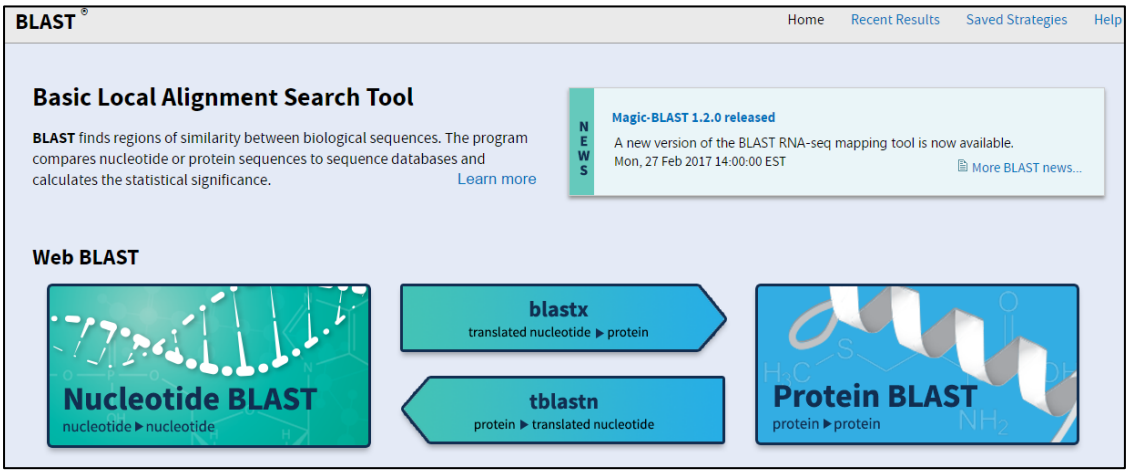
Comparing Influenza Protein Sequences Using BLAST (Basic Local Alignment Search Tool)

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**. This is also called the **reference sequence** when performing medical tests such as *BRCA1* genetic testing. The Query Sequence is the sequence that you are going to use to relate to or compare 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 or identify a sequence, such as when doing DNA barcoding or finding contamination in a DNA sequencing reaction. The results of BLAST are in the form of an **alignment** to find regions that are the same between your Query Sequence and your Subject Sequence or sequences.



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 patient or patients more closely match the current influenza vaccine or a previous influenza vaccine?
2. After reviewing your BLAST results, ELISA and PCR data from your group, and the data from your classmates, answer these questions: (A) If your patients did not receive the current influenza vaccine, do you believe that your patient or patients would have gotten sick or did get sick from

influenza if they had received the current influenza vaccine? (B) If your patient or patients did receive the current influenza vaccine, did it prevent them from getting sick?

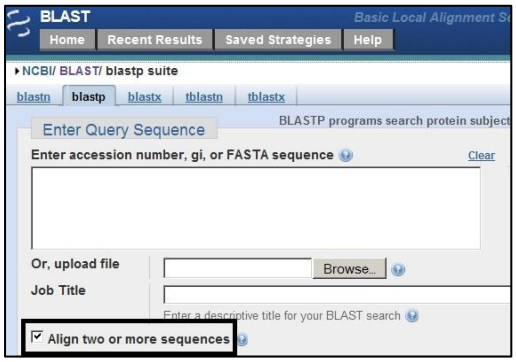
Patient Data:

List your patient ID's below and note which patient or patients in your group were infected with influenza. In other words, which patient(s) were positive for influenza by PCR? Which patient or patients in your group had anti-influenza antibodies? This is based on your ELISA results.

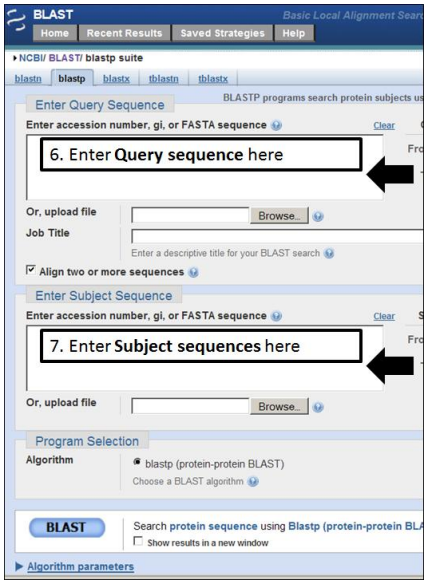
Patient ID	Influenza-Positive by PCR?	Anti-Influenza Antibody Positive by ELISA?

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, shown in the picture on the previous page.
4. Note: The default view in BLAST has only one sequence box, in which you 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 for 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**, from the sequence Word document 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 because your patients' sequences may be located further down in the Word document data set.

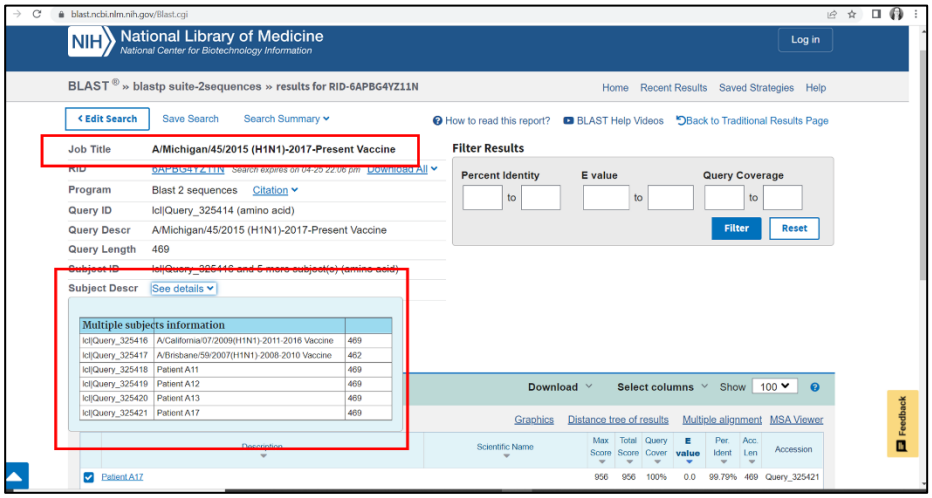
   a. A/California/07/2009(H1N1)-2011-2016 Vaccine

   b. A/Brisbane/59/2007(H1N1)-2008-2010 Vaccine

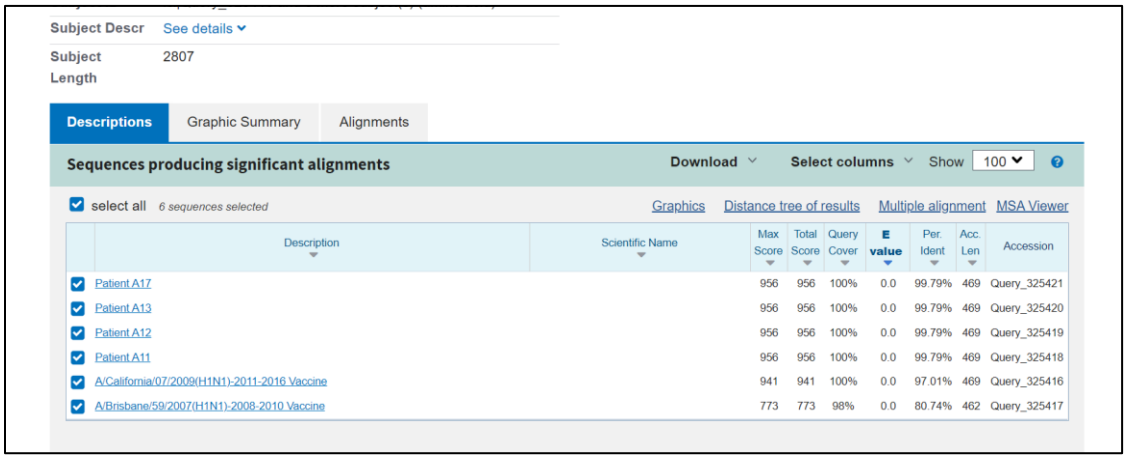
   c. The sequences for each of the patients that you analyzed that tested positive for influenza by PCR.

8. Click "**BLAST**."

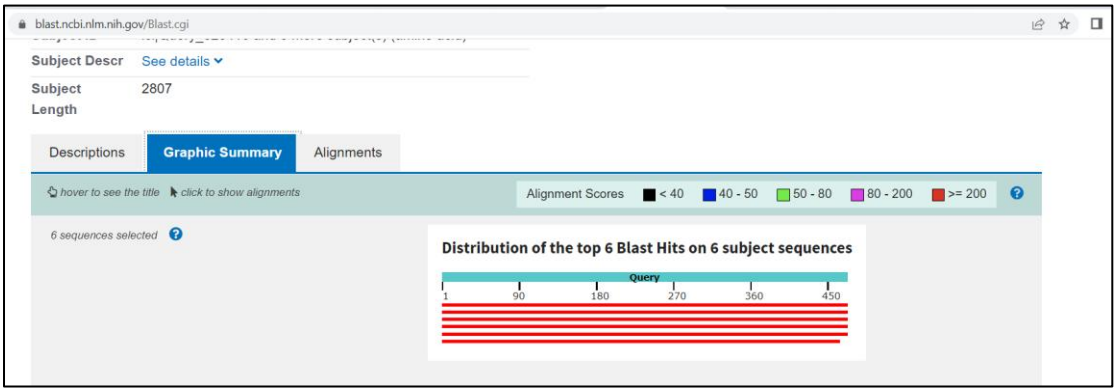
9. When your results appear, you will see a summary of your work, called the "Descriptions" tab in the lower, menu, on the left side (the first tab of that menu. The sequence that is your Query Sequence is listed as your "Job Title" in the top, left list. There are also additional details that you don't need to worry about for this analysis, but one useful thing in that list is a drop-down option next to the term "Subject Descr" that you can click to see the names of all of the Subject Sequences that you included in your BLAST analysis [just in case you want to make sure that you included everything that you need to include].

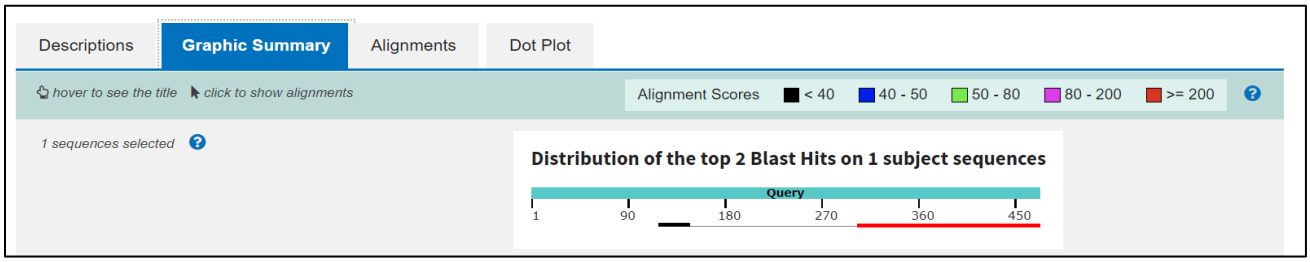


10. If you scroll down further, you will see a summary of the **BLAST Scores** for each of your sequences**.** We'll discuss BLAST scores later, but in general these are ways to *quantitatively* measure how "good" the match is between your Query Sequence and each Subject Sequence. See the example below.

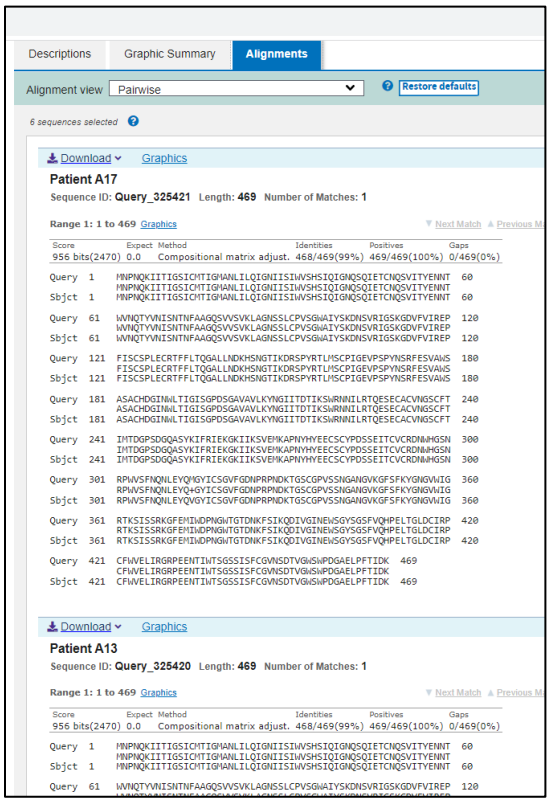


11. Click on the second tab, "Graphic Summary." You will see an image like the one below. The color refers to *how many nucleotides or amino acids of the Query Sequence and each Subject Sequence match* (i.e., are the same). In the top image, we see **red lines** – this means that each of these sequences match one another by ≥ 200 nucleotides or amino acids. In the second image below, we see that ≥ 200 amino acids of the Query Sequence and Subject Sequence match near the C-terminus [seen in **red**], and ≤ 40 only of the amino acids match near the N-terminus, as seen in **black**. We also get to see where these matches are, or are not





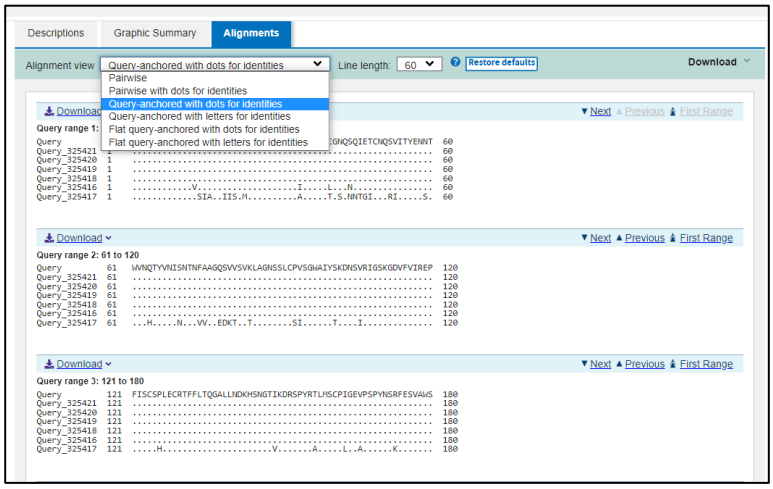
12. Next, click the "Alignments" tab. In this case, you will see which individual nucleotides (or amino acids) match between the Query Sequence and the Subject Sequence of Sequences. See the image below.



13. The Default settings in BLAST show you each alignment between the Query Sequence and each Subject Sequence, showing all of the one-letter amino acid abbreviations. Click on the Scroll to the top of the window and click on the "**Formating options**" dropdown menu.

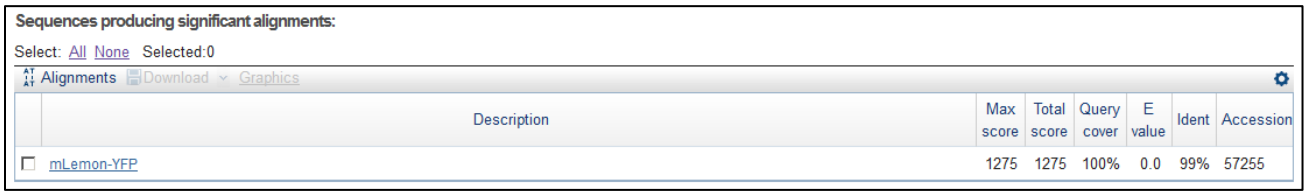
14. Some people find this format difficult to analyze: (a)Where exactly are the differences? (b) Are there any differences between the various Subject Sequences (if you included more than one Subject Sequence in your analysis, as we did here)?

15. There is a drop-down menu below the "Alignments" tab. From this menu, click "Query-anchored with dots for identitites." This will re-format your results, as shown below. (a) Query-anchored means that everything is being compared to the Query Sequence, which you see on the top line of each portion of the alignment. (b) If the Query Sequence matches any of the Subject Sequences, there wil be dot at that position. (c) If the Query Sequence and the Subject Sequence do not match, BLAST will show you the single letter abbreviation for the nucleotide or amino acid in the Subject Sequence that is different from the nucleotide or amino acid in the Query Sequence. Note: If the sequences are longer than 60 nucleotides or amino acids, the alignment will be displayed as multiple 'mini-alignments' – 1-60, 61-120, 121-180, etc. **Take a moment to scroll through your multiple sequence alignment to get a visual idea of how much the vaccine sequences and your patient sequence(s) vary from one another. Make a note of what you see.**



16. Click again on the "Descriptions" tab from the menu. You just looked at your sequences to get a general idea of how much they vary from one another, but how can we quantify this variation? That is why we now return to **BLAST scores.**

17. ***BLAST scores*** help us ***quantify*** the BLAST results. In the example below, a sequence named yellow fluorescent protein, "mLemon-YFP" has been compared to a Query Sequence (green fluorescent protein, or "GFP"). This comparison 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.


Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download Graphics

	Description	Max score	Total score	Query cover	E value	Ident	Accession
☐	mLemon-YFP	1275	1275	100%	0.0	99%	57255

The ***Max score*** and ***Total score*** are related to the length of the sequences that are compared to one another and how well they match one another. 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) within that matching region.

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. Imagine if you just grabbed two sequences and alignment them: how likely is it that they would match each other?

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:

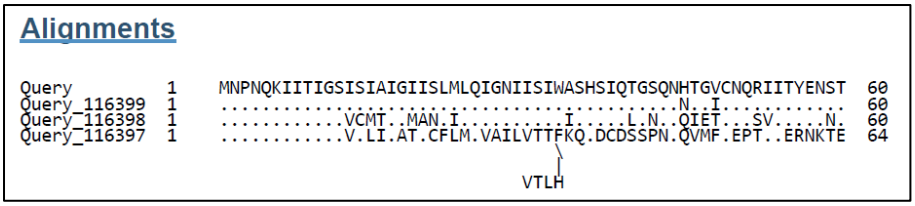
ATGGATACGT
TGA•••GATC

ATGCCGATTG
•G•G•A••••

18. 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 Query Sequence, which is the current vaccine strain of influenza, A/Michigan/45/2015 (H1N1).

Sample or 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			

19. Return to the "Alignments" tab. If BLAST did not save your previous selection, use the drop-down menu again to select "Query-anchored with dots for identities." 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.



20. Which patients that you studied, if any, received an influenza vaccine? List their patient ID(s) below and note which vaccine they received. This information is on your Patient Information Card.

21. The patients that received the current influenza vaccine did not have influenza sequence data available. Why do you think that is?

22. 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?

23. Based on your BLAST results, how well do you think that influenza vaccination status predicts the symptoms of the patients that you studied? For your BLAST results, 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."

24. Based on your BLAST results, ELISA and PCR data from your group and the data from your classmates, how well do you think that influenza vaccination status predicts the symptoms of the patients that you studied? For your BLAST results, 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."

25. Return to the "Descriptions" tab and select "Distance tree of results." This will show you a **cladogram**, similar to a **phylogenetic** tree, which is a visual representation of how similar or different sequences are. Each sequence is shown on an individual horizontal line, or **branch** of the tree. The length of the branch represents how similar or different the sequences are from one another. Take a screen shot of your tree, or draw a picture of it below.

26. Does the image of the Distance Tree of Results inform your conclusions about how similar or different the vaccine strains are from one another, and how similar or different they are from your patience sequence(s)? What does this mean? In other words, does it support or confirm your conclusions from the previous question?

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