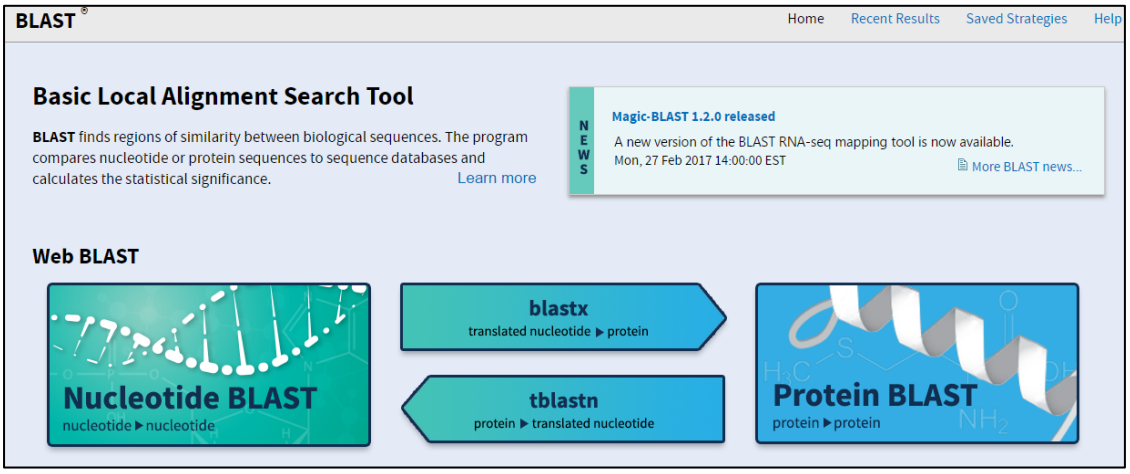
Comparing Influenza Protein Sequences Using BLAST (Basic Local Alignment Search Tool)

Background:

One tool in the bioinformatics is called **BLAST – Basic Local Alignment Search Tool.** BLAST can be used to look at differences in the sequences of two or more proteins (**Protein BLAST**) or nucleic acid molecules (**Nucleotide BLAST**). BLAST, begins with a **query sequence**, the sequence BLAST, begins with a **query sequence**, the sequence you are going to use to relate to other sequences. You can also compare a single query sequence to a collection of sequences in a database at the National Center for Biotechnology Information (NCBI). All of the results from your BLAST are called **subject sequences**. You use this when you are trying to find a sequence, such as when doing DNA barcoding or finding contamination in a sequencing reaction. The results of BLAST are in the form of an **alignment** to find regions that are the same to your first sequence and regions that are different than your first sequence.



In this experiment, you will do a **Protein BLAST**, looking at the following N1 influenza protein sequences:

- A/Brisbane/59/2007(H1N1), which was the influenza vaccine strain from 2008-2010;
- A/California/07/2009(H1N1), which was the influenza vaccine strain from 2011-2016;
- A/Michigan/45/2015 (H1N1), which has been the influenza vaccine strain since 2017; and
- The N1 sequences from each patient that was influenza-positive by PCR

You will use the current influenza vaccine strain, A/Michigan/45/2015 (H1N1), as your **query** (or reference) **sequence** and all of the other sequences will be your **subject sequences**.

Research Questions:

1. Based on your BLAST results, does the influenza strain that infected your patients more closely match the current influenza vaccine or a previous influenza vaccine?
2. Based on your BLAST results, ELISA and PCR data from your group and the data from your classmates, do you believe that your patients would have gotten sick from influenza if they had received the current influenza vaccine?

Patient Data:

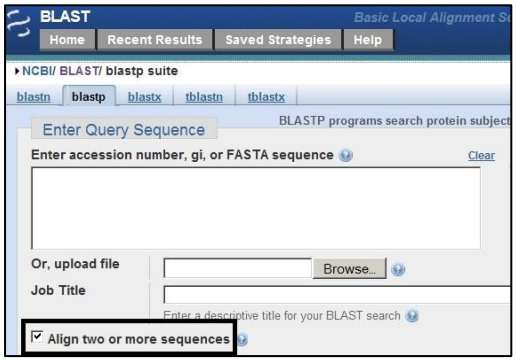
Which patient(s) in your group were infected with influenza? In other words, which patient(s) were positive for influenza by PCR?

Patient ID: _____ Patient ID: _____

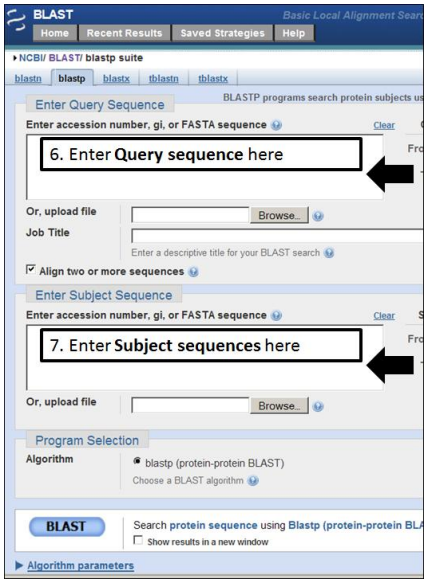
Patient ID: _____ Patient ID: _____

Procedure:

1. Download the "**N1 Protein Sequences**" file provided by your teacher.
2. Go to BLAST, either using your search engine or the URL: https://blast.ncbi.nlm.nih.gov/Blast.cgi
3. Because we are comparing protein sequences, click "**Protein BLAST**" on the BLAST homepage. (See the screenshot on the previous page.)
4. Note: The default view in BLAST has only one sequence box, in which to enter your Query Sequence. This is because BLAST is most frequently used to compare a query sequence to all of the sequences in the NCBI databases (i.e., sequence identification).



5. To make the "**Subject Sequences**" box display, click on the box "**Align two or more sequences**."
6. Copy and paste your query sequence, **A/Michigan/45/2015 (H1N1)-2017-Present Vaccine**, into the "**Enter Query Sequence**" box. Be sure to include the top line of text in the sequence, ">A/Michigan/45/2015 (H1N1)-2017-Present Vaccine" when you paste the text into the "**Enter Query Sequence**" text box.



7. Copy and paste ***ALL*** of the following subject sequences into the "**Enter Subject Sequence**" box, making sure to include the "<" caret and the sequence names when you copy and paste. You may have to copy and paste multiple times to copy and paste all of the sequences into the "Enter Subject Sequences" box.
    a. A/California/07/2009(H1N1)-2011-2016 Vaccine
    b. A/Brisbane/59/2007(H1N1)-2008-2010 Vaccine

c. The sequences for each of the patients that you analyzed that tested positive for influenza by PCR.

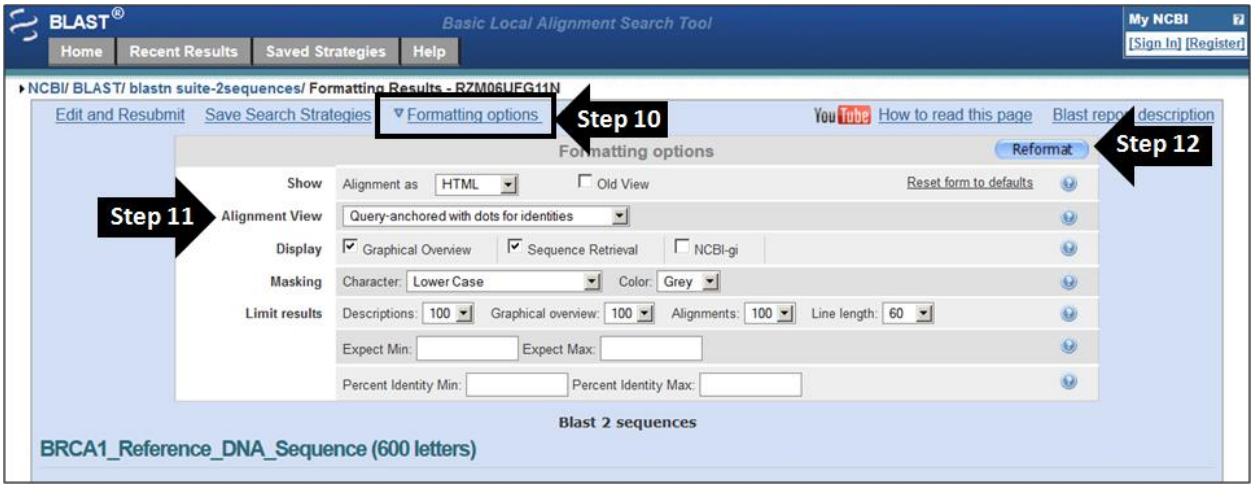
8. Click "**BLAST**."

9. Note: When your results appear, they will be in the ***pairwise comparison*** format. In other words, each subject sequence is aligned to the query sequence individually. This is a useful format to use when identifying an individual sequence (i.e., searching the NCBI databases with your query sequence), but to compare multiple sequences to one another, a ***multiple sequence alignment*** format is more helpful.

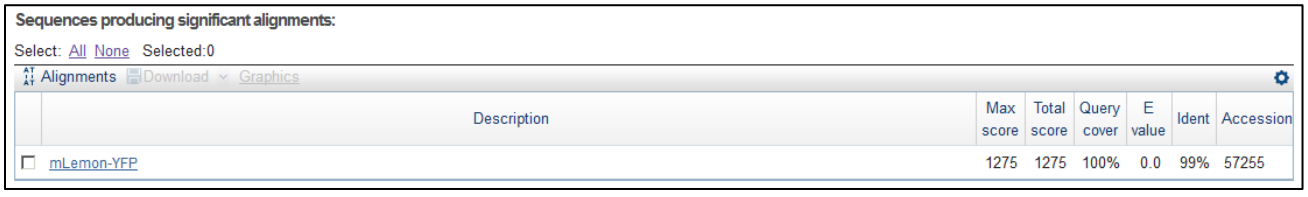
10. Scroll to the top of the window and click on the "**Formating options**" dropdown menu.

11. Click on "**Alignment View**" and select "**Query-anchored with dots for identities**."

12. Click "**Reformat**."



13. Scroll back down the page to see a data table with your BLAST scores, as well as the multiple sequence alignment of all of your neuraminidase protein sequences. ***BLAST scores*** help us ***quantify*** the BLAST results. In the example below, a sequence named "mLemon-YFP" has been compared to a query sequence and has a ***Max Score*** and ***Total Score*** of 1275, a ***Query Coverage*** of 100% and a ***Percent Identity*** of 99%. The ***e-value*** is 0.0.



The ***Max score*** and ***Total score*** are related to the length of the sequences compared and how well they match. Generally, the higher the score, the better the two sequences match each other. These scores are particularly helpful when comparing multiple sequences to each other.

Query coverage (abbreviated "Query cover") & ***Percent identity*** (abbreviated "Ident") quantify how much of the sequences match each other (***Query Coverage***), and how well they match (***Percent***

Identity). For example, a small portion of the sequences (25% query coverage) may match well (100% identity). Alternatively, 100% of the sequences may line up with one another (or "align"), but might share only 50% of the same nucleotides or amino acids (50% identity).

The ***e-value*** or ***expect value*** is an indication of how likely these results are based purely on chance. For example, if you have performed statistical analyses like chi-squares, you should be familiar with a ***p-value***. As with p-values, a low e-value mean you can be more confident in your results because they are not due to chance.

EXAMPLES:

30% Query Coverage, 100% Identity

3/10 bases (30%) match perfectly (100%)

ATG**GAT**ACGT

TGA**GAT**GATC

100% Query Coverage, 70% Identity

All 10 bases (100%) align, but only 70% match

AT**G**C**C**G**ACAG**

AG**G**G**C**A**ACAG**

The formatting option "**Query-anchored with dots for identities**" BLAST alignment would look like this, with a dot in the subject sequence at each position where it matches the query sequence:

ATGGATACGT

TGA•••GATC

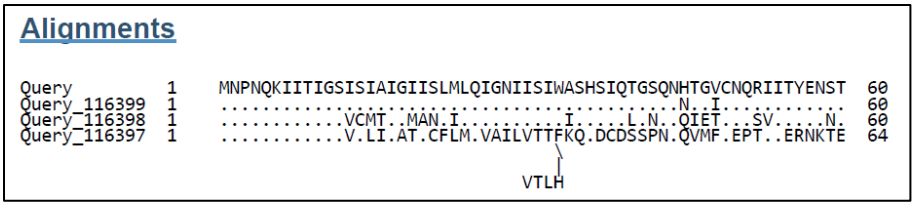
ATGCCGATTG

•G•G•A••••

14. Record in the data table below the patient ID for each sample you analyzed (for example, "E11"), the query coverage, the percent identity and the e-values. Remember that you are comparing these sequences to the current vaccine strain of influenza.

Patient ID	Query Coverage	Percent Identity	E-Value
A/California/07/2009(H1N1), the 2011-2016 Vaccine Strain			
A/Brisbane/59/2007(H1N1) the 2008-2010 Vaccine Strain			

15. Scroll through your multiple sequence alignment to get a general, visual idea of how well the different protein sequences match. Remember, the subject sequence is identical to the query sequence where there are dots. You may also see something like the image below, where one of the protein sequences has additional amino acids (insertions of V, T, L and H) relative to the query sequence.



16. Which patients that you studied, if any, received an influenza vaccine prior to 2017 (in other words, an influenza vaccine different that the current vaccine)? List their patient ID(s) below and note which vaccine they received. This information is on your Patient Information Card.

Patient ID	Sympotoms	Vaccination
A11	mild	old vaccine
A12	mild	old vaccine
A17	mild	old vaccine
B11	mild	old vaccine
B12	mild	old vaccine
B17	mild	old vaccine
C11	mild	old vaccine
C12	mild	old vaccine
C17	mild	old vaccine
D11	mild	old vaccine
D12	mild	old vaccine
D17	mild	old vaccine

17. How closely does the influenza vaccine that your patient(s) received match the current vaccine strain? Include in your answer ***percent identity***, ***query coverage*** and ***e-value***. Remember that an e-value is essentially a ***p-value***, or probability of "chance."

18. How well do the protein sequences from the patients that you studied match the current vaccine strain? Include in your answer ***percent identity***, ***query coverage*** and ***e-value***. Remember that an e-value is essentially a ***p-value***, or probability of "chance," as we saw with the coin toss and the ELISA data.

19. Based on your BLAST results, ELISA and PCR data from your group and the data from your classmates, do you believe that your patients would have gotten sick from influenza if they had received the current influenza vaccine?

20. What do you notice about the sequences from the infected patients? How well do they match the previous vaccine strains? How well do they match the current vaccine strain?

21. Based on your answer to the previous question, what recommendations do you have regarding influenza vaccination?