BIO/CMPSC 300 Introduction to Bioinformatics
Spring 2016
Kristen Webb and Janyl Jumadinova
http://cs.allegheny.edu/sites/jjumadinova/300

Lab Week 7 – Investigating Potential Virulence Factors in *E. coli*
Given: Friday March 11, 2016
Due: Friday April 1 by 2:30pm

Objectives:

- **Understand the use of a substitution matrix to score amino acid similarity in a protein sequence alignment.**
- **Gain experience using protein alignment to develop hypotheses about protein function based on sequence similarity.**
- **Know how protein alignment differs algorithmically from DNA alignment.**
- **Know how substitution matrix is developed and how different matrices might be used to produce better alignments in particular situations.**

Reading Assignment:

Chapter 5 in Exploring Bioinformatics textbook.

Required Deliverables (submitted through your Bitbucket repository):

Please ensure that your bitbucket repository contains a directory named 'lab7' and that your report and your program (s) must contain the names of all team members and the honor code pledge. Additionally, if you create a new team repository, you have to name it '300s2016-lastNames' and share it with both instructors.

1. **Report containing two one page summaries, one for each of two candidate virulence genes discussing their likely functions based on conserved domains and orthologous proteins.**
2. **Your report should also contain analysis of your BioPython results.**
3. **Python program implementing part III.**
4. **The outputs after running BLAST using the Python program. You should include the outputs as appropriately labeled separate files.**

General Guidelines for Labs

Work on the Alden Hall computers. If you want to work on a different machine, be sure to transfer your programs to the Alden machines and re-run them before submitting.
Keep all of your files! Don't delete your programs and reports after you hand them in---you might need them again later.
Back up your files regularly. Use a flash drive or Google Drive or whatever your favorite backup method is.
Review the Honor Code policy on the syllabus. Remember that you may discuss experiments and programs with others, but copying answers or programs is a violation of the Honor Code

Part I Deliverable: Choose two candidate virulence factors from table on the next page. For each factor you've selected, write a one-page summary discussing the factor's likely function based on conserved domains and orthologous proteins. Each summary should include a table with the organisms you've identified as having similar proteins, the known functions of those proteins, the quality of your BLAST matches, and the type of BLAST and substitution matrix used to generate the alignment. Based on the evidence accumulated, is it reasonable to identify your protein as a virulence factor? How would the function you have hypothesized for the protein contribute to the ability of EDL933 to cause disease? Comment on the strength of your evidence: How confident are you in assigning this function to your protein or in characterizing its role in virulence? Be sure to include other tools and resources used in your characterization.

Part I Background: *Escherichia coli* (*E. coli*) is a very well known species of bacteria due to it being a major contributor to cases of food poisoning. However, most strains of *E. coli* are harmless, or even beneficial, and reside in the large intestines of humans and other mammals. One strain in particular, named O157:H7 is a highly virulent pathogen known to cause serious or even potentially fatal disease if as few as 10 cells are ingested. What makes strain O157:H7 so different?

One key factor is O157:H7's acquisition of the gene for a toxin called Shiga toxin (Stx) not present in other *E. coli* strains. Stx binds receptors found in human kidney tissue but, importantly, is not found in cattle, enabling these animals to be symptom-free carriers of the bacteria. Genome sequencing has revealed many other differences between the O157:H7 genome and the genomes of "tame" *E. coli* inhabiting the human gut. At least some of these genes specific to O157:H7 are likely to encode **virulence factors**: proteins such as Stx that contribute to the ability of the organism to cause disease.

Identifying and studying novel virulence genes evolved in or acquired by highly pathogenic strains of *E. coli* such as O157:H7 could be crucial for dealing with this important foodborne disease. Understanding how these bacteria cause disease and why they have more severe effects than typical *E. coli* strains may lead us to new and better ways to treat and prevent disease.

One of the first completely sequenced genomes was that of *E. coli* strain K-12 substrain MG1655. This strain is a descendent of benign intestinal *E. coli* isolates. Subsequently, a number of different *E. coli* genomes have been sequenced, including O157:H7 strains. The first O157:H7 genome sequenced came from strain EDL933, isolated from contaminated ground beef from a McDonald's restaurant in Michigan. Once genomes were sequenced, a key question was to find out how they differed.

The degree of difference between the genomes of MG1655 and EDL933 is surprising: MG1655 has more than 500,000 bases of sequence not found in EDL933, whereas more than 1.3 million bases of sequence unique to EDL933 were identified, including about one-forth of it's 5,416 total genes. Thus, hundreds of distinct genes could be virulence factors for EDL933.

Bioinformatics allows us to develop hypotheses about the functions of proteins. Simply being present in in EDL933 but not MG1655 suggests that a gene could be a virulence factor. Evidence to strengthen this hypothesis can be acquired by using protein alignment to look for orthologs of putative virulence proteins that have been identified and studied in other organisms. Sequence similarity to a protein with a known virulence function or identification of protein domains suggestive of a virulence function are examples of such evidence.

Much is known about bacterial virulence, and based on that background knowledge, we would expect virulence factors to function in roles such as toxins, systems for delivering toxins to host cells, components of pili and other bacterial surface features allowing attachment to host cells, enzymes that break down host proteins, and proteins that sequester iron and other nutrients. However, it is important to bear in mind that even strong bioinformatics-based hypotheses require experimental testing – even minor sequence variations might result in altered functions or characterization of a gene with no obvious disease function might lead to the discovery of a new type of virulence factor.

Table 5.2 Candidate virulence genes from *E. coli* O157:H7 strain EDL933.

Gene Name[1]	NCBI ID	Gene Name[1]	NCBI ID
yadK	12512854	*ydgE*	12515577
yagW	12513076	*yeeJ*	12516151
ybbK	12513379	*yehC*	12516323
ybgP	12513628	*yhiF*	12518204
ycjZ	12515432	*ysaS*	12517366

[1]In bacterial genomes, gene designations beginning with *y* indicate genes whose identity is not yet sufficiently certain to merit a specific name.

Part I Activity: Using Protein Alignment to Explore Protein Function

BLAST – use BLAST to compare your protein sequence to known proteins in the NCBI database. A protein BLAST incorporates a substitution matrix to score amino acid similarity.

1. Choose a potential virulence factor to invest age from the table. Obtain the amino acid sequence in FASTA format using the NCBI protein database.

2. From the **BLAST home page**, choose protein blast to align an amino acid sequence query with database sequences. Paste the FASTA-formatted sequence into the BLAST query sequence box.

3. Use the Organism field to limit your search appropriately. For example, you could choose to limit the search to Gram-negative bacteria or even the Enterobacteria (the large family of intestinal bacteria to which *E. coli* belongs).

4. Add an additional Organisms field and use it to exclude *E. coli* from the search results – this prevents your results from being cluttered with high-scoring matches from EDL933 itself or other pathogenic *E. coli* strains.

5. At the bottom of the window, click Algorithm parameters to choose an appropriate substitution matrix: BLOSUM 62 is the default, but because the search is limited to relatively closely related organisms, perhaps it makes sense to try a matrix optimized for more closely related sequences such as BLOSUM 80 or PAM 70 (remember higher BLOSUM numbers and lower PAM numbers represent more similar sequences used to generate the matrix).

6. Run your BLAST search.

Now comes the important work of analyzing the results. Obviously, a high -scoring match (indicating a high degree of similarity between your query and some other protein) provides stronger evidence for a conserved function than a low-scoring match. Similarly, a good alignment along the whole length of the protein better supports functional conservation than a partial match. Review Chapter 4 of your textbook if necessary to refresh your memory of what the score and *e*-value mean.

7. If you find a good match, investigate the function of the putative ortholog: Is it found in a pathogenic bacterium? What is known about its function? Is there evidence that it is a virulence factor? Add this information to your one-page summary.

Conserved domains – a **domain** is a functional region of a protein. For example, an energy-requiring enzyme might have an ATP-binding domain as well as a substrate-binding domain where its catalytic function is carried out. A transcription factor would likely have a DNA-binding domain as well as a domain that interacts with RNA polymerase. Even if two proteins are not terribly similar overall, they might have a particular domain in common: Two DNA-binding proteins that have different functions might have similarity in their DNA-binding domains but be very different in a domain used for interactions with their distinct molecular partners.

While your BLAST search was running, you might have seen a page informing you that "conserved domains" have been detected in your query protein. If so, you should see a box at the top of your BLAST results page titled Putative conserved domains have been detected. BLAST looks for patterns in the query protein that resemble known functional domains and reports these results. The conserved domain box shows the regions of your protein that are similar to well-characterized functional domains; clicking on this display takes you to more information about the conserved domains and the other proteins that contain them. You can also run a conserved domain search directly without a BLAST search by searching NCBI'S Conserved Domains database.

8. Were any conserved domains been detected in your query protein? If so, investigate these domains and add this information to your one-page summary.

Substitution Matrices – What would happen if you changed the substitution matrix used in your search? You initially optimized it to give higher scores to substitutions likely to occur in closely related sequences, but what if you used a matrix like PAM 250 or BLOSUM 45 that is based on more distantly elated sequences? Although it is likely that the BLAST will still pick up the same high-scoring matches, there could be some less closely related proteins in the list, or you may notice changes in the score or *e*-value resulting from scoring mismatches.

9. Change the substitution matrix and rerun your BLAST search. Repeat for at least three different substitution matrices and inspect the results. Note any interesting alignments in your one page summary.

What would happen if you searched for matches to really distantly related organisms? Because the goal of this exercise is to identify potential virulence factors in *E. coli* it is appropriate to limit the matches to related bacteria, but perhaps you are curious to know whether your gene might have a human ortholog. Some bacteria-specific proteins have no identifiable human orthologs, whereas others have been conserved across this long span of evolutionary time. Still others are surprisingly similar to human proteins, leading to speculation about recent horizontal transfer between species.

10. Use BLAST to determine if your gene has a human ortholog with a substitution matrix chosen to score such distant relationships appropriately. Describe your findings in your one-page summary.

Part II: More Tools for Exploring Protein Function

Depending on your results for Part I, you may or may not have a strong, well-supported hypothesis regarding the function of your chosen genes. Below are brief descriptions of additional tools that you may use to further investigate your genes.

PSI-BLAST: PSI-BLAST is a variation of BLAST in which initial matches are used to refine the substitution matrix to identify even more distant matches. This is a good tool when you want to identify meaning alignments to distantly related proteins, such as when a simple BLAST search reveals no good orthologs. To use PSI-BLAST, start at the BLAST home page and choose protein BLAST as before, and then on the next page click on the PSI-BLAST button before start the search.

Pfam: Pfam is a database of protein families – groups of proteins already shown to be similar in structure and function. Particularly when a protein sequence of interest does not have a strong ortholog identifiable by a BLAST search or when the closest matches are partial or relatively low scoring, aligning the sequences with Pfam protein families may yield information about specific domains or regions of the protein. When matches are found, the Pfam database provides considerable information about the known functions, sequences, and structures of the matching families, including links to still more information.

MOTIF: Like Pfam, MOTIF looks for alignments between query amino acid sequence and functional domains and motifs (short sequence segments associated with some function). The difference here is that MOTIF is a “meta site” that allows you to search up to six databases at once.

DAS: The localization of a protein within a cell can also provide clues to possible functions. DAS (dense alignment surface) deals with one aspect of protein localization: whether the protein contains potential transmembrane domains that would suggest it is an integral membrane protein. Although there are various ways to approach this question (e.g. mapping hydrophobic amino acids), DAS uses an alignment-based approach in which it essentially looks for meaningful local alignments between the query protein and a set of unrelated known membrane proteins.

11. Try some or all of the tools mentioned above to further your understanding of your potential virulence factor genes. Add any new information you gain to your summary. Don’t forget to include which tools were used to find information and draw conclusions!

Part III: BLAST through BioPython

Now, that you have investigated two candidate virulence factors in depth, you will perform a more extensive search using BioPython.

For all candidate virulence factors from the table presented at the beginning of this document:

1. Use BioPython do an Entrez search for each candidate. Save sequence into a separate output file using SeqIO module. You may want to utilize programs given on March 3. Keep in mind, we are interested in proteins, not nucleotides, so you will need to make certain modifications.

2. For each candidate, run your chosen protein sequence through BLAST using BioPython. Note, you must run your BLAST results using Python for this portion of the lab. You may utilize any of the programs given in class (especially programs discussed on March 3 and March 10).

Use four different matrices for each BLAST run of each candidate. Compare different matrices and discuss trends. You may gather and analyze some statistics, such as frequency, e-value, score information, etc. to perform your analysis (utilize programs from March 3, e.g., BioPythonBlast3, BioPythonBlast4, ch5BlastAA). Report those results in a table similar to the one shown below, include your analysis summary and a table in your report.

Gene Name	**Stat1 ...**	**Stat2 ...**
yadK	…	…
yagW		
ybbK		
ybgP		
ycjZ		
ydgE		
yeeJ		
yehC		
yhiF		
ysaS		