CMPSC 300
Bioinformatics
Spring 2021

Lab 5 Assignment:
By Global Alignment or by Local Alignment?
Submit deliverables through your assignment GitHub repository and complete the Google Form.

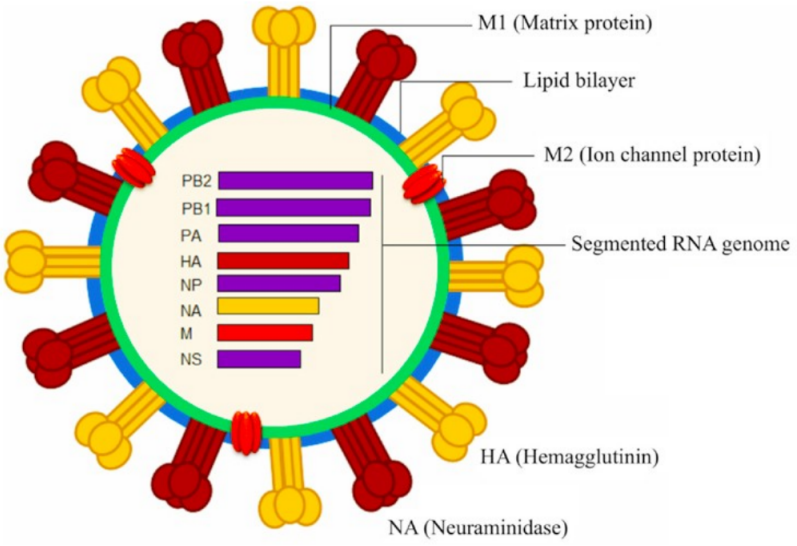

Figure 1: **Schematic diagrams of influenza A virus**: The segmented genome of influenza A virus encodes three envelope proteins (hemagglutinin, neuraminidase, and ion channel M2 protein), and internal nucleoprotein (NP), polymerases (PA, PB1, and PB2), matrix protein 1 (M1), and non-structural proteins (NS). The lipid bilayer is derived from host cell membrane.

Objectives

To understand the value of aligning genes and to recognize its practical applications. To gain more familiarity and experiencer with the use of Web-based alignment tools to explore sequence similarity and understand how to modify their parameters. To gain experience with how the Needleman-Wunsch algorithm optimally aligns any two sequences and to understand how this algorithm could be ued to learn about virus types.

Clone Your Assignment Repository

In this section, we will be using `Git` commands. It is suggested that the reader refer to online searches for help. For example, GitHub provides good documentation at the following link; `https://git.github.io/htmldocs/git.html`.

Handed out: 29^{th} March 2021

In many cases, you will be given a new repository containing assignment materials and you will save your files in this assignment repository as you continue to work on them. Copy and paste the assignment repository cloning command into your terminal to create your assignment repositories. Be sure to place your assignment repositories in a directory such as `cs300/` to keep your class materials organized by class.

Today's assignment repository can be found at the below link to a GitHub Classroom repository. Here you will work on your assignment and then push your work to the cloud where the instructor will be able to view your work for grading. Often, there will be files in your assignment repositories which you are to edit before you submit them by using the below commands for `git`.

https://classroom.github.com/a/2gjyji2a

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 locally
- As you are working on your lab, you are to commit and push regularly. The commands are the following.
  - `git add -A`
  - `git commit -m ‘‘Your notes about commit here’’`
  - `git push`

Check Your Submission

After you have pushed your work to your repository, please visit the repository the GitHub website (you may have to log-in) to verify that your files were correctly sent. Importantly, please check that GitHub Actions has checked your submission. For this, look for an orange dot that will turn into a red check mark to indicate errors, or a green check on the top line of your repository to indicate that all checks have passed.

Reading Assignment

Chapter 3 in the *Exploring Bioinformatics* textbook.

Part 1: Small Tasks using the Needleman-Wunsch Algorithm

In this part of the lab you are invited to further practice using a particular global alignment technique called the *Needleman-Wunsch* algorithm. In this lab, you are to create a report document that provides the answers or solutions to the following tasks:

Table 1: Use the lettering-system of this matrix to indicate the calculations of your own implementation of the NeedleMan-Wunsch algorithm for alignment. In your calculations document, please label each series of calculations according to the letter in the cell. There are instructions in the calculations file.

		A	**T**
	0	-1	-2
A	-1	(a)	(b)
T	-2	(c)	(d)
G	-3	(e)	(f)

1. **By Hand**: Compute the alignment of the two sequences shown in the Table 1 using the Needleman-Wunsch algorithm. The **match**, **mismatch** and **gap** scores are, 1, 0, -1, respectively. In your written work using the file writing/calc.md, please be sure to have all calculations listed on the relevant lines according to the letter of the cell. You will note that the recommended formatting of your answers is set-out in your submission file; `calc.md`.

2. **By software**: Use one of the software-based solutions that we implemented in class (i.e., the BioPython code from class or the online interactive demo at link, `http://experiments.mostafa.io/public/needleman-wunsch/`) to implement a global alignment of the sequences of files; `data/s1.fasta` and `data/s2.fasta`. Answer the below Questions-in-blue for part 1. USe the default **match**, **mismatch** and **gap** scores in the software that you use.

   (a) Which software did you use to conduct your analysis?

   (b) How similar were the two sequences of your input files; (`data/s1.fasta` and `data/s2.fasta`), which were applied to an alignment program?

   (c) Are the two sequences closely related to each other, in your opinion?

   (d) What proof do you have to suggest such a claim?

Part 2: Using Online Tools to Investigate the Influenza Virus

In the second part of the lab you are invited to explore online global and local alignment tools to investigate similarity in viruses.

Influenza Viruses

The genomes of the influenza viruses are divided into eight segments, each representing essentially the coding information for a single protein (Figure 1). Segment 4 contains the gene for hemagglutinin (*HA*), the viral surface protein essential for the initial interaction between the virus and its host cell. *HA* is one key determinant of which host(s) a particular virus can infect, because the virus cannot replicate or cause disease without being able to first bind to a host cell. The *HA*s of one of the major seasonal human viruses circulating before 2009, the 2009 H1N1 pandemic virus,

and the 1918 human pandemic virus are all classified as the H_1 type, whereas recent outbreaks of severe avian flu are caused by a virus with *HA* classified as H_5 . These classifications are based on binding of antibodies of known specificity, but sequence alignment provides much more detailed information about similarities and differences and where changes have occurred.

Influenza viruses have received a great deal of study, and the ability to compare many strains has led to significant advances in understanding what allows one virus to cause a more severe disease than another. The H_5N_1 "bird flu" virus makes an interesting case in point. The virus causes severe influenza in birds and has become established in populations of domestic chickens and turkeys. Human cases of influenza also occur sporadically, mostly in individuals heavily exposed to infected birds, such as poultry farmers. Interestingly, once a human case occurs, however, spread to another human is exceedingly rare, even among family members in close contact with the infected individual.

In a 2006 article by van Riel *et al.* [1], the authors demonstrated that the avian H_5N_1 virus binds to a form of sialic acid receptor that, in humans, is only found in lung tissue of the *lower* respiratory system. Human viruses, in contrast, bind to a form of the receptor common in the *upper* respiratory tract. Thus, it is difficult for H_5N_1 to infect humans because our respiratory defenses normally prevent the virus from reaching the lungs. However, a mutant strain in which the virus was altered to be able to bind to sialic acid receptors in the upper respiratory tract could be a very dangerous strain indeed for humans.

So far, no such H_5N_1 strains that infect humans efficiently have been observed. However, we might ask whether the strains that do make it into humans tend to have altered genes - if so, that would suggest that either adaptive mutations could be occurring within the human host or that the viruses that cause human infections are subpopulations that are already better adapted. There are many avian H_5N_1 sequences available and a number of sequences of the H_5N_1 viruses isolated from infected humans, so we can use sequence alignment to see whether these have essentially the same strain or one of noticeable differences (which may present more dangers to ourselves).

1. Find the listed files (below) that have been have been placed in the `data` directory of your repository.

   - `Influenza_A_Chicken_Vietnam2005_avianH5N1_segment4.fasta`
   - `Influenza_avianHong_Kong2007_avianH5N1_segment4.fasta`
   - `Influenza_A_ChinaGD012006_humanH5N1isolate_segment4.fasta`

2. You are to go to the EMBOSS *Needle* webpage at `www.ebi.ac.uk/Tools/emboss/` to perform a global alignment of the above three sequences. Note, there will be three alignments (of sequence pairs) to perform. For example for the sequences, a, b, and c, there will be comparisons of (a, b), (a, c) and (b, c). **Please save your results as a screen shot or as a text file to include in your report.**

3. **Questions-in-blue**: Add the following responses to your report:

   (a) How much similarity exists between the three pairs of alignments?

(b) Based on your results (which are too few to provide a comprehensive study), do you believe there is evidence that human adaptation is occurring in H_5N_1 viruses that might merit concern about human-to-human transmission in the near future?

(c) Statistics: What were the numbers of ***Lengths***, ***Similarities***, ***Gaps*** and ***Scores*** for each of your alignment tasks?

Required Deliverables

- Markdown file: **report.md**; Responses to PArt 1 and Part 2 Question-in-Blue from above.
- Markdown file: **calc.md**; Your calculations for the Needleman-Wunsch algorithm from above.

Grading

The grade that you receive for this lab assignment will be based on the following:

- 20% Your calculations in the file; `writing calc.md`.
- 70% Your work for Part1 and Part 2: Your responses to the above questions-in-blue (Part 1) and your analysis of the influenza sequences from Part 2 in the file; `writing report.md`.
- 10% Complete GitHub Actions CI build-pass corresponding to all the GatorGrader checks passing.

Please see the Technical Leaders or the instructor if you have questions about the assignment submission.

Please see the instructor if you have questions about assignment or its submission.

References

[1] D. Van Riel, V. J. Munster, E. De Wit, G. F. Rimmelzwaan, R. A. Fouchier, A. D. Osterhaus, and T. Kuiken, "H5n1 virus attachment to lower respiratory tract," *Science*, vol. 312, no. 5772, pp. 399–399, 2006.