CMPSC 300
Bioinformatics
Fall 2019

Lab 7:
Investigating Potential Virulence Factors in *E. coli*

GitHub starter link

https://classroom.github.com/a/UDC_MKY0

To use this link, please follow the steps below.

- Click on the link and accept the assignment.
- Once the importing task has completed, click on the created assignment link which will take you to your newly created GitHub repository for this lab.
- Clone this repository (bearing your name) and work on the practical locally.
- As you are working on your practical, you are to commit and push regularly. You can use the following commands to add a single file, you must be in the directory where the file is located (or add the path to the file in the command):
  - `git add -A`
  - `git commit -m ‘‘Your notes about commit here’’`
  - `git push`

  Alternatively, you can use the following commands to add multiple files from your repository:
  - `git commit <`*nameOfFile*`> -m ‘‘Your notes about commit here’’`
  - `git push`

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.

Part 1: 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 its 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

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.

Figure 1: This is Table 5.2 of *Exploring Bioinformatics: A Project-based Approach, second edition*, by Caroline St. Clair and Jonathan E. Visick, page 89.

gene with no obvious disease function might lead to the discovery of a new type of virulence factor.

Virulence Factor Description

Choose **two** (2) candidate virulence factors from Table 1. For each selected factor, write a one-page summary discussing its likely function based on its conserved domains and orthologous proteins. You will have to perform some research using databases from the National Center for Biotechnology Information (NCBI) and, likely, several primary source articles to follow. **Please be sure to properly cite your articles when they are used.**

Each summary should include a table of the organisms that you have identified where similar proteins are shared between them. You are also to add details in the report of the known functions of these proteins (available by consulting NCBI's databases), the per cent similarity according to your BLAST matches, and the type of BLAST used 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.

Activity Steps: Using Protein Alignment to Explore Protein Function

BLAST is a tool which permits researchers to locate sequence (or parts of the sequence) across all known sequences. These sequences are stored in a database which is constantly updated with new sequence information once it becomes available. When running a BLAST-search, all sequences which have some similarity to a sequence of interest are determined and presented to the researcher.

Use the protein BLAST tool at NCBI to compare your protein sequences to *all known proteins* listed in the NCBI protein database. A protein BLAST experiment incorporates a substitution matrix (`https://en.wikipedia.org/wiki/Substitution_matrix`) such as PAM (`https://en.wikipedia.org/wiki/Point_accepted_mutation`) or BLOSUM (`https://en.wikipedia.org/wiki/BLOSUM`) to score amino acid similarity. *Note: you may need to read your chapter for more information about these substitution matrices.*

1. Choose a potential virulence factor from Table 1. 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 NCBIS 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 2: 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. **You are invited to investigate these tools.**

1. *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 does not reveal any meaningful 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.

   - `https://blast.ncbi.nlm.nih.gov/Blast.cgi`

2. *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.

   - `http://pfam.xfam.org/search/sequence`

3. *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.

   - `http://www.genome.jp/tools/motif/`

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. Do not forget mention in your report which tools were used to find information and draw your conclusions.

Required Deliverables

All of the deliverables specified below should be placed into a new folder named 'lab06' in your Bitbucket repository (`cs300f2017-bbill`) and shared with the instructor by correctly using appropriate Git commands, such as `git add -A`, `git commit -m ''your message''` and `git push` to send your documents to the Bitbucket's server. When you have finished, please ensure that you have sent your files correctly to the Bitbucket Web site by checking the `source` files. This will show you your recently pushed files on their web site. Please ask questions, if necessary.

1. Part 1

- File:`writing/report.md`: A description of each chosen candidate virulence factor. Be sure to read above in Section "Virulence Factor Description" for details about this description.

2. Part 2

- File:`writing/report.md`: Answer the questions in the document regarding the new tools. These questions may be found in the markdown document that you are to edit.

You should see the instructor if you have questions about assignment submission.