CSCI 4181 / CSCI 6802 – Multiple Sequence Alignment (January 26, 2022)

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 **Friday, February 11, 2021**, by midnight.

Please submit your completed tutorial on **CSCI 4181/6802 Brightspace: Assessments → Assignment 1**. Although the assignment file can be in any format, the most compatible format is Word .docx. The preferred filename format is "BannerID_LastName_Assignment1.docx".

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

Note: 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 completely failed to make the point on Day X+7. We'll keep an eye out for this.

Part I: A tiny bit of data (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?

(b) How is a PAM250 matrix different from a PAM1 matrix?

Q2. (2 points) Why might two different amino acids have a score > 0 in the PAM matrix?

Okay, so here are two sequences:

S_1: MERA

S_2: MKVE

Q3. (2 points) 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 2: More data – investigating the SARS-CoV-2 spike protein (8 points total)

Like many viruses, the SARS-CoV-2 genome is pretty small: just under 30,000 nucleotides of RNA. Also like many viruses, SARS-CoV-2 packs its genes together very tightly, with almost no space

between genes. The business end of SARS is the spike (S) protein, which is responsible for binding to the ACE2 receptor of human cells and fusing with the cell membrane. If there's one thing RNA viruses like to do it's to mutate at a blistering pace, and our coronavirus is no exception, which is why we're constantly on the lookout for new variants. The most concerning of the mutations that lead to the emerging "variants" are found in the S protein, particularly in the receptor binding domain of S which does the actual linking up with ACE2.

An interesting question we can ask is whether the spike protein is homologous with any other proteins we know about. Since SARS-CoV-2 is closely related to a bunch of other coronaviruses, 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-coronavirus related)?

Let's explore this question using the NCBI RefSeq database and the BLAST algorithm. You can start by pointing your browser to https://www.ncbi.nlm.nih.gov/protein/?term=sars-cov-2+spike, which brings you to the main page for S. You'll see a bit of information here, including a low-resolution map of the genome. The genome includes "ORF1ab" and "ORF1" which encode a bunch of smaller proteins that are synthesized together and then chopped up post-translation. Right after ORF1ab is the S protein, which is followed in turn by a bunch of short but very important protein-coding genes.

Just above the genome map you will see a link to "RefSeq proteins" – click on this to go to our protein of interest. Here you will a bunch of information about the protein, including the primary reference, associated comments, coordinates, isolation source, 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 a bunch of 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.nature.com/articles/s41586-020-2180-5 has a good summary of the different pieces. This will become important once we start to explore the homology-search results.

On the right-hand side of the page you will see a "Run BLAST" link; click it to go to the submission page. This takes you to the protein-protein BLAST (BLASTP) page. Since we came from the spike 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. Guess what, there are a lot of spike protein sequences, and if we use this as a reference database 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 (2 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".

All right, 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 0.01 (of which there are zero as of January 26).

Click on the worst sequence match 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.

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

Your results from Run 2 should be pretty striking, consisting of:

(1) E-values of 0 (definitely homologous!),
(2) Questionable e-values (> 0.001),
(3) A handful with e-values between those two extremes.

The sequences in group (3) is interesting because it gives 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 “questionable” matches. Based on this alignment, do you think the query and subject proteins are significantly similar to one another and therefore homologous? Why or why not?

Keep that alignment (do a screenshot, download, or copy-and-paste into a text editor) as we will use it in the next question.

Let’s do one more round of PSI-BLAST. A few of our “questionable” e-values 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. How does this alignment compare to the one in the previous question? Do you think this sequence is homologous with the original spike 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 3: Actually, I’d like to align more than two sequences, thanks (6 points)

In this part of the exercise we are going to use a heuristic multiple sequence alignment program (MUSCLE) to build a multiple alignment of protein sequences. Our data source for this step will be the PFAM database (http://pfam.xfam.org/), which contains alignments and hidden Markov models for a large number of different homologous sets of proteins.

You will choose your own family of interest by doing the following:

(i) Go to “Browse”: http://pfam.xfam.org/browse
(ii) Choose a letter to browse within “Families”

(iii) In the resulting table, choose a record of type "Family", with average length 150 or greater. You'll be chopping the family down to eight sequences, so either choose a family with this number, or choose a number with more that you will cut down later.

(iv) Follow the ID or Accession link of the chosen family type.

(v) Select "Alignments" on the left-hand side of the page, and on the resulting tab, find the line "You can also download a FASTA format file containing the full-length sequences for all sequences in the full alignment." Click this to get the .gz file.

(vi) Open the resulting .gz archive (Windows users: you may need a third party file compression utility like 7zip or WinRAR). Extract the FASTA file, open it in a text editor, and choose eight sequences (plus their headers).

We will use the MUSCLE Web service at https://www.ebi.ac.uk/Tools/msa/muscle/ to perform the multiple alignment. You can paste the sequences from your saved text file directly into the "input sequences" window.

They don't offer much in the way of options for the Web version, so just hit 'Submit' and wait for your run to complete. You could also give them your e-mail address and receive the results that way, but aligning six sequences will go pretty quickly.

While you are waiting for the answer to come back, you can think about the following question:

Q8. (4 points) MUSCLE is a progressive alignment method. What are the key steps in the progressive method, and why might it be better than simultaneous alignment of all sequences?

Once the alignment is finished, you will see a nice output window that stacks the residues (another term for amino acid) from all six sequences on top of another. You may also see a ton of gaps as well, depending on how dissimilar the sequences are.

Q9. (2 points) Choose any column from the alignment that has at least six residues (in other words, is mostly not gaps). Write down the amino acids form this column and show the formula for the sum-of-pairs (SP) score. Using PAM250 again, calculate the SP score for this column. What is the maximum theoretically possible SP score for a column with this number of residues, under the PAM250 scoring scheme? (Hint: Recall that if you have N items, there will be N*(N-1)/2 unique pairs).

PAM250 matrix

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
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