogenetic tree comparing the gene sequence encoding FGARAT from 100 representative organisms from numerous taxonomic orders. The organisms in blue are the organisms which have the heterotetrameric form of FGARAT. The organisms in the red are the organisms which have the heterodimeric form of FGARAT. The organisms in the black are the organisms which have the monomeric form of FGARAT. The phylogenetic tree was constructed by the program MegAlign.

firmicute. *S. thermophilum* has a heterodimeric form of the gene protein FGARAT. *M. thermoacetica* on the contrary, has a heterotetrameric protein form. Using a multiple sequence alignment with *S. thermophilum*, and *M. thermoacetica*, one noticed *S. thermophilum* showed more sequence homology with selected firmicutes over selected actinobacteria. This data supports an argument for horizontal gene transfer. Figure 3, shows the multiple sequence alignment with *S. thermophilum*, *M. thermoacetica*, *Thermoanaerobacter tengcongensis* (firmicute), *Acidothermus cellulolyticus* (actinobacterium), and *Streptomyces coelicolor* (actinobacterium). *S. thermophilum* shows a much greater homology to the sequences of the firmicutes rather than its relative actinobacteria.

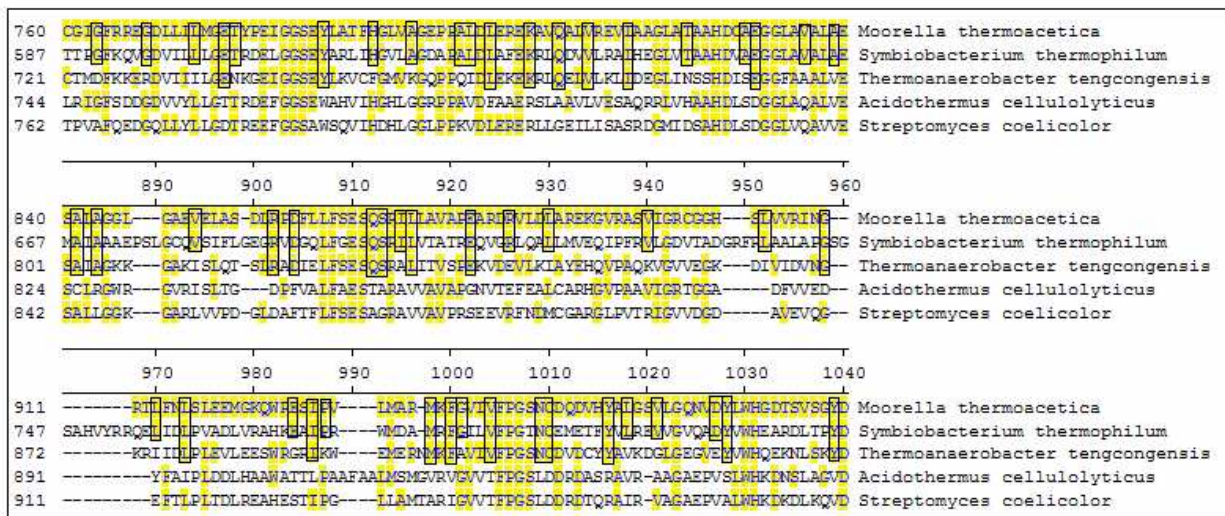


Figure 3. Multiple sequence alignment outlining sequences which *S. thermophilum* share with the firmicutes while none of the actinobacteria selected had or only one had. The black boxes are circling the amino acids. There are many conserved sequences as well.

Not only do the sequences match up closer to the firmicutes but *S. thermophilum* and *M. thermoacetica* grow at the same temperature and possibly inhabit the same niche. *S. thermophilum* is generally found in soils, feces, and animal feeds while *M. thermoacetica* inhabits aquatic environments (NCBI). Therefore, they have the potential

to transfer genes whether it be via direct transfer from the environment (transformation), virus mediated, or a plasmid.

Syntrophomas wolfei was another microbe which stood out on the phylogenetic tree. It had a monomeric gene form however it clustered as an outgroup of archaea which have a heterotetrameric gene form. Research to date has observed a high rate of horizontal gene transfer among archaeal genomes to prokaryotes (Koonin et al 2001). *S. wolfei*'s monomeric form only had a PurL protein sequence of 367 amino acids. This could be a sequencing error resulting from a frame shift mutation so an early stop codon was read during sequencing and thus truncating the resulting gene. An NCBI BLAST search had no significant similarities for any gene sequencing resembling *purS* or *purQ*. However, from the sequencing data available, the amino acid sequenced from *S. wolfei* had a much greater homology to the archaea than the firmicutes. *Figure 4*, was a multiple sequence alignment comparing the amino acid SSLQ's of *S. wolfei*, *Haloarcula marismortui* (archaea), *Haloquadratum walsbyi* (archaea), *Methanococcoides burtonii* (archaea), *Methanosarcina acetivorans* (archaea), *Clostridium acetobutylicum* (firmicute), *Clostridium tetani* (firmicute), and *Clostridium perfringens* (firmicute).

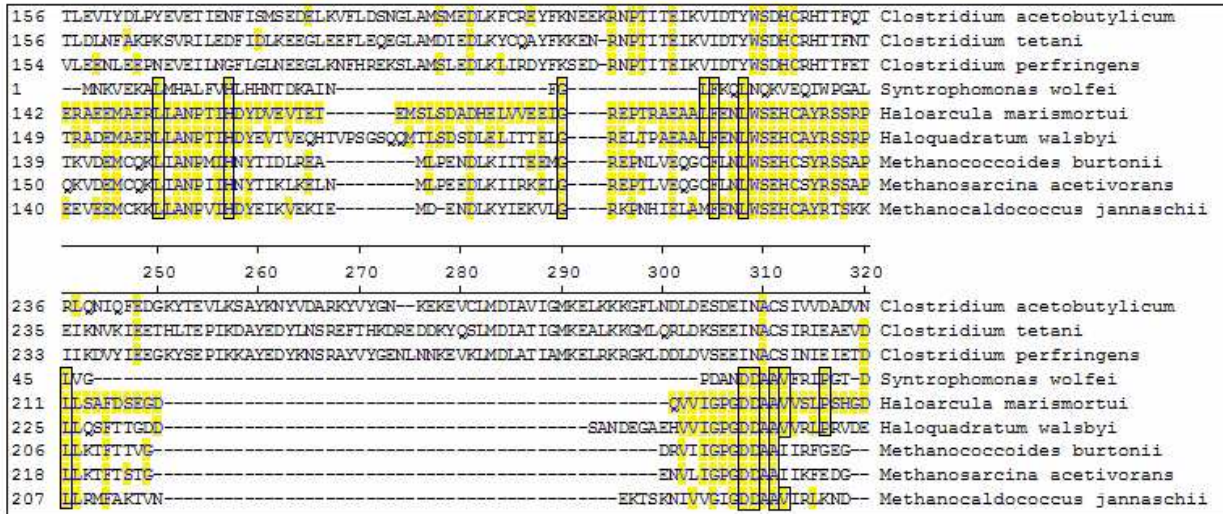


Figure 4. Multiple sequence alignment comparing *S. wolfei*'s FGARAT genomic sequence to that of selected archaea and firmicutes. From what sequence is in the NCBI database, *S. wolfei* shows more sequence homology with the archaea than that of the firmicutes. The black boxes show amino acids shared by the archaea and *S. wolfei* which are not homologous in the selected firmicutes.

The black boxes show the amino acids homologous between the archaea and *S. wolfei* which are not seen in the firmicutes. Furthermore, additional analysis of Figure 4 showed when *S. wolfei* did not have an amino acid homologous to the archaea its amino acid sequence still remained unique from the firmicutes; there were a few sequence exceptions. *S. wolfei* inhabits multiple anaerobic environments where it possibly could encounter *M. burtonii*, *M. acetivorans*, or *M. jannaschii* all of which are aquatic halophiles (NCBI).

The gamma proteobacterium, *Legionella pneumophila* (heterodimeric protein form) clustered with the archaea on the phylogenetic tree. In addition, NCBI reported unusual heterodimeric protein sizes for PurL and PurQ. NCBI reported the PurL protein sequence to contain 780 amino acids which is representative of a typical heterotetrameric gene form. NCBI reported the PurQ protein sequence to contain 419 amino acids which does not correspond to any PurQ protein size in the database.

Figure 5 was the multiple sequence alignment of *L. pneumophila* with *Escherichia coli* (gamma proteobacteria), *Salmonella typhimurium* (gamma proteobacteria), *Baumannia cicadellinicola* (gamma proteobacteria), *Picrophilus torridus* (archaea), *Thermoplasma acidophilum* (archaea), and *Archaeoglobus fulgidus* (archaea). There were no clear distinctions whether *L. pneumophila* had more homologous amino acid sequences with the archaea it clustered with on the phylogenetic tree, or whether it does with the archaea using the whole SSLQ sequence.

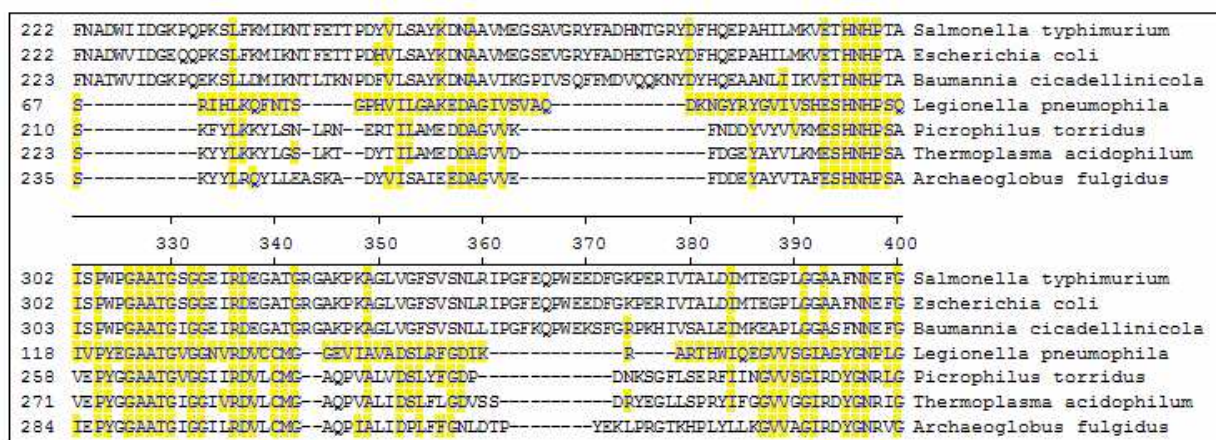


Figure 5. Multiple sequence alignment of the SSLQ comparing *L. pneumophila* to selected gamma proteobacteria and archaea. There is no real distinction as whether *L. pneumophila* should cluster more with the gammas or the archaea.

Another observation from the multiple sequence alignment was *L. pneumophila* did not align with the rest of the SSLQ's until position 172. This would correspond to missing PurS subunits which are approximately 80 amino acids long which would equal 160 amino acids at the beginning of the SSLQ sequence. In addition, *L. pneumophila* ran past the remaining SSLQ's by 130 amino acids at the C-terminus. This showed *L. pneumophila* had a unique gene structure for encoding FGARAT. A multiple sequence alignment of the PurQ sequences revealed *L. pneumophila* had roughly equal homology between the gamma proteobacteria and the archaea. However, it had 150 amino acid

runoff past where the gamma proteobacteria and the gamma sequences ended in the alignment. *Figure 6* showed the PurQ sequence alignment. The FGARAT 3D protein sequence had been researched from *S. typhimurium* and deposited in the Protein Data Bank under the accession number IT3T (Anand et al 2004). *Figure 7* was the picture of the 3D FGARAT protein structure from *S. typhimurium*. A BLAST search of the PurQ sequence yielded no significant results to any FGARAT sequence other than other strains of *L. pneumophila*.

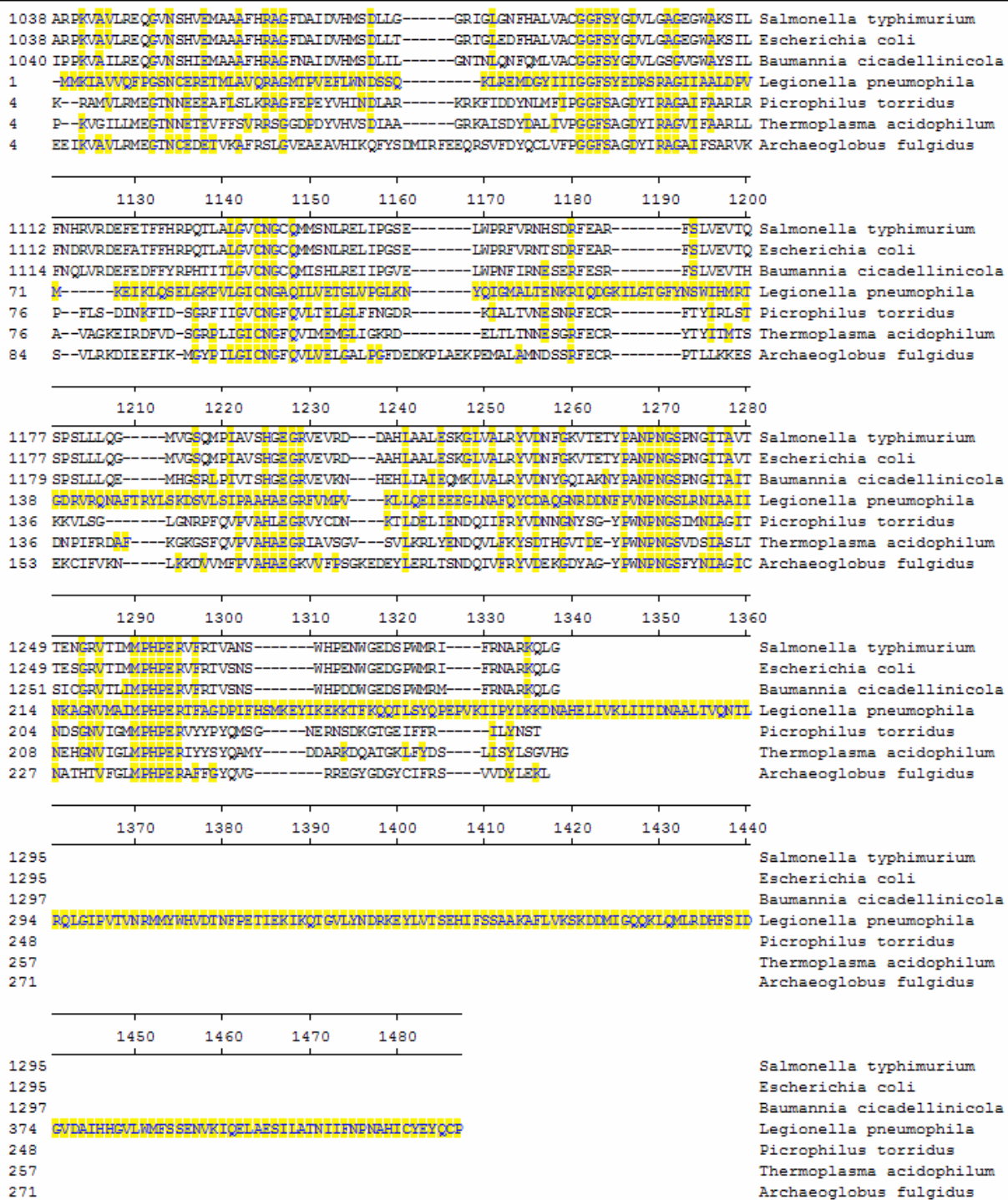


Figure 6. PurQ multiple sequence alignment of *L. pneumophila* with selected gamma proteobacteria and archaea. There is no distinction between whether *L. pneumophila* has more homology with the gammas over the archaea. The 150 amino acid run off is visualized at the end of the alignment.

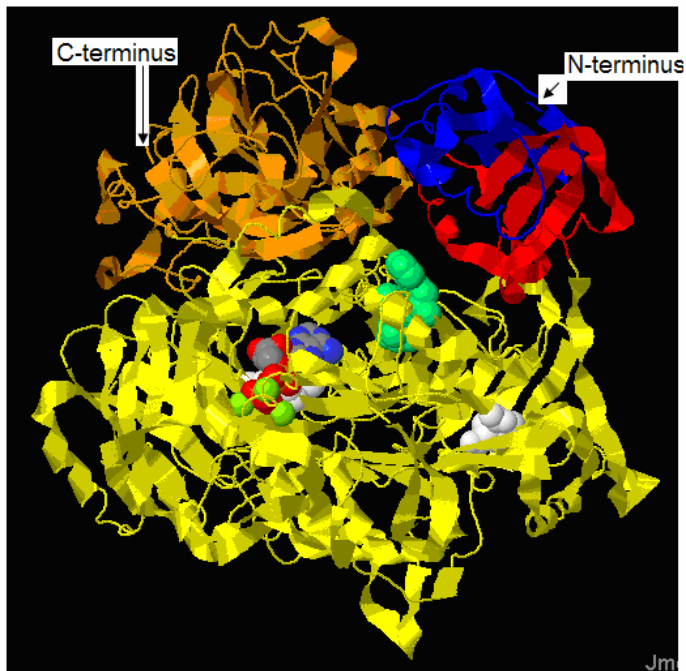


Figure 7. 3D FGARAT structure from *S. typhimurium*. The blue and red subunits correspond to the 2 purS gene segments. The yellow subunit corresponds to the 1 purL gene segments. The orange subunit corresponds to the 1 purQ gene segment. The protein channel leading to the proteins active site runs down along the purS's to the opening in the purL where the white and green molecules are shown.

The multiple sequence alignments revealed *L. pneumophila* did not align with any PurS sequences. However, at its C-terminus end of the PurQ subunit, there was an additional 150 amino acids. These amino acids could quite possibly wrap around the protein where the PurS subunits are located as seen in *Figure 7* in order to protect the ammonium ion channel leading to FGARAT's active site. This would help to explain the unusual

PurQ length as reported in the NCBI database. To determine if the amino acid runoff at the end of PurQ, one could study the structure of FGARAT in *L. pneumophila*.

Figure 2 showed the monomeric forms of FGARAT being separated in the evolutionary history by the heterodimeric gene forms. Noticing the microbes in the upper group of the phylogentic tree (the clostridiums, *Fusobacterium nucleatum*, *Bifidobacterium longum*, *Methanocorpusculum labreanum*, *Streptococcus agalactiae*, *Streptococcus pyogenes*, and *Corynebacterium diphtheriae*) one noticed a general pattern that the above live symbiotically with eukaryotes either as normal flora or as pathogens. One would be lead to believe one possible mode these microbes gained their unique form via horizontal gene transfer from eukaryotes to prokaryotes. However,

the multiple sequence alignment, *Figure 8*, with *Homo sapiens* and all the microbes maintaining the monomeric gene form refuted this hypothesis. *Figure 8* revealed the lower group of organisms with the monomeric form showed more homology with *Homo sapiens* than the pathogenic/symbiotic group. This data helped to support convergent evolution as the means for the presence of two monomeric gene sequence forms. Within the first subunit group, there is a chance for horizontal gene transfer to occur. *Clostridium acetobutylicum*, *Clostridium tetani*, *Clostridium perfringens*, *Streptococcus agalactiae*, and *Streptococcus pyogenes* are all members of the firmicutes. They are all known to cause pathogenic diseases in eukaryotes and thus have the opportunity to transfer genes while being in close proximity to one another in their environment. In addition, *Bifidobacterium longum* and *Fusobacterium nucleatum* are members of the normal intestinal flora of humans (NCBI 2007). Thus providing the perfect opportunity to interact and exchange genes. In order to gain a better understanding of the monomeric protein evolution, one should compare more than one representative organism from each order to increase the sample size. If the sample size were increased, one may be able to study the protein sequences to look for evidence of horizontal gene transfer in the first group of the monomeric protein form.

134	EKIKKYVINPVDGREASLDKPE TLEV TYDLPYVEVTIENFISMSDE	-----LKVFLDSNGIAMSMDLKFCEYFKN	Clostridium acetobutylicum
134	NKIKDYCINDVDSDREGSINKENTLDLNFAPKPSVRILEDFIDLKEEG	-----LEEFLEQEGLAMDIEDLKYQAVFKK	Clostridium tetani
133	NKIKSYIYNPVDGREVSP--LSKVLEENLEE PNEVE IING FLGLNEEG	-----LRNFHREKS LAMSEDLKLRDYFKS	Clostridium perfringens
136	NKIKKFIYNPIEMREKDLVSLKKEELFN--SEVITYDNFIFLDVVD	-----LEKIRIDLGLSMSFEDLKFQGHVKE	Fusobacterium nucleatum
130	DI IKHYVINPVEAREASLETKE TLKT QVFPVKVETIAGFNEMDAEA	-----GQKFTIDERG LAMDLADIE FCQKYTSE	Bifidobacterium longum
130	LA IKKYVINPVESREASLEKPAITLSVSYHI PTTVETLTCFTKLDKDG	-----LAGLIRELGLAMDEDIRFSQQYFIQ	Methanococcus labreanum
133	EAVKNYLNPVDSRFKDI TLP--LEVQAFVSDKTI SNLDFEFTYQADD	-----FAAYKAEQGLAMEVDLLFIQDYFKS	Streptococcus agalactiae
149	EAVKNYLNPVDSRFKDI TLP--LEEQAFVSDKTI PNLDFEFTYQADD	-----FATYKAEQGLAMEVDLLFIQDYFKS	Streptococcus pyogenes
132	ANIRDYLINPVEAGEKNLDV--LRAPSMGD IAPLKQYDFLELDQAG	-----LSALLASEGMMSLADIKLIQCYVRT	Corynebacterium diptheriae
139	AIATLATLHDRMTEQHFT----HPIQSFSPESMPE PINGPINILGE	-----GRILAEKANQELGLALSDWLDLFFYTKRFQ	Homo sapiens
125	QQVTAELHDRMTEVTFALD--DAEQ--LFAHHPPTVTSVDLLGGQ	-----RQALIDANRLGLALAEDEIDYLDQAFI	Escherichia coli
125	RQVAAELHDRMTEVTFSSLT--DAEK--LFIHQPAFVSSVDLLGGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Salmonella typhimurium
126	KKLICLHDRMTEVTFEKLE--HAQK--LFTQYQPE PSKI IDILGAG	-----RTALEKANQIFGLALTEENINYLEFTFT	Baumannia cicadellinicola
131	RQUNTLHDRMTEVTFNDFA--QAST--LFASSPEGELTADIE SGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Colwellia psychrerythraea
125	IEILKAIHDRMTEVTFDFE--SASA--LFAVSEPAFYTEVDLLTGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Vibrio fischeri
150	ATLKDQLHDRMTEVTLNHET--EAL--LFTQKPKALTTIDILNGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Haemophilus influenzae
185	ADVAARLHDRMTEVTFGEMN--EAAA--LFAHHEPRPFTVQVVLGGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Aeromonas hydrophila
131	ERLKFVHDRMTEVTFVGRLE--DAAI--LFAHHPPTVTSVDLLGGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Myxococcus xanthus
125	AAIPAAI HDRMTEVTFEIA--GAEL--LFAHAEPKMTVDLLAGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Hahella chejuensis
138	QAIADSLHDRMTEVTFVGLS--QAAS--LFSHRAQPKPLTAVDILGGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Pseudomonas syringae
131	AALRPLHDRMTEVTFASLT--DAQK--LYHTAEPALPLSTVDLLGGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Methylobacillus flagellatus
127	AAVPLPLHDRMTEVTFVLANLE--AAEA--LFHHYEPKPMISVEVLGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Azoarcus
125	EAVPALHDRMTEVTFVLSRGG--E--EM--LFRQREPE PLQYVPLMGGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Methylococcus capsulatus
127	ERLVSVHDMTEVTFVFAHPD--ETEA--LFCRHEFVPLTTVDLLGGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Nitrosococcus oceani
125	SRIEACI HDRMTEVTFVLSLA--EATL--LFHHSEPGMLNE IDLTLGRG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Nitrosomonas europaea
128	EKILLPLHDRMTEVTFVVDAA--ALPG--LFSFSPGALRQVPTQGGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Magnetococcus
103	QMLLPLHDRMTEVTFVAGLG--DAAA--LFGVHPPRPLETVDLLAGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Thiobacillus denitrificans
141	EQVAALHDRMTEVTFVTDRA--HAEAGLFTSILGAPLQTIIDVLTGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Polaromonas naphthalenivorans
162	AQIGILLHDRMTEVTFVMDRA--SAYG--LFAELPVMMAHVVDLHGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Rhodoferrax ferrireducens
139	AAVAAALHDRMTEVTFVVASRE--DARH--LFDLPAKPLATVDVLAEG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Burkholderia pseudomallei
124	KQIQDIHDMTEVTFVTSCKD--DLYR--LFSVIAPEKLEFVNVLEKG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Francisella tularensis
128	QQWAALHDRMTEVTFVDFQ--TASK--LFHHLESETFSTVDVLLGGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Neisseria meningitidis
155	DISLKCVDYDMTEVTFVLYLEP--PNIMSIFTHEEPKPLVHVPLT PKDTKQSPKDL SKANTELGLALSDGEMEYLIHAFVE	-----RQALIDANRLGLALAEDEIDYLDQAFI	Saccharomyces cerevisiae
125	EKVASELYDPLTESLLFDAE--DLAQ--LFGHPAKTFNDIPVLEKGG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Coxiella burnetii
122	AAVAKLHDMTEVTFVLLGSAA--AAEA--LNVDFPGQLERIP-LD	-----RQALIDANRLGLALAEDEIDYLDQAFI	Xanthomonas campestris
204	KEFAMVHDMTEVTFVYVIT----QKLVSPFETNVVFEVVK--YVFMVK	-----RQALIDANRLGLALAEDEIDYLDQAFI	Arabidopsis thaliana
143	DQFLELHDMTEVTFVYVIT----TPIKSFTDGIIPKAVV--YIPVVEE	-----RQALIDANRLGLALAEDEIDYLDQAFI	Dictyostelium discoideum
107	QEFIRAHDMTEVTFVYVIT----QFITTFETGIKPAEVY--DVDLMSG	-----RQALIDANRLGLALAEDEIDYLDQAFI	Desulfotalea psychrophila
87	---EASHKMLLEVIYQN----LTQDLYNIHQPEFIK-----	-----YIDDI AAFNKQEGIALSQEEIDYLDQAFI	Cytophaga hutchinsonii
88	---FHDFFPMISQKYEN----LDQDIFDIHVNPAFII-----	-----EVDDESFNKQEGIALNAEEVDYLDQAFI	Gramella forsetii
103	---NADHDFMLQRMKGC----LNQNVFTITNROPEAII-----	-----YIDDEAVNKEKQEGIALSKEMDYLDQAFI	Bacteroides fragilis

Figure 8. Multiple Sequence Alignment of the two monomeric forms comparing the amino acid sequences with Homo sapiens amino acid sequences. The lower group shows more homology with Homo sapiens than the pathogenic/symbiotic group does.

CONCLUSIONS

Phylogenetic analysis study of FGARAT genetic sequences supported gene evolution events including horizontal gene transfer and convergent evolution.

Horizontal gene transfer was supported in the multiple sequence alignments evaluating the evolution of *S. thermophilum*'s and *S. wolfei*'s gene sequences. Koonin et al state horizontal gene transfer is most pronounced when, "a particular organism shows the strongest similarity to a homolog from a distant taxon," (2001). Multiple sequence alignment diagrammed *S. thermophilum*'s SSLQ amino acid sequence was most closely

homologous to the SSLQ amino acid sequence observed in the firmicutes it clustered with in *Figure 2* rather than its fellow actinobacteria. The multiple sequence alignment performed with *S. wolfei* supported a strong homology of amino acid sequence to the archaea rather than its taxonomic firmicutes. Convergent evolution as a means of prokaryotic evolution was supported in the evidence compiled for the formation of the two monomeric gene forms. The data also supports the formation of four gene types rather than the three initially hypothesized: one heterotetrameric, one heterodimeric, and two monomeric forms.

For future studies, one would be able to add the new completed microbial genomes to keep the table updated. Since this project was started in the Spring 2007 semester, roughly an additional 20 microbial genomes have been completed. Once the table is updated, one can continue with the phylogenetic analysis.

Possible errors which could have affected the results of the study include errors in sequencing the genomes. If a frame shift were to occur during sequencing the whole gene sequence could be truncated or extended depending when a stop codon was reached. Gene sequences could have been altered since the list was compiled from Sehi's studies if a scientist noticed a sequencing error and went back to fix the mistake. Therefore, some of the sequences which were saved on the Excel spreadsheet may have been altered. This could alter the phylogenetic tree and how the MegAlign program compiled the most parsimonious tree.

APPENDIX

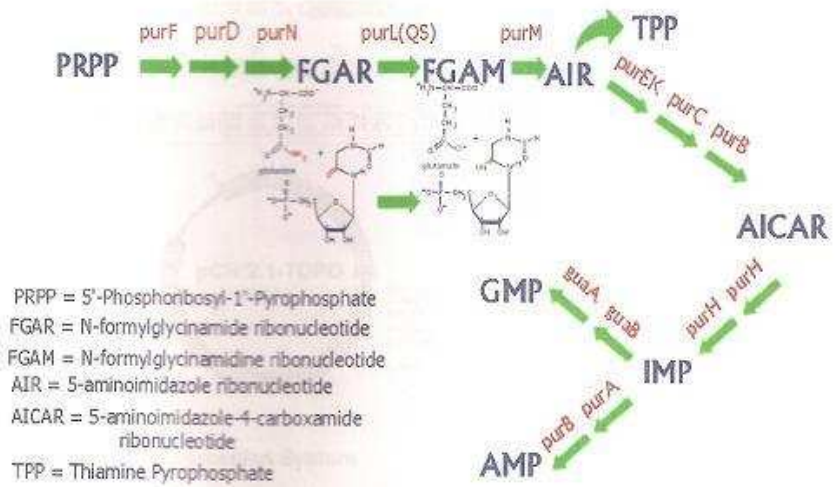


Figure 1. Full steps in the Purine Biosynthetic Pathway. *purL* is outlined in step 4 in the conversion of FGAR to FGAM. In organisms displaying the small *purL* gene, they utilize *purQ* and *purS* in addition to *purL*. (Walsh 2005).

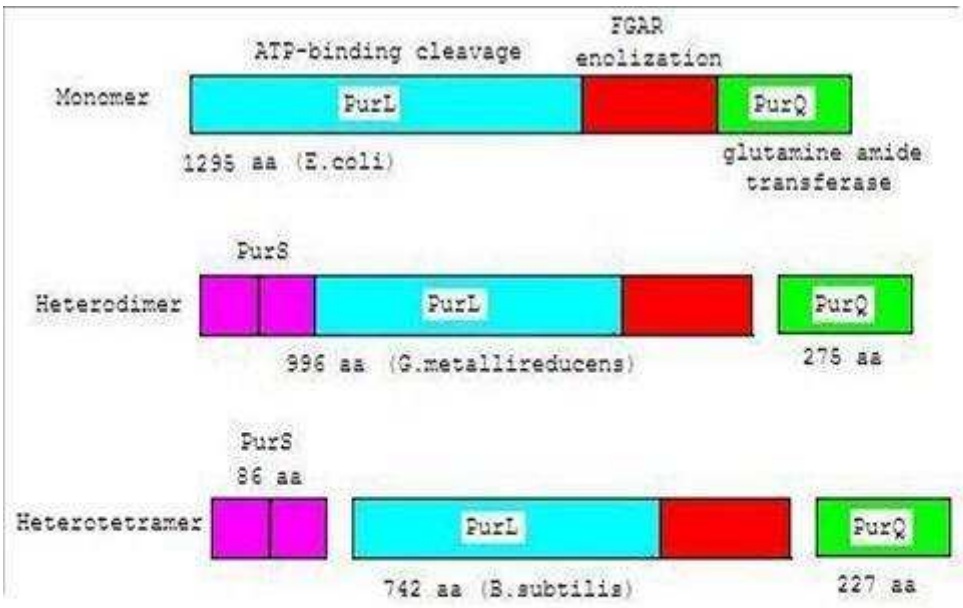
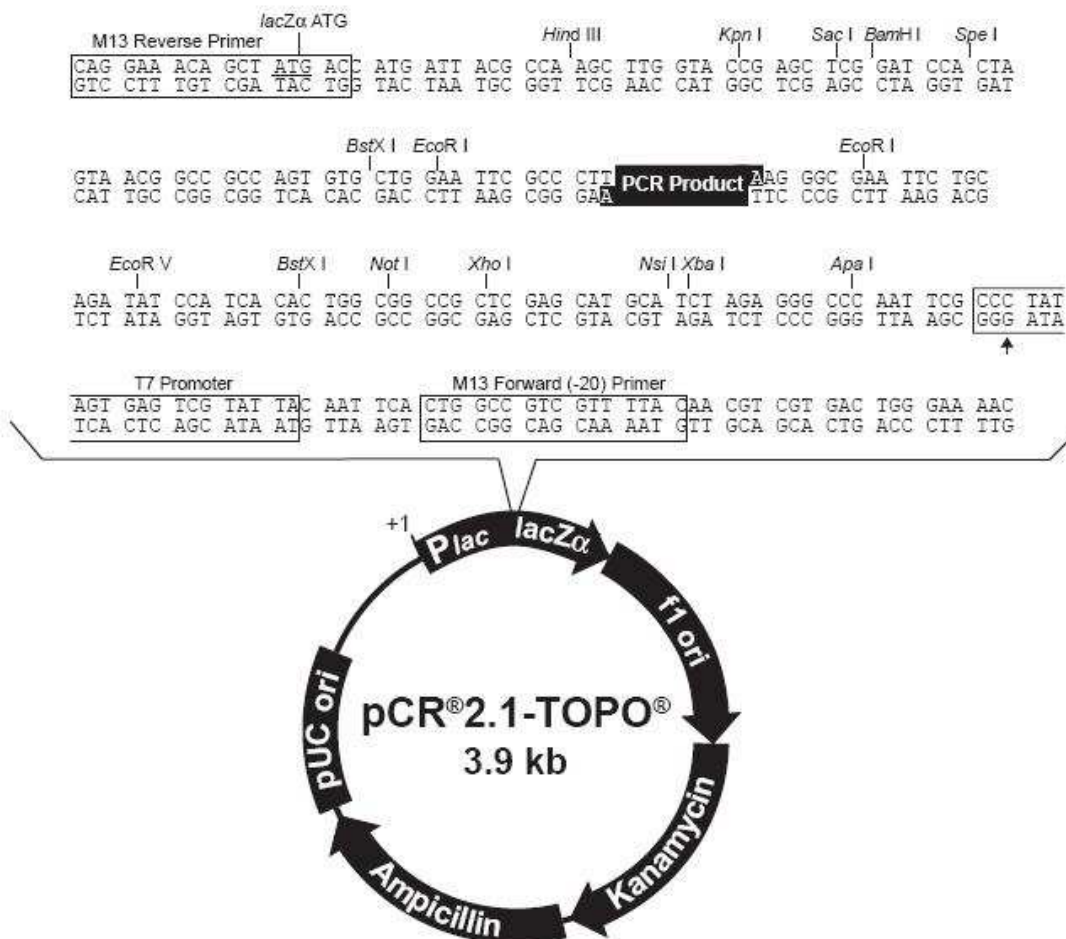


Figure 2. Diagram outlining all 3 forms of *purL*. The top form is large *purL* also designated monomeric *purL*. The bottom diagram depicts small *purL* also designated heterotetrameric *purL*. The middle diagram depicts medium *purL* also designated heterodimeric *purL*.



**Comments for pCR[®]2.1-TOPO[®]
3931 nucleotides**

- LacZα* fragment: bases 1-547
- M13 reverse priming site: bases 205-221
- Multiple cloning site: bases 234-357
- T7 promoter/priming site: bases 364-383
- M13 Forward (-20) priming site: bases 391-406
- f1 ori*: bases 548-985
- Kanamycin resistance ORF: bases 1319-2113
- Ampicillin resistance ORF: bases 2131-2991
- pUC origin: bases 3136-3809

Figure 3. pCR2.1 TOPO cloning vector from Invitrogen. The sequence of the multiple cloning site with the restriction enzyme sites.

Table 1. 100 Representative Sequences used for Phylogenetic Analysis						
Organism	Taxonomy	subunit	PurL	PurQ	Pu rS	ord er
<i>Corynebacterium diphtheriae</i>	Actinobacteria, Actinomycetales	1	1238	-	-	L
<i>Bifidobacterium longum</i>	Actinobacteria, Bifidobacteriales	1	1244	-	-	L
<i>Methanocorpusculum labreanum</i>	Archaea, Euryarchaeota, Methanomicrobiales	1	1231	-	-	L
<i>Bacteroides fragilis</i>	Bacteroidetes, Bacteroidales	1	1249	-	-	L
<i>Gramella forsetii</i> KT0803	Bacteroidetes, Flavobacteriales	1	1225	-	-	L
<i>Cytophaga hutchinsonii</i>	Bacteroidetes, Sphingobacteriales	1	1231	-	-	L
<i>Saccharomyces cerevisiae</i>	Eukaryota, Fungi	1	1358	-	-	L
<i>Homo sapiens</i>	Eukaryota, Metazoa	1	1338	-	-	L
<i>Dictyostelium discoideum</i>	Eukaryota, Mycetozoa	1	1355	-	-	L
<i>Arabidopsis thaliana</i>	Eukaryota, Viridiplantae	1	1387	-	-	L
<i>Clostridium acetobutylicum</i>	Firmicutes, Clostridia	1	1255	-	-	L
<i>Clostridium perfringens</i>	Firmicutes, Clostridia	1	1266	-	-	L
<i>Clostridium tetani</i>	Firmicutes, Clostridia	1	1258	-	-	L
<i>Syntrophomonas wolfei</i>	Firmicutes, Clostridia, Clostridiales	1	367	-	-	L
<i>Streptococcus pyogenes</i>	Firmicutes, Lactobacillales	1	1257	-	-	L
<i>Streptococcus agalactiae</i>	Firmicutes, Lactobacillales	1	1203	-	-	L
<i>Fusobacterium nucleatum</i>	Fusobacteria, Fusobacteriales	1	1249	-	-	L
<i>Rhodoferrax ferrireducens</i>	Proteobacteria, Beta, Burkholderiales	1	1409	-	-	L
<i>Polaromonas naphthalenivorans</i>	Proteobacteria, Beta, Burkholderiales	1	1340	-	-	L
<i>Burkholderia pseudomallei</i>	Proteobacteria, Beta, Burkholderiales	1	1356	-	-	L
<i>Thiobacillus denitrificans</i>	Proteobacteria, Beta, Hydrogenophilales	1	1291	-	-	L
<i>Methylobacillus flagellatus</i>	Proteobacteria, Beta, Methylophilales	1	1297	-	-	L
<i>Neisseria meningitidis</i>	Proteobacteria, Beta, Neisseriales	1	1320	-	-	L
<i>Nitrosomonas europaea</i>	Proteobacteria, Beta, Nitrosomonadales	1	1304	-	-	L
<i>Azoarcus</i> sp. EbN1	Proteobacteria, Beta, Rhodocyclales	1	1310	-	-	L
<i>Desulfotalea psychrophila</i>	Proteobacteria, Delta, Desulfobacteriales	1	1267	-	-	L
<i>Myxococcus xanthus</i>	Proteobacteria, Delta,	1	1302	-	-	L

	Myxococcales					
<i>Aeromonas hydrophila</i>	Proteobacteria, Gamma, Aeromonadales	1	1357	-	-	L
<i>Colwellia psychrerythraea</i>	Proteobacteria, Gamma, Alteromonadales	1	1323	-	-	L
<i>Baumannia cicadellinicola</i>	Proteobacteria, Gamma, Candidatus Baumannia	1	1297	-	-	L
<i>Nitrosococcus oceani</i>	Proteobacteria, Gamma, Chromatiales	1	1300	-	-	L
<i>Salmonella typhimurium</i>	Proteobacteria, Gamma, Enterobacteriales	1	1295	-	-	L
<i>Escherichia coli</i>	Proteobacteria, Gamma, Enterobacteriales	1	1295	-	-	L
<i>Coxiella burnetii</i>	Proteobacteria, Gamma, Legionellales	1	1296	-	-	L
<i>Methylococcus capsulatus</i>	Proteobacteria, Gamma, Methylococcales	1	1288	-	-	L
<i>Hahella chejuensis</i>	Proteobacteria, Gamma, Oceanospirillales	1	1298	-	-	L
<i>Haemophilus influenzae</i>	Proteobacteria, Gamma, Pasteurellales	1	1320	-	-	L
<i>Pseudomonas syringae</i>	Proteobacteria, Gamma, Pseudomonadales	1	1313	-	-	L
<i>Francisella tularensis</i>	Proteobacteria, Gamma, Thiotrichales	1	1290	-	-	L
<i>Vibrio fischeri</i>	Proteobacteria, Gamma, Vibrionales	1	1303	-	-	L
<i>Xanthomonas campestris</i>	Proteobacteria, Gamma, Xanthomonadales	1	1348	-	-	L
<i>Magnetococcus</i> sp. MC-1	Proteobacteria, Magnetococcus	1	1295	-	-	L
Human herpesvirus 4	Viruses, Herpesviridae	1	1318	-	-	L
<i>Methanococcus maripaludis</i>	Archaea, Euryarchaeota, Methanococci	2	989	272	-	QL
<i>Methanospirillum hungatei</i>	Archaea, Euryarchaeota, Methanomicrobiales	2	979	282	-	LQ
<i>Dehalococcoides ethenogenes</i>	Chloroflexi, Dehalococcoidetes	2	953	255	-	LQ
<i>Rhodopirellula baltica</i>	Planctomycetes, Planctomycetales	2	1009	292	-	QL
<i>Anaplasma marginale</i>	Proteobacteria, Alpha, Rickettsiales	2	1016	260	-	QL
<i>Bdellovibrio bacteriovorus</i>	Proteobacteria, Delta, Bdellovibrionales	2	1009	239	-	LQ
<i>Desulfovibrio vulgaris</i>	Proteobacteria, Delta, Desulfovibrionales	2	1009	269	-	LQ
<i>Geobacter sulfurreducens</i>	Proteobacteria, Delta, Desulfuromonadales	2	996	275	-	
<i>Syntrophobacter fumaroxidans</i>	Proteobacteria, Delta, Syntrophobacteriales	2	1009	269	-	LQ

<i>Legionella pneumophila</i>	Proteobacteria, Gamma, Legionellales	2	780	419	-	QL
<i>Treponema denticola</i>	Spirochaetes, Spirochaetales	2	766	270	-	QL
<i>Symbiobacterium thermophilum</i>	Actinobacteria, Symbiobacterium	2	778	235	-	LQ
<i>Acidobacteria bacterium Ellin345</i>	Acidobacteria, Acidobacteriales	4	768	231	80	SQ L
<i>Solibacter usitatus</i>	Acidobacteria, Solibacteres, Solibacterales	4	742	232	81	N/A
<i>Corynebacterium glutamicum</i>	Actinobacteria, Actinomycetales	4	762	223	81	LQ S
<i>Corynebacterium jeikeium</i>	Actinobacteria, Actinomycetales	4	839	223	84	SQ L
<i>Streptomyces coelicolor</i>	Actinobacteria, Actinomycetales	4	752	226	90	SQ L
<i>Acidothermus cellulolyticus</i> 11B	Actinobacteria, Actinomycetales	4	754	225	81	LQ S
<i>Rubrobacter xylanophilus</i>	Actinobacteria, Rubrobacteriales	4	727	220	74	SQ L
<i>Aquifex aeolicus</i>	Aquificae, Aquificales	4	745	227	77	SQ L
<i>Pyrobaculum aerophilum</i>	Archaea, Crenarchaeota, Thermoprotei	4	697	212	84	SQ L
<i>Archaeoglobus fulgidus</i>	Archaea, Euryarchaeota, Archaeoglobi	4	765	271	80	SL Q
<i>Haloarcula marismortui</i>	Archaea, Euryarchaeota, Halobacteria	4	720	228	84	SQ L
<i>Methanothermobacter thermautotrophicus</i>	Archaea, Euryarchaeota, Methanobacteria	4	714	214	84	SQ L
<i>Methanocaldococcus jannaschii</i>	Archaea, Euryarchaeota, Methanococci	4	733	230	83	SQ L
<i>Methanosarcina acetivorans</i>	Archaea, Euryarchaeota, Methanomicrobia	4	715	232	88	SQ L
<i>Methanopyrus kandleri</i>	Archaea, Euryarchaeota, Methanopyri	4	724	226	84	SQ L
<i>Methanococcoides burtonii</i>	Archaea, Euryarchaeota, Methanosarcinales	4	715	231	82	QS L
<i>Pyrococcus abyssi</i>	Archaea, Euryarchaeota, Thermococci	4	705	223	84	SQ L
<i>Picrophilus torridus</i>	Archaea, Euryarchaeota, Thermoplasmata	4	741	248	75	SL Q
<i>Thermoplasma acidophilum</i>	Archaea, Euryarchaeota, Thermoplasmata	4	759	257	79	QL S
<i>Haloquadratum walsbyi</i>	Archaea, Halobacteria,	4	725	228	91	LS

	Halobacteriales					Q
<i>Salinibacter ruber</i>	Bacteroidetes, Sphingobacteriales	4	754	235	92	QS L
<i>Chlorobium tepidum</i>	Chlorobi, Chlorobiales	4	759	234	84	LQ S
<i>Synechococcus elongatus</i>	Cyanobacteria	4	777	221	74	
<i>Gloeobacter violaceus</i>	Cyanobacteria, Gloeobacteria	4	774	232	88	SQ L
<i>Nostoc sp. PCC 7120</i>	Cyanobacteria, Nostocales	4	782	224	92	SQ L
<i>Trichodesmium erythraeum</i>	Cyanobacteria, Oscillatoriales	4	775	231	87	SQ L
<i>Prochlorococcus marinus</i>	Cyanobacteria, Prochlorales	4	803	217	90	SQ L
<i>Deinococcus radiodurans</i>	Deinococcus-Thermus, Deinococcales	4	747	266	84	LQ S
<i>Thermus thermophilus</i>	Deinococcus-Thermus, Thermales	4	725	227	84	SQ L
<i>Bacillus subtilis</i>	Firmicutes, Bacillales	4	742	227	84	SQ L
<i>Carboxydotherrmus hydrogenoformans</i>	Firmicutes, Clostridia	4	728	234	81	SQ L
<i>Thermoanaerobacter tengcongensis</i>	Firmicutes, Clostridia	4	733	224	82	SQ L
<i>Moorella thermoacetica</i>	Firmicutes, Clostridia, Thermoanaerobacteriales	4	733	236	##	LQ S
<i>Enterococcus faecalis</i>	Firmicutes, Lactobacillales	4	739	224	83	LQ S
<i>Caulobacter crescentus</i>	Proteobacteria, Alpha, Caulobacterales	4	739	220	79	SQ L
<i>Sinorhizobium meliloti</i>	Proteobacteria, Alpha, Rhizobiales	4	743	223	80	SQ L
<i>Silicibacter pomeroyi</i>	Proteobacteria, Alpha, Rhodobacterales	4	719	222	76	LS Q
<i>Gluconobacter oxydans</i>	Proteobacteria, Alpha, Rhodospirillales	4	734	233	80	SQ L
<i>Candidatus Pelagibacter ubique</i>	Proteobacteria, Alpha, Rickettsiales	4	730	227	80	SQ L
<i>Zymomonas mobilis</i>	Proteobacteria, Alpha, Sphingomonadales	4	734	221	77	LS Q
<i>Rhizobium etli</i>	Proteobacteria, Alpha, Rhizobiales	4	743	223	80	
<i>Anaeromyxobacter dehalogenans</i>	Proteobacteria, Delta, Myxococcales	4	759	218	81	SQ L
<i>Campylobacter jejuni</i>	Proteobacteria, Epsilon, Campylobacterales	4	728	215	81	SQ L
<i>Leptospira interrogans</i>	Spirochaetes, Spirochaetales	4	745	219	82	SQ L
<i>Thermotoga maritima</i>	Thermotogae, Thermotogales	4	603	213	82	SQ L

REFERENCES

- American Type Culture Association (2006). ATCC medium: 1657 M-14 medium. Online < <http://www.atcc.org/common/documents/mediapdfs/1657.pdf>>.
- Anand, R., Hoskins, A., Bennett, E., et al. (2004). A Model for the *Bacillus subtilis* Formylglycinamide Ribonucleotide Amidotransferase Multiprotein Complex. *Biochemistry* 43:10343-10352.
- Anand, R., Hoskins, A., Stubbe, J., et al. (2004). Domain Organization of *Salmonella typhimurium* Formylglycinamide Ribonucleotide Amidotransferase Revealed by X-ray Crystallography. *Biochemistry* 43:10328-10342.
- Braak, K.V.D. (2002). Haemocytic defence in black tiger prawn (*Penaeus monodon*). PhD thesis – Wageningen Institute of Animal Sciences. The Netherlands. p. 1-168.
- Deysach, B. (2005). On Beauty. *A Journal of the Built & Natural Environments*. 16: Online < <http://www.terrain.org/essays/16/deysach.htm>> [Note: Picture of the African Savannah tree watermarked behind “Part Two” title].
- Ebbole, D.J., Zalkin, H. (1986). Cloning and Characterization of a 12-Gene Cluster from *Bacillus subtilis* Encoding Nine Enzymes for *de novo* Purine Nucleotide Synthesis. *The Journal of Biological Chemistry* 262: 8274-8287.
- Freeman, S., Herron, J.C. (2007). Estimating Evolutionary Trees. *Evolutionary Analysis*. Ed. 4. Pearson Education, Inc. NJ. 111-140.
- Freeman, S., Herron, J.C. (2007). Phylogenomics and the Molecular Basis of Adaptation: 15.2 Lateral Gene Transfer. Ed 4. Pearson Education, Inc. NJ. 584-591.

- Fuerst, J.A., Sambhi, S.K., Paynter, J.L., Hawkins, J.A., Atherton, J.G. (1991). Isolation of a Bacterium Resembling *Pirellula* Species from Primary Tissue Culture of the Giant Tiger Prawn (*Penaeus monodon*). Applied and Environmental Microbiology 57:3127-3134.
- Gogarten, J.P., Doolittle, W.F., Lawrence, J.G. (2002). Prokaryotic Evolution in Light of Gene Transfer. Molecular Biology and Evolution. 19:2226-2238.
- Hoskins, A., Anand, R., Ealick, S., et al. (2004). The Formylglycinamide Ribonucleotide Amidotransferase Complex from *Bacillus subtilis*: Metabolite-Mediated Complex Formation. Biochemistry 42:10314-10327.
- Invitrogen. 2006. TOPO TA Cloning – Five minute cloning of Taq Polymerase – amplified PCR Product. p. 5.
- Invitrogen. 2006. TOPO TA Cloning – Five minute cloning of Taq Polymerase – amplified PCR Product. p. 9-10.
- Invitrogen. pCR2.1-TOPO 3.9kb. Invitrogen life technologies. Online http://www.invitrogen.com/content/sfs/vectors/pcr2_1topo_map.pdf. [Online diagram of the pCR2.1 TOPO cloning vector to include in the appendix from Invitrogen's online website].
- Klug, W.S., Cummings, M.R., Spencer, C.A. (2006). DNA Structure and Analysis. Concepts of Genetics. 8th Edition. 231-262.
- Koonin, E.V., Makarova, K.S., Aravind, L. (2001). Horizontal gene transfer in prokaryotes: quantification and classification. Annual Reviews Microbiology. 55:709-742.

- Lindsay, M.R., Webb, R.I., Fuerst, J.A. (1997). Pirellulosomes: a new type of membrane-bounded cell compartment in planctomycete bacteria of the genus *Pirellula*. *Microbiology* 143:739-748.
- Maegawa, T., Karasawa, T., Ohto, T., et al. (2002). Linkage between toxin production and purine biosynthesis in *Clostridium difficile*. *Journal of Medical Microbiology* 51:34-41.
- NCBI. (2007). The National Center for Biotechnology Information. Online <http://www.ncbi.nlm.nih.gov/>.
- Newman, J. (2004). *Biology 110 Laboratory Manual*.
- Newman, J. (2000). *Biology 110 lecture*.
- Newman, J. (2005). *Genetics 222W Laboratory Manual*.
- Newman, J. (2007). *Microbiology lecture*. [Note: quote at the end of the thesis was taken from March 12, 2007 lecture notes]
- Novagen. (2006). *pET System Manual*. Ed 11: 1-80.
- Qiagen. 2002. *QIAprep Miniprep Handbook*. 22-23.
- Saieb, F., Adley, C. *Biofilm Formation*. *Elements7 Issue*. Online <http://www.ul.ie/elements/Issue7/Biofilm%20Information.htm>. [Note: used the website for the biofilm *Figure* in the introduction].
- Schendel, F.J., Mueller, E., Stubbe, J. (1988). Formylglycinamide Ribonucleotide Synthetase from *Escherichia coli*: Cloning, Sequencing, Overproduction, Isolation, and Characterization. *Biochemistry* 28: 2459-2471.
- Schlesner, H., Rensmann, C., Tindall, B.J., Gade, D., Rabus, R., Pfeiller, S., Hirsch, P. (2004). Taxonomic heterogeneity within the *Planctomycetales* as derived by

DNA-DNA hybridization, description of *Rhodopirellula baltica* gen. nov., sp. nov., transfer of *Pirellula marina* to the genus *Blastopirellula* gen. nov. as *Blastopirellula marina* comb. Nov. and emended description of the genus *Pirellula*. International Journal of Systematic and Evolutionary Microbiology 54: 1567-1580.

Sehi, E., Newman, J. (2006). Unpublished work.

Wagner, M., Horn, M. (2006). The *Planctomycetes*, *Verrucomicrobia*, *Chlamydiae*, and sister phyla comprise a super phylum with biotechnological and medical relevance. Current Opinion in Biotechnology 17:241-249.

Walsh, E. (2005). Lycoming College Honor's Project.

“For the last three and a half billion years, evolution has
been taking notes.”

- Dr. Eric Lander