CSC 314, Bioinformatics Lab #10: **Name:________________________**
BLAST and Conserved Domain Identification

Escherichia coli (*E. coli*) is a rod-shaped bacterium that is prevalent in the human gut. Although most strains are harmless, some strains cause food poisoning in their hosts. One such strain is EDL933, which was isolated from contaminated ground beef from a McDonald's in Michigan following an outbreak in 1982. An important biological question is why does the EDL933 strain cause disease, while harmless strains such as MG1655 do not. *Bioinformatics* can be used to help answer this question, by identifying *virulent* proteins that are present in the disease-causing strain but not in the harmless strains.

Through gene sequencing and a bioinformatics analysis, approximately 1000 genes have been identified that are present in the pathogenic EDL933 strain but that are not present in the harmless MG1655 strain.

In this lab, you will use the protein Basic Local Alignment Search Tool (BLAST; https://blast.ncbi.nlm.nih.gov/Blast.cgi) along with the Conserved Domain Identification search to identify potential virulence factors in the EDL933 strain.

Potential virulence factors can be identified by considering proteins that (a) are similar to proteins (have homologs) in strains of bacteria known to cause disease, and/or (b) have conserved domains with functions known to be associated with disease. Functions associated with disease include Type III or Type IV secretion systems, which deliver toxins; enzymes that break down host proteins; and features that allow attachment to host cells, among others. Pathogenic strains include *Salmonella enterica*, a bacterium that causes Salmonellosis, a disease associated with vomiting, diarrhea, fever, and abdominal cramps; and the *Shigella* bacteria, which can cause dysentery. You may assume that other strains are not pathogenic.

Assignment

Three protein sequences are provided in the file sequences.txt.

1. Answer the questions below for each sequence:

   A. Perform a protein-protein BLAST search against the *reference proteins* database, limiting your search to *enterobacterales (taxid:91347)* and excluding all *Escherichia (taxid:561)* bacteria. Also exclude non-redundant RefSeq proteins. For the top hit, identify the protein name, the species, and accession number. (Note: the top hit may be a MULTISPECIES protein, which is a consensus across multiple species; if this is the top hit, then answer this question based on the second hit).

B. For the top hit, how many matches (regions of the protein) were identified? For the top match, specify its E-value, its percent identity, and percent similarity scores.

C. Based on your answer to (B) above, do you think the matches may be due to chance? Why or why not?

D. Based on the species of the top hit, do you believe that the protein may be a virulence factor in EDL933? Why or why not?

E. Look at the conserved domains, and answer the following for the first *pfam* match (You should only look at the first *pfam* match for this question). What is the *pfam* accession # and domain name? Does this domain indicate that the protein may be a virulence factor in EDL933? Why or why not?

Part II. Based on the above analysis for each protein, which protein is most likely to be a virulence factor, and why?

Note: The top matches for each of the three sequences are listed below. If your top matches are different, then your results are incorrect, and you should see me if you are unable to get the correct ones.

1. Unknown protein #1: NP_709290.2
2. Unknown protein #2: WP_094318076.1
3. Unknown protein #3: WP_014657490.1