CSCI 4181 / CSCI 6802 – Alignment and Distant Homology (due Wednesday, February 23, 2025)

Overview

This is the first of four assignments in the course that will give you practical experience with biological data. The due date for this assignment is **Tuesday, February 23, by midnight**.

- Please submit your completed assignment on CSCI 4181/6802 Brightspace: Assessments -> Assignment 1 - Alignment and Distant Homology.
- Please submit your assignment as **one file in PDF format**.
- Please name your file following this naming scheme: **BannerID_LastName_Assignment1.pdf**.

Questions that you need to answer are numbered and marked in **boldface**. The point value of each question is indicated next to the question. Be sure to read the questions thoroughly and address each part.

The goal of this assignment is to test your theoretical knowledge of sequence alignment and give you some practical experience with methods as well.

Academic Integrity

All assignments in this course are to be completed individually. The following are specifically forbidden and will lead to a submission to the Academic Integrity Officer:

- Submission of any responses (text or otherwise) that were in any way produced or partially produced by generative AI.
- Any text reproduced in part or in whole from online or other sources without attribution. Answers must be in your own words unless otherwise specified.
- Collaboration on any assignment.

I am happy to answer any questions you may have about academic integrity. If you are unsure, please ask.

Note on datasets: The RefSeq database that we will use is updated periodically, and e-values change. I once set a tutorial that worked on Day X, but due to a database update the assignment no longer worked correctly on Day X+7. We'll keep an eye out for this (as well as ensuring the NCBI website stays up).

Running the assignment:

- Part 1 of the assignment does not involve running software.
- Part 2 relies entirely on online tools.
- In Part 3 you will use the **Mash** software to perform genomic comparisons.

The easiest way to complete Part 3 is to do so on the Timberlea Computer Science servier (login details below). Everything has been tested on this server and later assignments will also be based around it. However, you are welcome to use your own system. The Mash software is Linux based and runs at the command line on Linux and Mac systems. You can also run it in Windows using Windows Subsystem Linux. We have provided a compiled Linux binary but on other systems you may have to recompile it yourself or retrieve an alternative binary from the Mash site.

The Linux Environment

The server you can optionally use, timberlea.cs.dal.ca, is a Linux server that is accessible to students in Computer Science. Everyone registered in a CS course is automatically assigned a CSID (see here for details https://csid.cs.dal.ca/). If you don't have the login credentials, please let us know, and we will ensure you get set up.

You will need to run a few Linux commands in this assignment. A couple of key ones are:

`mkdir` - Make a new directory

`cd` - Change to a directory

`ls -1` - List directory contents (that's the letter 'l', not the number '1')

`ls -lrt` - As above, but list directory contents in order of time ('t'), reversed ('r'), so the most-recently updated file is at the bottom of the listing. This can be fun if you want to obsessively track the progress of a run.

Very very very useful tip: If you start typing the name of a file or directory, you can autocomplete it by hitting the tab key. If there is some ambiguity, you may need to type a couple more letters. This can save a LOT of time. Also, you can cycle through previous commands using the up arrow.

You will need to connect to **timberlea** using the secure shell 'ssh' protocol. You can do this directly from the command line on a Linux or Mac OS system with the following command:

```
ssh <your CS ID>@timberlea.cs.dal.ca
```

On Windows you can use the 'putty' (https://www.putty.org/) package or ssh directly from Windows Subsystem Linux.

If I want to copy the file 'blah.txt' from my home directory on timberlea, I would type from my own computer:

```
scp finlaym@timberlea.cs.dal.ca:blah.txt .
```

That dot at the end means 'copy to my current location' on the current computer.

If you're on a Windows machine, you can use WinSCP (https://winscp.net/eng/download.php) or scp directly from Windows Subsystem Linux.

Part I: A tiny bit of data

Q1-Q3, 6 points total

Q1. (2 points)

(a) What do the scores in the PAM250 matrix (e.g., https://www3.nd.edu/~aseriann/CHAP7B.html/sld017.htm, shown on the last page of this assignment) represent? How are they calculated?
(b) How is a PAM250 matrix different from a PAM1 matrix?

Q2. (2 points)

Why might the score for a given amino acid pair be different in a BLOSUM matrix than it is in an entropy-matched PAM matrix?

Q3. (2 points)

Here, is an amino acid sequence:
Sequence 1: MKVE

Select any **two amino acids** from Sequence 1 and replace them with different amino acids from the same category (e.g., polar, nonpolar, acidic, or basic) to create Sequence 2.

Your task in this question is to generate an alignment of the original Sequence 1 with your updated Sequence 2.

Show the dynamic programming matrix (generated using the global Needleman- Wunsch method) for these two sequences, given the attached PAM250 scoring scheme and a linear gap penalty of -3 per gap position. Also indicate the optimal alignment path through the matrix and the resulting pairwise sequence alignment.

Part II: More data: investigating the *Salmonella* invasion (Sip-D) protein

Q4-Q9, 8 points total

Part 2a: PSI-BLAST

The bacterial type III secretion system (T3SS) is crucial to produce biotechnologically relevant proteins and facilitating protein delivery in synthetic biology applications. The T3SS forms a needle complex embedded in the membrane, capped by translocon proteins, extending into the extracellular space (Wikipedia, very detailed review).

========== molecular biology bit ==========

In *Salmonella enterica*, the needle tip complex comprises three "translocon" proteins that can transfer proteins and other molecules from one side of the cell membrane to the other: these are named SipB, SipC, and SipD. Antibodies against SipD hinder *Salmonella* invasion, suggesting SipD as a potential target for blocking SPI-1-mediated virulence. The N-terminal domain of SipD facilitates effector secretion at post-transcriptional and post-translational levels.

Additionally, SipD from *S. Typhi* acts as a potential antigen for diagnosing typhoid fever, aiding in timely patient diagnosis and treatment.

========== end molecular biology bit ==========

This raises an interesting question whether the cell invasion protein is homologous with any other proteins we know about. It's reasonable to expect that we will get some results if we compare against a reference database, but will we turn up anything surprising (e.g., non-*Salmonella* related)?

Let's explore this question using the NCBI RefSeq database and the BLAST algorithm.

Note: PSI-BLAST may throw the occasional error. If this occurs, please try refreshing your browser and continuing your search.

You can start by pointing your browser to https://www.ncbi.nlm.nih.gov/protein/?term=cell+invasion+protein+SipD which brings you to the main page for SipD in the NCBI Protein database. You'll see a search list of hits for SipD across organisms, with their cross references for all NCBI databases. To the right you will see "Results by taxon" and an option to filter the search by "Top Organisms". To the left, there is a tab called "Source databases" where you could attempt to filter the search results by "RefSeq".

You should see this box near the top of the results screen:

Click on "sipd reference sequences Protein (1)" list to go to our protein of interest. On selecting the first search hit, you will find information about the protein, including the authors and references, associated comments, coordinates, features, source organism, and finally the protein itself. Yep, it's a bunch of letters symbolizing the twenty different amino acids – no surprise there. Down the right-hand side of the page are links to other databases that make NCBI an extremely valuable resource. They don't have a fantastic domain breakdown of the protein, but Figure 1a in https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6689963/figure/F1/ has a good summary of the different pieces. This will become important once we start to explore the homology-search results.

Near the top right-hand corner of the page under "Analyze Sequence", you will see a "Run BLAST" link; open this in a new tab (we'll want to keep this sequence for later) to the submission page. This takes you to the protein-protein BLAST (BLASTP) page. Since we came from the SipD protein, the query window is already populated with the accession number of the protein, but we could have also pasted the raw amino-acid sequence. We will now set up a PSI-BLAST query.

Orange indicates options that must be changed from the default. Be sure to make each of these changes or you might not be able to find the proteins we are looking for! Non-default parameters are highlighted on the webpage, so you can see which parameters have been changed.

The default database to search against is the massive 'nr' repository. This database is huge, and if we use it as our reference the first 5000 hits will all be identical to the query, which is not so useful. Instead, **we will use the reference protein sequence database (here called refseq_protein)**, which uses representative sequences extracted from the behemoth nr. There are other options such as the much smaller set of proteins in the Protein Data Bank (PDB) which have known structures, but this carries the risk of missing any weird stuff we might want to discover since PDB is a relatively small database.

Under 'program selection' choose PSI-BLAST.

Under 'Algorithm Parameters' you have a few options, although if you're serious about BLAST you will want to download the command-line version in the NCBI BLAST+ suite, which gives you much finer control over parameter options and lets you BLAST huge numbers of sequences against each other.

Set "Max target sequences" to 5000. We're going to cast the net widely here.

Set “Expect threshold” to 1. This is the maximum e-value of hits that will be shown to us.

Q4 (1 points). What is the meaning of the expectation value? In particular, what does an expectation value of 1 correspond to?

Turn on filtering of low-complexity regions.

Q5 (2 points). Why is it important to filter out low-complexity regions? Find an example of a low-complexity region in the literature, and give the sequence of the region, the organism it is found in, and a citation to the paper where you found this sequence.

At the bottom of the screen, select “Show results in a new window”.

Let’s submit the query. It will likely take 60 seconds or more to run each iteration (depending in part on the time of day!), but you will eventually get a list of matches to the query sequence. You will see a tab called “Graphical Summary” which has an interactive graphical overview of the hits (significantly matching proteins from the database) with an indication of the bit score S and where statistically significant local alignments are found. The “Descriptions” tab gives us two lists of proteins: those that match with a corresponding e-value less than our PSI-BLAST threshold ($e < 0.005$), and those that match with an e-value between 0.005 and 1 (a relatively small number).

Uncheck all the sequences. Click on the **worst** sequence match that satisfies our PSI-BLAST threshold (i.e., first list, e-value slightly less than or equal to 0.005) under the “Descriptions” tab to see the alignment between the worst hit and our query protein. Hmmm.

The “Alignments” tab gives us, you guessed it, alignments returned by BLAST for the query sequence against the matching database sequences. Your first hit should be pretty much identical to the query sequence (because it *is* the query sequence); as you go further down, you will see more and more differences between query and subject.

Finally, back at the top if you click "Search Summary", you will get a list of parameters and statistics about the query you just performed:

- Values of K and lambda that are used to estimate the significance of database matches,
- The number of sequences that were considered at every step of the BLAST procedure,
- The matrix and gap penalties that were used for the query.

The upper right-hand corner has a “Filter Results” pane where you can limit the display of sequences to only those that satisfy % identity, e-value, or query coverage ranges. Query coverage is the percentage of the query sequence that is actually aligned with the database match; low coverage can be a warning sign that only a small part of your sequence is homologous to the match. This can occur when only parts of two sequences are homologous, for example when they both share a DNA-binding domain but nothing else. It is very common to apply a query coverage filter to the BLAST results.

Let’s filter our sequences with this in mind. **Set the percent identity range to “50 to 100” and the query coverage to the same**. We won’t be building profiles from any of the worse matches, but those worse matches will likely come up in our next PSI-BLAST iteration anyway. In developing this assignment I had 257 sequences matching this threshold; your results may vary a bit.

Now, find the line that says, ‘Run PSI-BLAST iteration 2 with max 5000’. Click 'Run' and ponder the next question while you wait for BLAST to run.

Q6. (1 point) How does PSI-BLAST use information from the first round of hits to generate the second round of hits?

Your displayed results from Run 2 will still be filtered by % ID and cover. **Click Reset right next to “Filter” to show the entire set of results**.

New matches not detected in Round 1 will be highlighted in yellow.

The sequences with e-values near the PSI-BLAST threshold are interesting because they may give us some context for “divergent but still homologous”. Take a look at the alignment of this sequence with the query.

Q7. (1 point) Look at one of the matches with an e-value just above the e-value threshold. Based on this alignment, do you think the original query and newly matching proteins are significantly similar to one another and therefore homologous? Why or why not?

Let’s do one more round (Iteration 3) of PSI-BLAST. Some of the poorest matches will be included in the model, which could have interesting consequences. And yep, there are more matches.

Q8. (1 point) Now look at one of the really bad matches (again, just above the e-value threshold). How does this alignment compare to the one in the previous question? Do you think this sequence is homologous with the original SipD protein?

Keep in mind that I do not know the answer to the last two questions – I have an opinion, but the question can be argued either way.

Part III: Computing DNA sequence distances using Mash

Q10-Q12, 6 points total

In this part of the exercise, we will use a dimensionality reduction method with a tool called MASH. MASH is designed to compare and group whole genomes and metagenomes on a large scale. It uses a method called MinHash to compress large sequences into smaller, simplified "sketches." This allows for tasks like quickly choosing the best reference genome for mapping reads or identifying poor-quality or mislabeled samples that don't cluster as expected. To do this, MASH breaks the sequence into smaller pieces called k-mers by sliding a window of length k across it.

Download Mash and files requires for this question from the course website: maguire-lab.github.io/bioinformatics_algorithms_2026/static_files/assignments/assignment1_materials.tar.gz

We will need to use the commands explained in section "The Linux Environment"

If using Timberlea: Using scp in the terminal on your local computer where your files are stored to copy them onto the server.

```
scp assignment1_materials.tar.gz
ssnmurthy@timberlea.cs.dal.ca:/users/grad/ssnmurthy
```

(replacing with your csid and timberlea home directory)

Note that this is a single command, do not hit Enter to separate the two parts.

Enter your password when prompted and press enter.

tar is an archived format, so to extract the folder and binary the command is:

```
tar -xvf assignment1_materials.tar.gz
```

you will see a folder with the name `assignment1_materials` with a binary "mash" inside it and input files called `assignment1_genomes.fasta`

Source Data: We have prepared a single file, "assignment1_genomes.fasta" that contains the sequences of chromosomes and plasmids of ten members of the taxonomic family Enterobacteriaceae. This family comprises many genera, most of which are anaerobes (they don't like oxygen) that live in our guts. Familiar names such as *Escherichia* and *Salmonella* are complemented with less well-known but equally nasty genera such as *Citrobacter* and *Klebsiella*, plus the occasional insect symbiont (?). We are going to use Mash to explore the distances between these and start to gain insights into (i) similarity relationships which anticipate the phylogenetics module, and (ii) how Mash works.

You can get a quick summary of what's in the file by running:

```
grep ">" assignment1_genomes.fasta
```

Which will show only the header lines for the file. We have shortened the header lines for readability but you can use the NCBI accession numbers to access the original record. Any row with "pXXX" is a plasmid, every other row is a chromosome.

Once you login to the server (or on your own local shell), enter the folder containing the mash binary using

```
cd assignment1_materials
```

```
./mash -h
```

which will display the Mash help manual with all its parameter options and defaults.

It's time to run Mash. The Mash "triangle" command estimates the distance of each input sequence to every other input sequence. It outputs a lower-triangular distance matrix in relaxed Phylip format which we save into an output file output.txt to open and interpret later.

```
mash triangle -i -C assignment1_genomes.fasta > output.txt
```

The command used above, "triangle" has a few parameters, including:

`-i` Sketch individual sequences, rather than whole files, e.g. for multi-fastas of single-chromosome genomes or pair-wise gene comparisons.

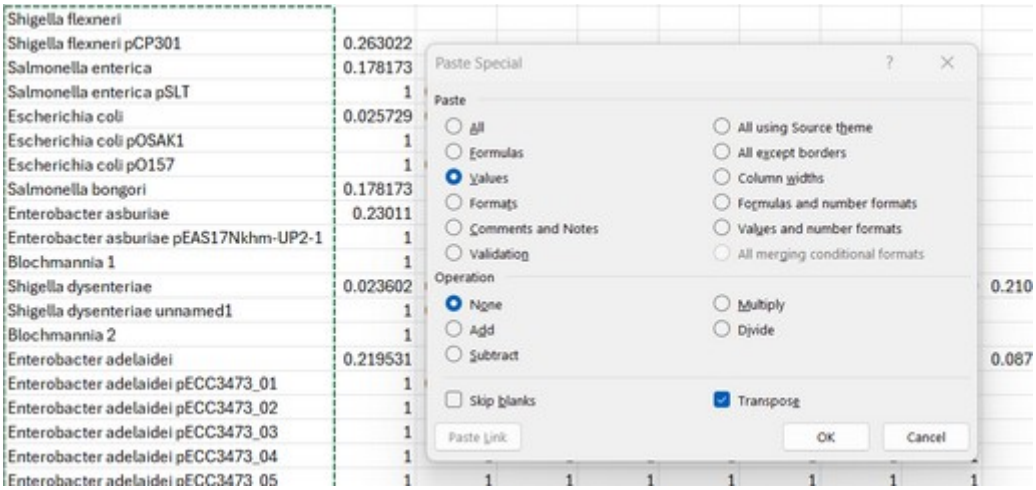
`-k` <int> *K*-mer size. Hashes will be based on strings of this many nucleotides. [default *k* = 21)

`-s` <int> Sketch size. Each sketch will have at most this many non-redundant min-hashes. (default s = 1000)

Mash is **fast**. These genomes are small and few in number relative to many datasets we and others work with, but even this set can take hours to analyze in a more comprehensive and precise way. Mash is great for a first look but is not a substitute for proper phylogenetic analysis.

"output.txt" is a tab-separated file containing the distance matrix between all pairs of DNA sequences in the original file. The value in the upper left-hand corner of the triangular matrix represents the distance between the first two sequences, and so on. Open this file in a spreadsheet and look at the produced distance values. It will make life a lot easier if you copy the row names as column headers. You can do this by copying the rows and pasting them across the top using the "transpose" option.

First step (this should look similar in other spreadsheet applications):



Final result:

Shigella flexneri	Shigella fle	Shigella fle	Salmonella	Salmonella	Escherichi	Escherichi	Escherichi	Salmonella	Enterobac	Enteroba
Shigella flexneri pCP301	0.263022									
Salmonella enterica	0.178173	1								
Salmonella enterica pSLT	1	0.295981	1							
Escherichia coli	0.025729	0.295981	0.191731	1						
Escherichia coli pOSAK1	1	1	1	1	1					
Escherichia coli pO157	1	0.147707	1	0.263022	1	1				
Salmonella bongori	0.178173	1	0.094981	1	0.178173	1	1			
Enterobacter asburiae	0.23011	1	0.210897	1	0.219531	1	1	0.23011		
Enterobacter asburiae pEAS17Nkhm-UP2-1	1	1	1	1	1	1	0.295981	1	1	
Blochmannia 1	1	1	1	1	1	1	1	1	1	
Shigella dysenteriae	0.023602	0.295981	0.191731	1	0.017127	1	1	0.182269	0.210897	
Shigella dysenteriae unnamed1	1	0.154223	1	0.23011	1	1	0.111073	1	1	
Blochmannia 2	1	1	1	1	1	1	1	1	1	
Enterobacter adelaidei	0.219531	1	0.219531	1	0.219531	1	1	0.219531	0.087751	
Enterobacter adelaidei pECC3473_01	1	0.263022	1	1	1	1	1	1	1	
Enterobacter adelaidei pECC3473_02	1	1	1	1	1	1	1	1	1	0.14770

Q10a. How can one obtain a Mash distance of 1 when comparing two sketches? (1 point)

Q10b. What is the minimum distance between any pair of DNA sequences in the set? What does this tell us about those two sequences? (1 point)

As stated above, the default *k*-mer size is 21, which may or may not be suitable for any given dataset. Try re-running with a different *k*-mer size that reduces the number of distances of 1 in the dataset. You may see warnings produced; for the purposes of this assignment we will not worry about them and focus on the produced output.

Q11. How might increasing or decreasing the k-mer size affect the accuracy and sensitivity of the results? Indicate what k-mer size you used in your second run and describe how the distance values have changed. (2 points)

The other primary way we can tweak the distances is to change the sketch size. Try running with a smaller sketch size and assess the impact on the distances.

Q12. What is the impact of the smaller sketch size on the resulting distances? Why does this change occur? (2 points)

PAM250 matrix

C Cys	12																		
S Ser	0	2																	
T Thr	-2	1	3																
P Pro	-3	1	0	6															
A Ala	-2	1	1	1	2														
G Gly	-3	1	0	-1	1	5													
N Asn	-4	1	0	-1	0	0	2												
D Asp	-5	0	0	-1	0	1	2	4											
E Glu	-5	0	0	-1	0	0	1	3	4										
Q Gln	-5	-1	-1	0	0	-1	1	2	2	4									
H His	-3	-1	-1	0	-1	-2	2	1	1	3	6								
R Arg	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6							
K Lys	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5						
M Met	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6					
I Ile	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5				
L Leu	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6			
V Val	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4		
F Phe	-4	-3	-3	-5	-5	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9	
Y Tyr	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10
W Trp	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W