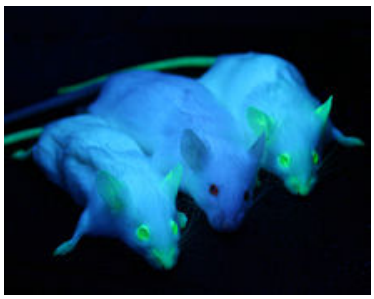
Comparing Sequences of Fluorescent Proteins Using BLAST (Basic Local Alignment Search Tool)



Mice expressing GFP under UV light (left & right), compared to normal mouse (center). Source: Wikipedia.

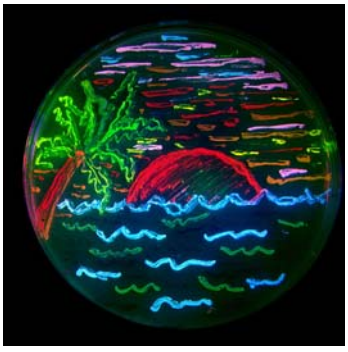
Researcher Background:

Fluorescent proteins have become a valuable tool in recent years among scientists in many different fields of biology. Often, these glowing proteins are linked to other proteins to confirm protein expression, identify where specific proteins exist in the cell, and to track cell movement.

Green fluorescent protein (GFP) was isolated from the jellyfish *Aequorea victoria* and is comprised of 238 amino acids that form 11 beta(β)-sheets (1, 2). These β-sheets fold into a barrel-shaped protein called a β-can (due to its resemblance to a soup can (3)). GFP glows bright green when exposed to light in the blue or ultraviolet range.

Many things in biology are hard to see without the addition of stains or dyes. GFP quickly revolutionized many fields of biology as scientists realized that they could use GFP to see things like proteins and cells in a way that had never been possible before.

However, scientists quickly realized that having a **rainbow of fluorescent proteins** would dramatically increase their usefulness. Some scientists used **intentional mutation** of the *gfp* gene to change its fluorescent properties, while others **set out across the globe** to try to discover other animals that make fluorescent proteins similar to GFP. As the image to the right shows, some of these scientists have been successful!



The diversity of genetic mutations is illustrated by this San Diego beach scene drawn with living bacteria expressing 8 different colors of fluorescent proteins. Source: Wikipedia.

One group of fluorescent proteins is referred to as the "mFruits" due to the names given to these fluorescent proteins, such as:

- mBlueberry (Blue Fluorescent Protein, or BFP)
- mLemon (Citrine/Yellow Fluorescent Protein, or YFP)
- mGrape1 (Purple Fluorescent Protein, or PFP)
- and many others, all with similarly 'fruity' names…

Some fluorescent proteins, like the mFruits, are **monomeric** -- they are composed of only one polypeptide chain.

Research Questions: The cloning and protein purification experiments you have been conducting in the laboratory involve mTomato, also called **red fluorescent protein (RFP).**

(1) Is red fluorescent protein (RFP) related to its famous cousin, GFP, or is it from a different source entirely?
(2) What other fluorescent proteins, if any, are closely related to GFP and RFP?

Updated September 10, 2014 by Dina Kovarik, Digital World Biology

INSTRUCTIONS

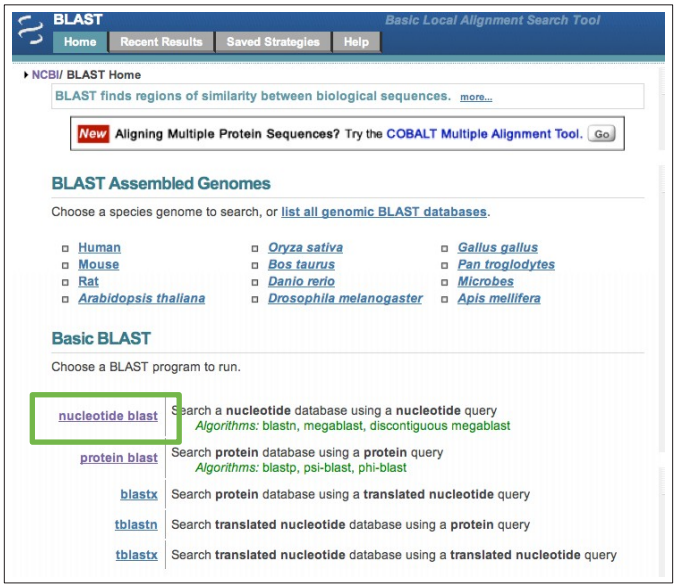
Aim: To understand how to analyze DNA and protein sequences using the bioinformatics tool BLAST (Basic Local Alignment Search Tool). These analyses will help you better understand the origin and diversity of fluorescent proteins used in biological research.

PART I: Aligning DNA Sequences

1. Access the file "**DNA Sequences of Fluorescent Proteins**" from the Digital World Biology website: http://digitalworldbiology.com/dwb/abe-resources

2. Go to the BLAST website at the National Center for Biotechnology Information (NCBI): http://blast.ncbi.nlm.nih.gov/Blast.cgi

3. Select "**nucleotide blast**," as shown in the image to the right. We will use a nucleotide BLAST (or "**blastn**") because we will be comparing a DNA sequence (sequence of nucleotides) to a DNA sequence (sequence of nucleotides). Note that there are options for other types of BLASTs, such as for comparing protein sequences to protein sequences.



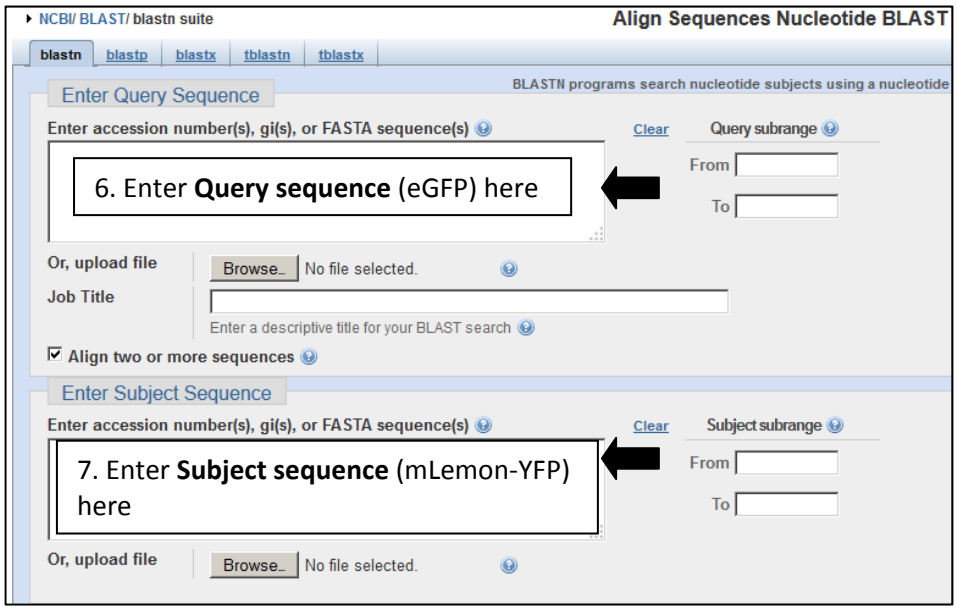
4. From the nucleotide blast page (see below), click the box to choose the option to "**Align two or more sequences**."

5. A second text box will appear.

6. Copy and paste the DNA sequence "**eGFP**" into the top box (the "**Query Sequence Box**"). Be sure to include the caret (">") symbol and the sequence name.

A **query sequence** is the sequence to which other sequences are compared when performing a BLAST alignment. In this experiment, we are comparing sequences to GFP, so GFP is our query sequence.

NOTE: The "**e**" in "**eGFP**" stands for "enhanced." This GFP glows brighter than other forms of GFP.

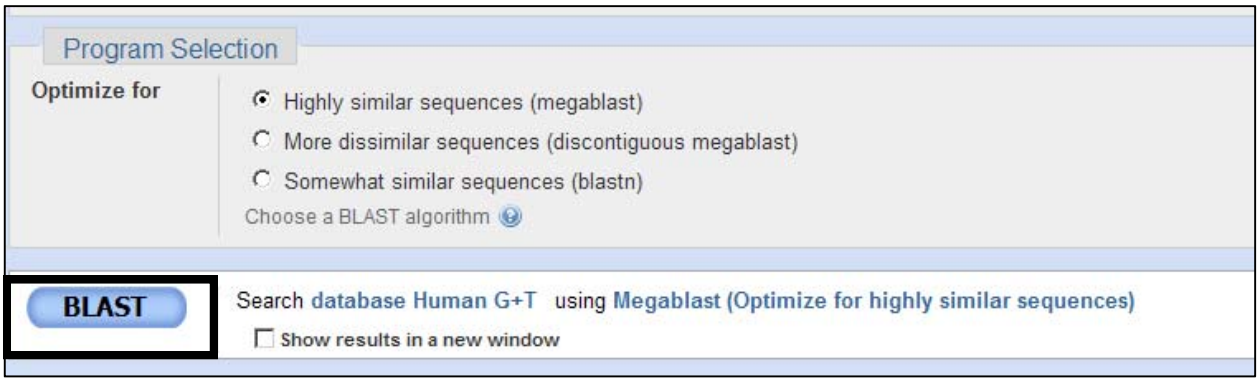


7. Copy and paste the DNA sequence "**mLemon-YFP**" into the bottom box (the "Subject Sequence Box"). Be sure to include the caret (">") symbol and the sequence name.

A **subject sequence** is the sequence being compared when performing a BLAST alignment. In this experiment, the mLemon-YFP gene is our subject sequence. YFP is sometimes called "Citrine," which is a shade of yellow and a semi-precious gemstone.

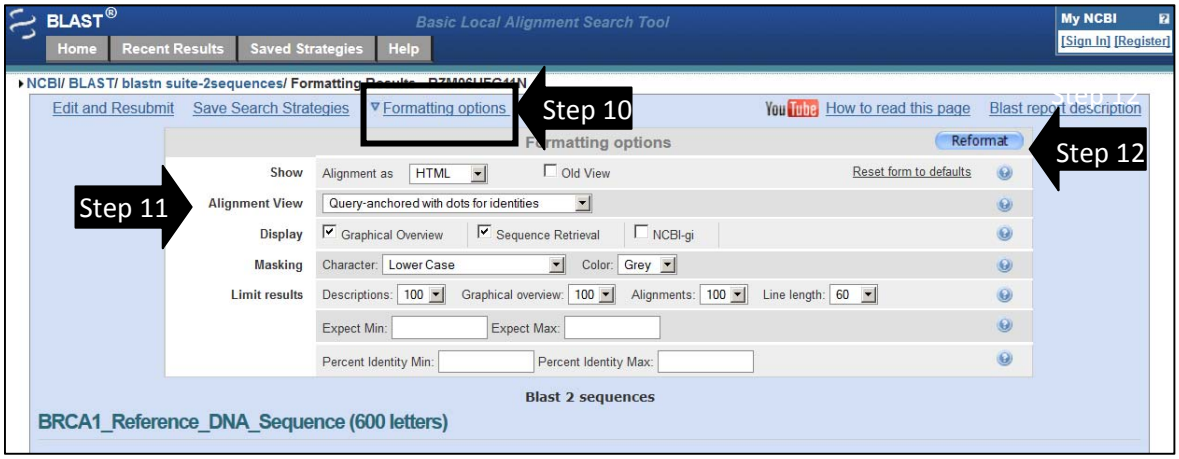
NOTE: "**YFP**" stands for **"yellow fluorescent protein**" and "**m**" stands for "**monomeric**."

8. Click "**BLAST**."



9. When your search is complete, you will see a window with the BLAST results, showing an alignment of the two DNA sequences you entered above.

10. Click the "**Formatting Options**" link located near the top of the page.

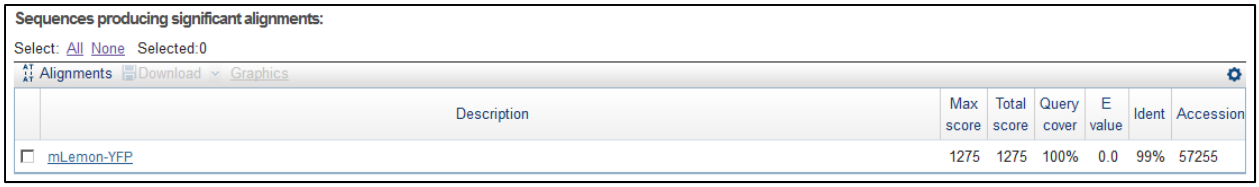


11. Find the "**Alignment View**" and use the drop-down menu to choose "**Query-anchored with dots for identities**." The query is the eGFP sequence. The query-anchored view shows the eGFP sequence at the top with the subject sequence (mLemon-YFP) aligned below. Dots are used to show nucleotides that are identical and letters are used to show nucleotides that differ.

12. Click the "**Reformat**" button.

13. Scroll down the page to see if there are positions where the eGFP and mLemon-YFP sequences differ. In other words, look for a place where there is a letter instead of a dot, showing that there's been a change in the nucleotide at that position. Note the numbers at the ends of the lines refer to the position of the nucleotide.

14. **BLAST scores** help us ***quantify*** the BLAST results.



The **Max score and Total score** are related to the length of the sequences compared. 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 & **Percent identity** quantify how much of the sequences match, and how well they match. For example, a small portion of the sequences (for example, *25% query coverage*) may match well (*100% identity*). Alternatively, 100% of the sequences may align to one another, but might share only 50% of the same nucleotides (50% identity).

The **e-value** or **expect value** is an indication of how likely these results are based purely on chance. Similar to a p-value, low e-values mean you can be more confident in your results.

EXAMPLES:

30% Query Coverage, 100% Identity
3/10 bases (30%) match perfectly (100%)

ATG**GAT**ACGT
TGA**GAT**GATC

100% Query Coverage, 50% Identity
All 10 bases (100%) align, but only 50% match

AT**G**C**C**G**A**TT**G**
AG**G**G**C**A**A**CA**G**

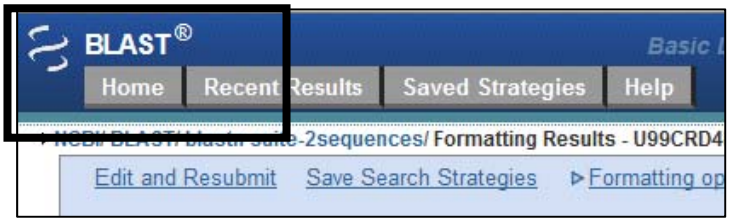
The "Query-anchored with dots for identities" BLAST alignment would look like this, with a dot in the subject sequence at each position where it matches the query sequence:

ATGGATACGT
TGA•••GATC

ATGCCGATTG
•G•G•A•CA•

15. **Based on your BLAST results, do eGFP and mLemon-YFP appear to be closely related to one another? Why or why not? Your answer should include the "Query Coverage" and "Max Identity" scores obtained from your BLAST results.**

16. Return to the BLAST homepage. You can re-enter the URL in step 2, or you can click the "**BLAST**" logo or the "**Home**" button in the upper left corner of the screen.



17. Perform another nucleotide BLAST alignment as explained above, **Steps 3-14**. Use ">mTomato-RFP" as the **query sequence** and "mGrape-PFP" as the **subject sequence**.

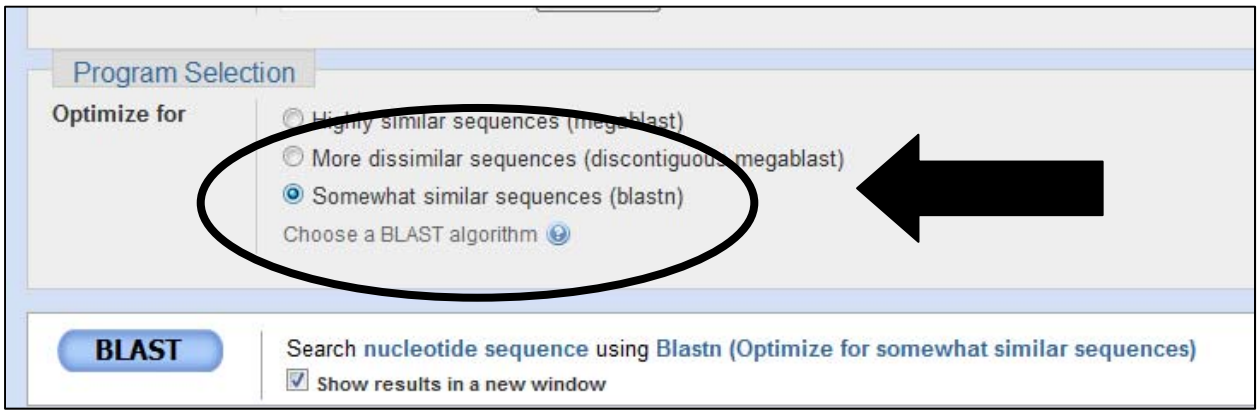
RFP = Red Fluorescent Protein

PFP =Purple Fluorescent Protein

18. **Do these two sequences appear to be closely related to one another? Why or why not? Your answer should include the "Query Coverage" and "Max Identity" scores obtained from your BLAST alignment.**

19. Perform a third nucleotide BLAST alignment as explained above, **Steps 3-14**. Use "eGFP" as the **query sequence** and "mTomato-RFP" as the **subject sequence**.

SPECIAL NOTE: ***BEFORE*** clicking the blue "**BLAST**" button, choose "**Somewhat similar sequences (blastn)**" from the "**Program Selection**" menu.



20. **Do these two sequences appear to be closely related to one another? Why or why not? Your answer should include the "Query Coverage" and "Max Identity" scores obtained from your BLAST alignment. Compare your answer to this question with your answers to Question 15 [eGFP and mLemon YFP comparison] and Question 18 [pTomato-RFP and mGrape-PFP comparison].**

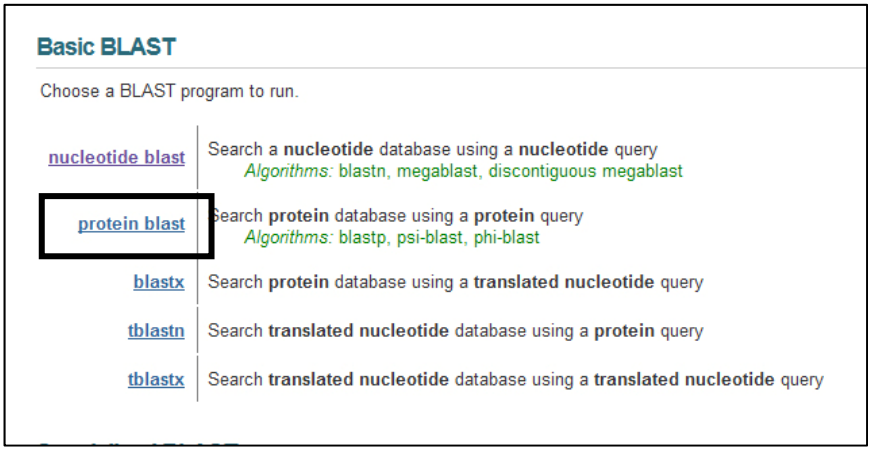
PART II: Aligning Protein Sequences

Scientists often find it useful to compare both DNA and protein sequences. Multiple codons can code for a single amino acid, and different amino acids can have similar chemical properties. Therefore, we will use **Protein BLAST** to compare the protein sequences of different fluorescent proteins, and compare these results with our nucleotide BLAST results.

21. Access the file "Protein Sequences of Fluorescent Proteins" from the Digital World Biology website: http://digitalworldbiology.com/dwb/abe-resources

22. Go to the BLAST website at the National Center for Biotechnology Information (NCBI): http://blast.ncbi.nlm.nih.gov/Blast.cgi [or just return to the BLAST homepage]

23. Select "**protein blast**," as shown in the image to the right. We will use a protein BLAST (or "**blastp**") because we will be comparing a protein sequence (sequence of amino acids) to a protein sequence (sequence of amino acids).



24. From the protein blast page, click the box to choose the option to "**Align two or more sequences**."

25. A second text box will appear.

26. Copy and paste the protein sequence "**eGFP**" into the "**Query Sequence Box**" (top box). Be sure to include the caret (">") symbol and the sequence name.

27. Copy and paste the protein sequence "**mLemon-YFP**" into the "**Subject Sequence Box**" (bottom box). Be sure to include the caret (">") symbol and the sequence name.

28. Click "**BLAST**."

29. When the BLAST results page appears, click the "**Formatting Options**" link located near the top of the page.

30. Find the "**Alignment View**" and use the drop-down menu to choose "**Query-anchored with dots for identities**."

31. Click the "**Reformat**" button.

32. Scroll down the page to see if there are positions where the eGFP and mLemon-YFP protein sequences differ. In other words, look for a place where there is a letter instead of a dot, showing that there's been a change in the amino acid at that position. Note the numbers at the ends of the lines refer to the positions of the amino acids.

33. **Do eGFP and mLemon-YFP appear to be closely related to one another? Why or why not? Your answer should include the "Query Coverage" and "Max Identity" scores obtained from your BLAST alignment.**

34. **Are your results similar to your nucleotide comparison of these two sequences in Question 15? Explain how they are or are not similar.**

35. Perform another protein BLAST alignment as explained above, **Steps 22-32**. Use ">mTomato-RFP" as the **query sequence** and "mGrape1-PFP" as the **subject sequence**.

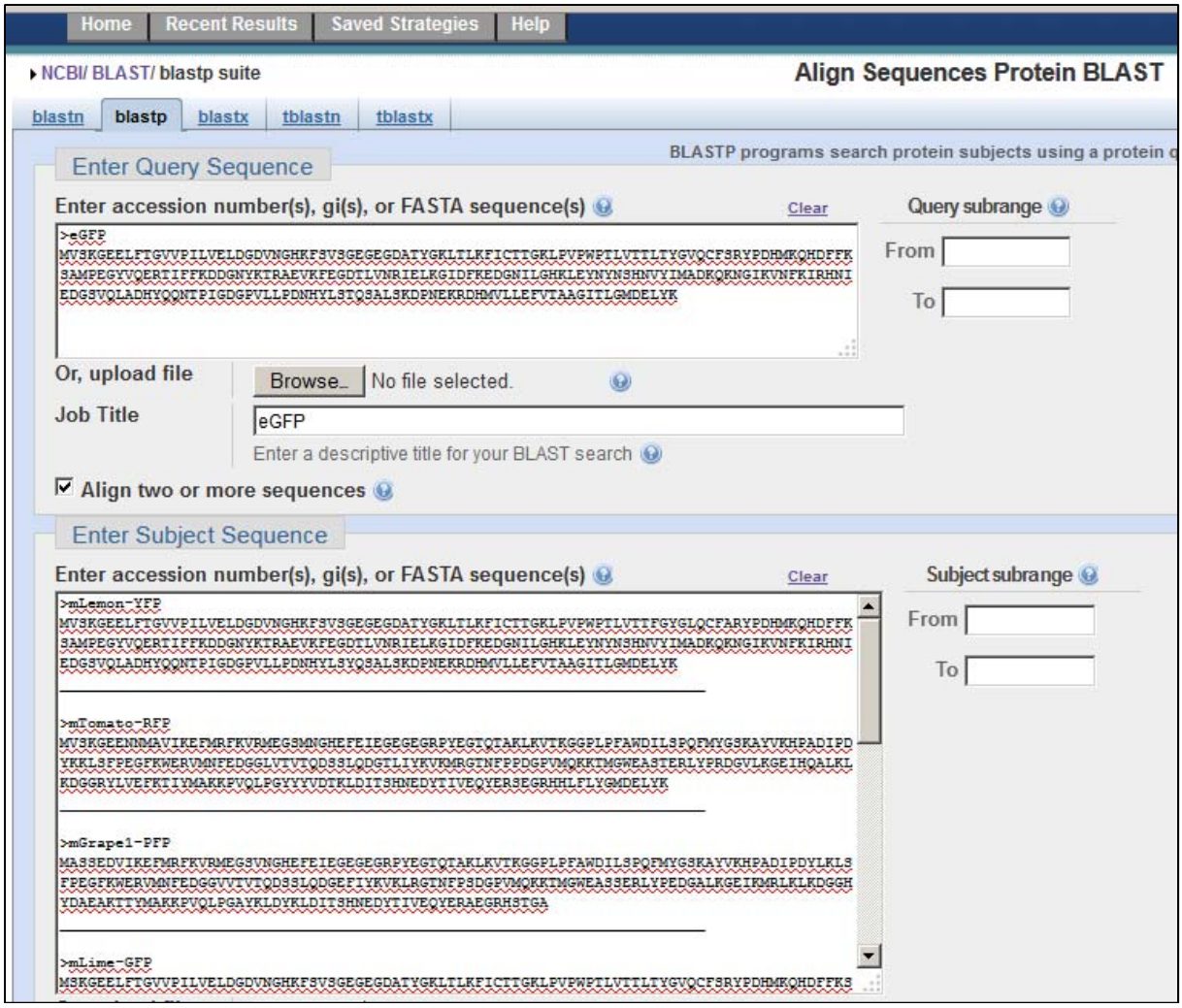
36. **Do mTomato-RFP and mGrape-PFP appear to be closely related to one another? Why or why not? Your answer should include the "Query Coverage" and "Max Identity" scores obtained from your BLAST alignment.**

37. **Are your results similar to your nucleotide comparison of these two sequences in Question 18? Explain how they are or are not similar.**

38. You can compare MANY sequences at one time with BLAST. Perform a final protein BLAST alignment will ALL of the fluorescent protein sequences. Use "eGFP" as the **query sequence** and all of the remaining protein sequences as the **subject sequences**:

    a. mLime-GFP [a light green fluorescent protein]
    b. mBlueberry-BFP [Blue Fluorescent Protein]
    c. mTangerine1.5 [an orange fluorescent protein]
    d. mCherry-RFP [a Red Fluorescent Protein]
    e. mOrange-OFP [an Orange Fluorescent Protein]

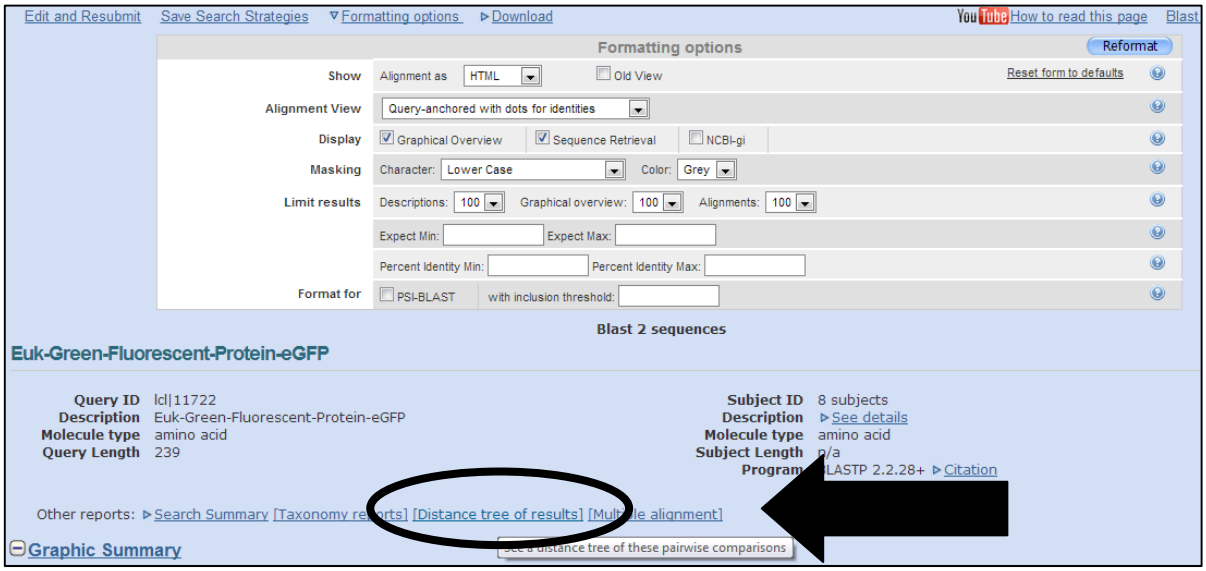
Simply paste ALL of the subject sequences in the bottom **subject sequence box** (see example on the next page)**.** Be sure to include the caret (">") and sequence names for each protein sequence. Don't worry about lines and line breaks – BLAST will ignore these.



39. **Based on the Query coverage and Max identity scores you obtained, which fluorescent proteins appear to be most closely related to eGFP? Fill in the table below to support your answer.**

Protein	Query Coverage	Max Identity (%)	Closely related to eGFP?
mLemon-YFP			
mTomato-RFP			
mGrape1-PFP			
mLime-GFP			
mBlueberry-BFP			
mTangerine1.5			
mCherry-RFP			
mOrange-OFP			

40. Another way to visualize your results is by using a **distance tree**. Similar to a phylogenetic tree, a distance tree is a graphical representation of relationships – in this case, which sequences are most similar to one another. BLAST can perform this analysis for you using the comparisons that you have already made. Click "**Distance Tree of Results**" to view your tree.



41. **Draw your tree in the space below.**

42. **Does the phylogenetic tree support the conclusions that you made in Questions 39 and 40 about the relatedness of various fluorescent proteins to eGFP? Which fluorescent proteins appear to be most closely related to mTomato-RFP? Explain your answer.**

References:

1. Prendergast F, Mann K (1978). "Chemical and physical properties of aequorin and the green fluorescent protein isolated from Aequorea forskålea". *Biochemistry* 17 (17): 3448–53. doi:10.1021/bi00610a004. PMID 28749.

2. Tsien R (1998). "The green fluorescent protein" (PDF). *Annu Rev Biochem* 67: 509–44. doi:10.1146/annurev.biochem.67.1.509. PMID 9759496

3. Yang, F., Moss, L.G., and Phillips, G.N.J. (1996) The molecular structure of green fluorescent protein, *Nat. Biotechnol.*, **14**, 1246-1251.

Additional Information about GFP and other Fluorescent Proteins:

1. Robert E. Campbell (2008). Fluorescent Proteins. Scholarpedia, 3(7):5410. http://www.scholarpedia.org/article/Fluorescent_proteins

2. Green fluorescent protein. Wikipedia. http://en.wikipedia.org/wiki/Green_fluorescent_protein

3. Marc Zimmer. (2011). Green Fluorescent Protein: A Molecular Microscope. http://www.photobiology.info/Zimmer.html

4. Zeiss. Fluorescent Protein Technology. http://zeiss-campus.magnet.fsu.edu/articles/probes/index.html

5. Chudakov et al. (2010). Fluorescent Proteins and Their Applications in Imaging Living Cells and Tissues. APS Physiological Reviews. 90: 1103-1163) DOI: 10.1152/physrev.00038.2009. http://physrev.physiology.org/content/90/3/1103