BIF 701: Identification of Potential Virulence Factors by Utilizing pBLAST to Compare Proteins Sequences from Virulent and Non-virulent strains of *Escherichia coli*

By Christopher Eeles
136079183

1. Introduction:

1.1. BLAST

Basic Local Alignment Tool (BLAST) is a database search program developed by the National Centre for Biotechnology Information (NCBI) to utilize a heuristic model for producing results in the comparison of biological sequence data.[6.1.] By using a heuristic—a relatively quick method for producing approximate solutions to complex problems—the tool can provide a platform for researchers to easily compare and align newly discovered sequences of DNA, RNA or protein with those already known.[6.1.,6.2.] BLAST utilizes the Block Substitution Matrix (BLOSUM), generated from mutation data on highly conserved protein sequences, to calculate a score for alignments between two or more sequence clusters.[6.3.] BLOSUM matrices cluster similar sequences based on percentage identity (C) then compare between clusters; this allows reduction of the bias in current databases for known species and sequences.[6.3.] BLAST searches output the total score for each alignment, the percentage identity between sequences, the percentage positive (i.e., with similar amino acids), the number of gaps, as well as the expected mutation rate in the random model (E).[6.4.] Additional visual data includes a colour coded chart indicating the proportion of sequences within a given score range, along with the sequences of each respective alignment and whether each position is a match, positive, or negative.[6.1.]

1.2. BLAST Search Parameters

By changing the value of C for the BLOSUM matrix used in the search we can screen for various magnitudes of conservation between the compared sequences; gap penalty can also be modulated to the same end[6.3.,6.4.] Scoring choices such as these allow us to tailor the result depending on our needs; *e.g.* use of BLOSUM-50 for sequences expected to be distantly related vs BLOSUM80 when comparing those more closely related.[6.3.] We can also limit our search to strains, species, families, etc., of an organism; vary the class of protein, DNA or RNA; the number of bases or residues in a sequence; the molecular weight; and many other features which may be relevant to our inquiry.[6.4.] More general parameters include maximum number of results, expected threshold (E), word size (seed size) and max query range (limits number of highly aligned results).[6.4.] While not an exhaustive list of parameters, this sample displays the flexibility of BLAST as a tool for exploring relationships between biological sequence data.

1.3. BLASTp

BLASTp is an iteration of the BLAST system designed specifically for protein-protein alignment.[6.1.] Given that protein sequence comparison is the foundation of the BLOSUM matrix, it is logical as a basis for the BLAST system as well. Since sequences of proteins are more highly conserved (20 amino acids vs 4 nucleotides), their comparison may yield insightful results for nucleotides by reverse engineering the protein data into the parental RNA and DNA sequences.[6.3.] In fact, the BLASTp algorithm was modified for use in other modules such as BLASTx (nucleotide to protein) and TBLASTn (protein to nucleotide) by translating the query on the fly.[6.1.] With the accumulation of massive databases on protein and nucleotide sequences, annotated for function, species, etc., discoveries in any sub-field yield interesting and relevant results for others.[6.3.] Innovations such as these have contributed to the exponential growth of biological databases and hold a wealth of valuable information to be utilized in basic and applied science with the proper analysis and interpretation.

2. Purpose:

The purpose of this lab is to introduce the application of BLASTp sequence alignment to infer the function of proteins and the genes which produced them.[6.5.] Using these tools, we will investigate potential virulence factor genes in *E. coli* strain 0157:H7 and align them with BLASTp databases to identify homologous proteins in other species.[6.5.] These homologs will be used to infer the function, class and family of the protein in question. Subsequently the protein from *E. coli* 0157:H7 EDL99 will be compared to sequences from *E. coli* K-12 MG1655—a known non-pathogenic strain—to identify potential differences which could account for their relative pathogenicities.[6.5.] While lack of a homolog in *E.col* K-12 would be strong evidence for the gene of interest producing a virulence factor, it is possible that sequence variation could alter the expression of the gene or the efficiency of the protein produced.[6.5.] Additionally, even a direct homolog could be an indirect virulence factor by conferring a function which supports the production or distribution of other virulence factors (e.g., endotoxin) or even supports survival of the bacterium in its environmet.[6.6.]

The gene of interest in this analysis will be ybgP (NCBI ID: 12513628), a candidate virulence factor in *E. coli* 0156:H7 EDL99.[6.5.]

3. Methods:

3.1. BLASTp was accessed from the NCBI website at https://blast.ncbi.nlm.nih.gov/Blast.cgi.

3.2. The initial search was conducted in Non-redundant Protein Sequences (NRPS) Database utilizing standard parameters (BLOSSOM62, etc.) to identify the superfamily to which the gene and produced protein belong. This information will be used to identify the function of the protein/gene as well as to investigate whether the gene is common to other pathogenic strains or species.

3.3. Secondary search utilized BLOSSUM80 and RefSeq Database while limiting comparisons to *E. coli* K-12 MG1655. Increasing the C value of the BLOSSUM matrix increases the required minimum percent identity for clustering and thus allows more specific comparisons between the gene of interest and the database results (*i.e.*, smaller clusters therefore more clusters to compare to). Selection of RefSeq over NRPS to increase the quality of the data for comparison: it is a smaller set but contains only data from curated (higher quality) databases.[6.7.]

3.4. Results will be analyzed and discussed in next section.

4. Discussion:

The search from 3.2. yielded many hits with high scores and percent identities (Fig. 1):

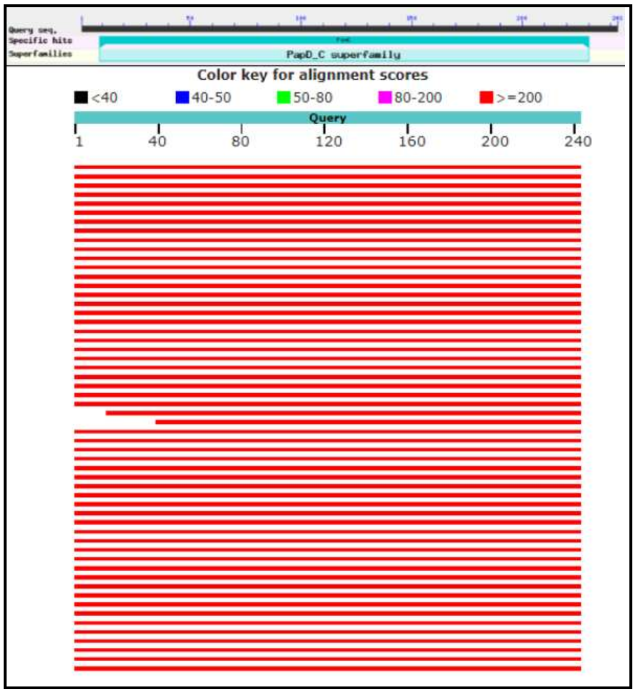
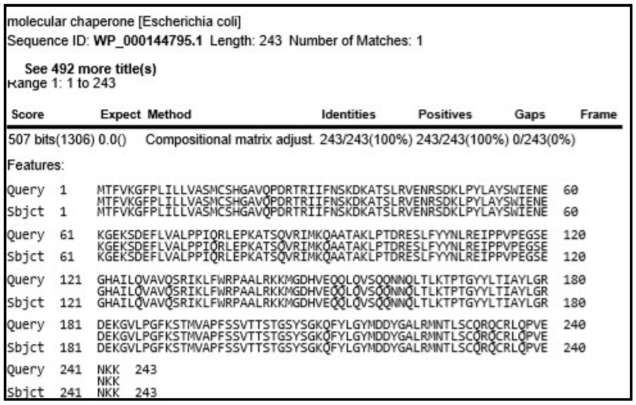


Figure 1

The expressed protein appears to be a member of the PapD_C superfamily, specifically the FimC variety. These proteins are part of the chaperone/usher pathway used by Gram-negative bacteria to assemble adhesive pili or fimbriae.[6.7.] Specifically, FimC is a molecular chaperone that assists in periplasmic protein folding in assembly of type 1 and P pili. Type 1 and P pili play a key role in adhesion host epithelia that experience intermittent fluid flow, such as the urinary and gastrointestinal tracts, which is necessary for pathogen growth in such environments. [6.7.] Within *E. coli* 0157:H7 FimC also plays a role in biogenesis of long, polar fimbriae which appear to be involved in formation of bacterial microcolonies.[6.8.]



```
molecular chaperone [Escherichia coli]
Sequence ID: WP_000144795.1  Length: 243  Number of Matches: 1

   See 492 more title(s)
Range 1: 1 to 243

Score             Expect  Method                        Identities      Positives      Gaps        Frame
507 bits(1306) 0.0()    Compositional matrix adjust. 243/243(100%) 243/243(100%) 0/243(0%)
Features:
Query  1    MTFVKGFPLILLVASMCSHGAVQPDRTRIIFNSKDKATSLRVENRSDKLPYLAYSWIENE  60
            MTFVKGFPLILLVASMCSHGAVQPDRTRIIFNSKDKATSLRVENRSDKLPYLAYSWIENE
Sbjct  1    MTFVKGFPLILLVASMCSHGAVQPDRTRIIFNSKDKATSLRVENRSDKLPYLAYSWIENE  60

Query  61   KGEKSDEFLVALPPIQRLEPKATSQVRIMKQAATAKLPTDRESLFYYNLREIPPVPEGSE  120
            KGEKSDEFLVALPPIQRLEPKATSQVRIMKQAATAKLPTDRESLFYYNLREIPPVPEGSE
Sbjct  61   KGEKSDEFLVALPPIQRLEPKATSQVRIMKQAATAKLPTDRESLFYYNLREIPPVPEGSE  120

Query  121  GHAILQVAVQSRIKLFWRPAALRKKMGDHVEQQLQVSQQNNQLTLKTPTGYYLTIAYLGR  180
            GHAILQVAVQSRIKLFWRPAALRKKMGDHVEQQLQVSQQNNQLTLKTPTGYYLTIAYLGR
Sbjct  121  GHAILQVAVQSRIKLFWRPAALRKKMGDHVEQQLQVSQQNNQLTLKTPTGYYLTIAYLGR  180

Query  181  DEKGVLPGFKSTMVAPFSSVTTSTGSYSGKQFYLGYMDDYGALRMNTLSCQRQCRLQPVE  240
            DEKGVLPGFKSTMVAPFSSVTTSTGSYSGKQFYLGYMDDYGALRMNTLSCQRQCRLQPVE
Sbjct  181  DEKGVLPGFKSTMVAPFSSVTTSTGSYSGKQFYLGYMDDYGALRMNTLSCQRQCRLQPVE  240

Query  241  NKK   243
            NKK
Sbjct  241  NKK   243
```

Figure 2

A perfect match was found for the sequence of interest in the reference sequence of a molecular chaperone in *E. coli* (Fig. 2). This provides evidence for the role of the ybgP gene in production of chaperonins across the entire species *Escherichia coli*; this may indicate that this protein is not a virulence factor in and of itself, but instead plays a general role in adhesion of *E. coli* bacterium to their respective environments. A multispecies match was also found (99% identity) in the *Enterobacteriaceae* family of Gram-negative bacteria; this family contains known pathogens such as *Salmonella, Shigella* and *Enterbacter* along with many harmless species.[6.9.] This further substantiates the hypothesis of a general role for ybgP in bacterial adhesion to their respective environments.

The search from 3.3. yielded significantly fewer hits with lower scores and percent identities (Fig. 3):

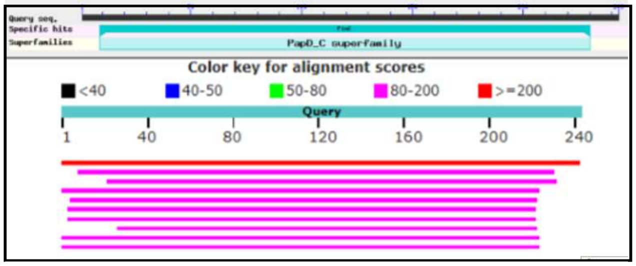
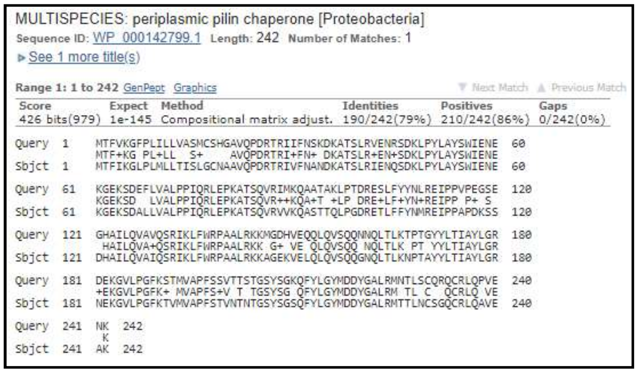


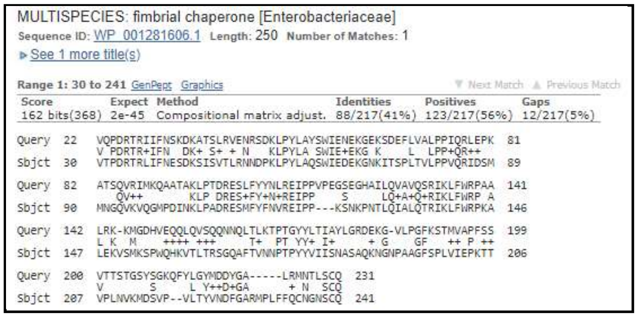
Figure 2

A high percentage identity (79%) and percent positive (86%) was found for a periplasmic pilin chaperone in the phylum *Proteobacteria*, which contains the *Enterobacteriaceae* family identified in the previous search (Fig. 4). This extends the role of ybgP and similar genes to a much wider range of species and supports the hypothesis of a general role in bacterial adhesion. Interestingly, when limiting the search to *E. coli* K-12 MG1655, the multispecies match for *Enterobacteriaceae* dropped significantly to 41% identity and 56% positive (Fig. 5). While this match is for a fimbrial chaperone (vs a pilin chaperone for the previous search) it has been established in earlier analysis that FimC likely plays a role in both these processes. A potential explanation for such a change would be that in *E. coli* K-12, the ybgP gene is modified to have reduced expression levels or altered protein efficiency in non-pathogenic strains. This change may be adaptive as the environment inside the human body is significantly more hostile than most external environments; this would likely increase selection pressures for both adhesion (via pili) and colony formation (via fibria). An interesting question for further inquiry, assuming this explanation is correct, may be to study the phylogeny of *E. coli* to determine the direction of the adaptation (*i.e.*, from pathogens to non-pathogens or vice versa).



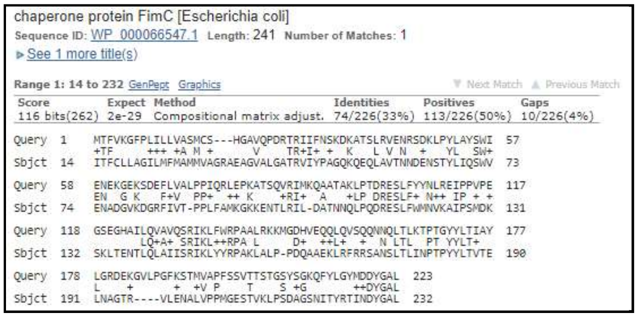
```
MULTISPECIES: periplasmic pilin chaperone [Proteobacteria]
Sequence ID: WP_000142799.1  Length: 242  Number of Matches: 1
▸ See 1 more title(s)

Range 1: 1 to 242  GenPept  Graphics                          ▼ Next Match  ▲ Previous Match
Score           Expect  Method                         Identities     Positives      Gaps
426 bits(979)   1e-145  Compositional matrix adjust.   190/242(79%)   210/242(86%)   0/242(0%)

Query  1    MTFVKGFPLILLVASMCSHGAVQPDRTRIIFNSKDKATSLRVENRSDKLPYLAYSWIENE  60
            MTF+KG PL+LL  S+    AVQPDRTRI+FN+ DKATSLR+EN+SDKLPYLAYSWIENE
Sbjct  1    MTFIKGLPLMLLTISLGCNAAVQPDRTRIVFNANDKATSLRIENQSDKLPYLAYSWIENE  60

Query  61   KGEKSDEFLVALPPIQRLEPKATSQVRIMKQAATAKLPTDRESLFYYNLREIPPVPEGSE  120
            KGEKSD  LVALPPIQRLEPKATSQVR++KQA+T +LP DRE+LF+YN+REIPP P+ S
Sbjct  61   KGEKSDALLVALPPIQRLEPKATSQVRVVKQASTTQLPGDRETLFFYNMREIPPAPDKSS  120

Query  121  GHAILQVAVQSRIKLFWRPAALRKKMGDHVEQQLQVSQQNNQLTLKTPTGYYLTIAYLGR  180
             HAILQVA+QSRIKLFWRPAALRKK G+ VE QLQVSQQ NQLTLK PT YYLTIAYLGR
Sbjct  121  DHAILQVAIQSRIKLFWRPAALRKKAGEKVELQLQVSQQGNQLTLKNPTAYYLTIAYLGR  180

Query  181  DEKGVLPGFKSTMVAPFSSVTTSTGSYSGKQFYLGYMDDYGALRMNTLSCQRQCRLQPVE  240
            +EKGVLPGFK+ MVAPFS+V T TGSYSG QFYLGYMDDYGALRM TL C  QCRLQ VE
Sbjct  181  NEKGVLPGFKTVMVAPFSTVNTNTGSYSGSQFYLGYMDDYGALRMTTLNCSGQCRLQAVE  240

Query  241  NK  242
             K
Sbjct  241  AK  242
```

Figure 4



```
MULTISPECIES: fimbrial chaperone [Enterobacteriaceae]
Sequence ID: WP_001281606.1  Length: 250  Number of Matches: 1
▸ See 1 more title(s)

Range 1: 30 to 241  GenPept  Graphics                         ▼ Next Match  ▲ Previous Match
Score           Expect  Method                         Identities     Positives      Gaps
162 bits(368)   2e-45   Compositional matrix adjust.   88/217(41%)    123/217(56%)   12/217(5%)

Query  22   VQPDRTRIIFNSKDKATSLRVENRSDKLPYLAYSWIENEKGEKSDEFLVALPPIQRLEPK  81
            V PDRTR+IFN  DK+ S+ + N   KLPYLA SWIE+EKG K    L  LPP+QR++
Sbjct  30   VTPDRTRLIFNESDKSISVTLRNNDPKLPYLAQSWIEDEKGNKITSPLTVLPPVQRIDSM  89

Query  82   ATSQVRIMKQAATAKLPTDRESLFYYNLREIPPVPEGSEGHAILQVAVQSRIKLFWRPAA  141
               QV++       KLP DRES+FY+N+REIPP    S     LQ+A+Q+RIKLFWRP A
Sbjct  90   MNGQVKVQGMPDINKLPADRESMFYFNVREIPP---KSNKPNTLQIALQTRIKLFWRPKA  146

Query  142  LRK-KMGDHVEQQLQVSQQNNQLTLKTPTGYYLTIAYLGRDEKG-VLPGFKSTMVAPFSS  199
            L K  M    ++++ +++     T+  PT YY+ I+      + G    GF   ++ P ++
Sbjct  147  LEKVSMKSPWQHKVTLTRSGQAFTVNNPTPYYVIISNASAQKNGNPAAGFSPLVIEPKTT  206

Query  200  VTTSTGSYSGKQFYLGYMDDYGA-----LRMNTLSCQ  231
            V       S     L Y++D+GA      + N  SCQ
Sbjct  207  VPLNVKMDSVP--VLTYVNDFGARMPLFFQCNGNSCQ  241
```

Figure 5



```
chaperone protein FimC [Escherichia coli]
Sequence ID: WP_000066547.1  Length: 241  Number of Matches: 1
▸ See 1 more title(s)

Range 1: 14 to 232  GenPept  Graphics                         ▼ Next Match  ▲ Previous Match
Score           Expect  Method                         Identities     Positives      Gaps
116 bits(262)   2e-29   Compositional matrix adjust.   74/226(33%)    113/226(50%)   10/226(4%)

Query  1    MTFVKGFPLILLVASMCS---HGAVQPDRTRIIFNSKDKATSLRVENRSDKLPYLAYSWI  57
            +TF     +++ +A M +      V    TR+I+ +  K   L V N  +   YL  SW+
Sbjct  14   ITFCLLAGILMFMAMMVAGRAEAGVALGATRVIYPAGQKQEQLAVTNNDENSTYLIQSWV  73

Query  58   ENEKGEKSDEFLVALPPIQRLEPKATSQVRIMKQAATAKLPTDRESLFYYNLREIPPVPE  117
            EN  G K   F+V  PP+  ++ K    +RI+  A   +LP DRESLF+ N++ IP + +
Sbjct  74   ENADGVKDGRFIVT-PPLFAMKGKKENTLRIL-DATNNQLPQDRESLFWMNVKAIPSMDK  131

Query  118  GSEGHAILQVAVQSRIKLFWRPAALRKKMGDHVEQQLQVSQQNNQLTLKTPTGYYLTIAY  177
                   LQ+A+ SRIKL++RPA L     D+  ++L+  +  N LTL  PT YYLT+
Sbjct  132  SKLTENTLQLAIISRIKLYYRPAKLALP-PDQAAEKLRFRRSANSLTLINPTPYYLTVTE  190

Query  178  LGRDEKGVLPGFKSTMVAPFSSVTTSTGSYSGKQFYLGYMDDYGAL  223
            L    +      +  +V P    T    S +G       ++DYGAL
Sbjct  191  LNAGTR----VLENALVPPMGESTVKLPSDAGSNITYRTINDYGAL  232
```

Figure 6

The alignment for the FimC chaperone protein yielded even weaker results at 33% identity and 50% positive (Fig. 6). This supports the general role hypothesis and may substantiate prior speculation about environmental adaptation. For example, if changes in 50% of the protein affect its level of activity or modify binding affinity for specific targets it may alter number and/or structure of pili and fimbriae produced by the cell. This could result in specialization of the cells structure to optimize adhesion and microcolony formation for their specific environmental niche. While these hypotheses doubtlessly require more bioinformatic analysis and eventually wet-lab testing, they demonstrate the kind of information which even a superficial examination of sequence alignment data can produce.

5. Conclusion:

The complexity of biological systems necessitates the use of computer assisted analyses such as BLAST to glean useful information about the structure and function of elements within them. In analyzing the ybgP gene from *E. coli* 0157:H7 EDL99 and its associated protein product, we were able to generate valuable clues about its role in the virulence and pathogenicity of *E. coli* strains. We discovered that the protein expressed is likely FimC, which functions as a periplasmic chaperonin in the chaperone/usher pathway gram-negative bacteria, *E. coli* included, use for biosynthesis of essential organelles such as pili and fimbria. We identified homologous proteins both within the *Escherichia coli* species and larger grouping such as the *Enterobacteriaceae* family and *Proteobacteria* phylum. While FimC does not play as direct a role in virulence as say endotoxin production, it nonetheless facilitates a necessary condition for survival and flourishing of bacterial pathogens. Homology within *E. coli* and between species of pathogenic and non-pathogenic bacterium indicates that the production of pili and fimbriae is likely common to many families and phyla, and that similar proteins are necessary for their growth and survival. Variation between the sequences of *E. coli* K-12 MG1655 and 0157:H7 EDL99 may explain difference in pathogenicity as well as reflect adaptation environment specific selection pressures. While our analysis is not conclusive, it does suggest a role for ybgP in bacterial virulence and therefore it could rightly be called a virulence factor. Further investigation of the chaperone/usher pathway in bacterial pathogens may yield valuable insights about potential drug targets for disrupting pathogen growth, survival and virulence; therefore, further investigation is warranted within bioinformatics as well as molecular biology/genetics and, eventually, medical science more generally.

6. References:

6.1. Madden, T. (2002). The BLAST Sequence Analysis Tool. *The NCBI Handbook* [Internet], (Md), 1–15. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK153387/.

6.2. Wikipedia community. (2018, August 14). Heuristics (computer science). In *Wikipedia*. Retrieved September 15, 2018 from https://en.wikipedia.org/wiki/Heuristic_(computer_science).

6.3. Zvelebil, M., Baum, J.O. (2008). Producing and Analyzing Sequence Alignments. In *Understanding Bioinformatics* (pp. 71 – 113). Garland Science: New York, NY.

6.4. NCBI. (N.d.). BLAST Search Parameters. In *BLAST topics* [Internet]. Retrieved on September 15, 2018 from https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp#Matrix.

6.5. School of Biological Sciences and Applied Chemistry. (2018). BIF701 Lab 1 Sequence Alignment: Using Protein Alignment to Investigate Function of Virulence Factors. Toronto, ON: Seneca College.

6.6. NCBI. (N.d.). Refseq non-redundant proteins. In *RefSeq* [Internet]. Retrieved on September 16, 2018 from https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/.

6.7. Sarowar, S. *et al*. (2016). The *Escherichia coli* P and Type 1 Pilus of Assembly Chaperones PapD and FimC Are Monomeric in Solution. *Journal of Bacteriology, 198*(17), pp. 2360 – 2369. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4984555/pdf/zjb2360.pdf.

6.8. Haque, S. (2010). Role of selected fimbrial adhesins in pathogenesis of acid-induced enterohemorrhagic *Escherichia coli* 0157:H7. *Theses and dissertations*, paper 1384. Ryerson University: Toronto, ON.

6.9. Wikipedia community. (2018, August 24). Enterobacteriaceae. In *Wikipedia.* Retrieved September 16, 2018 from https://en.wikipedia.org/wiki/Enterobacteriaceae.

6.10. Wikipedia community. (2018, August 3). Proteobacteria. In *Wikipedia.* Retrieved September 16, 2018 from https://en.wikipedia.org/wiki/Proteobacteria.