**Jesi Weed**

**Genetics Lab**

**GFP Bioinformatics**

Our goal for this exercise is to introduce you to some powerful tools and databases for studying biology. You will find nucleotide and protein sequences in online databases, be able to learn quite a bit of information about the sequences, and get to visualize the active site of the GFP protein.

We will perform a number of bioinformatics tasks relating to GFP:

* Retrieve the sequence of the pGLO plasmid and perform a restriction digest on the computer to see what our gels should look like
* Learn about the GFP protein and the gene that codes for it
* Compare the native GFP gene from jellyfish with the version on the pGLO plasmid we used in lab
* View the molecular structure of the GFP protein and highlight the active site

During this exercise, we will use several tools and databases, which are briefly described:

* GenBank is a database of nucleotide sequences. It contains over 100 million nucleotide sequences. Each sequence has an entry in the database that contains the nucleotide sequence and a variety of other information such as the scientific publications describing the sequence, the organism the sequence comes from, and features of the sequence itself such as coding regions, etc.
* NEBcutter is an online tool for performing restriction digests of DNA sequences. Simply enter the sequence, choose which enzymes you are interested in, and the program will tell you where the enzyme will cut and what size fragments are produced.
* UniProt is a database of proteins. It contains the amino acid sequences along with a wealth of information about the structure and function of each protein.
* BLAST is a program that finds similar regions within sequences. The program we will use is called bl2seq which is a special version that compares only the 2 sequences we will enter. This program is quite useful for quickly finding just the portions of 2 sequences that match up.
* The Protein Data Bank (PDB) is a database of protein structures. The structure of a protein determines its function and the structure typically the most difficult aspect of a protein to determine. PDB contains the structures of proteins and tools for analyzing and viewing the structures.

**Part I – A Computational Restriction Digest**

Go to NCBI GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/>) and find pGLO plasmid sequence using the accession number U62637. Enter this number in the search box at the top and click Go. On the results page, click on Nucleotide (near the top on the left side of the page). This page shows the GenBank record for the pGLO plasmid sequence (which for some reason is called pBAD-GFPuv here instead of pGLO, but it’s the same thing). Browse through the information on this page and answer the following questions:

1. How many nucleotides are in the sequence? There are 5371 nucleotides in the sequence.

2. What genes are on this plasmid and what is the function of each?

araC – Codes for a regulatory protein. The protein controls the expression of at least 6 genes

BLA- Confers resistance to ampicillin

Gfpuv = The GFP varient that causes the green fluorescent protein

Near the top of the page, click on the link that says FASTA. This displays the nucleotide sequence of the pGLO plasmid in FASTA format, a simple format consisting of an identifier line and the nucleotide sequence only. Select the entire sequence, including the identifier line that starts with “>gi|1490531|gb” and copy it.

Go to the NEBcutter website (<http://tools.neb.com/NEBcutter2/>) to perform a restriction digest of the pGLO plasmid online. Paste in the FASTA format sequence for the pGLO plasmid, click on the button that says the sequence is circular, and click submit. The results page shows the restriction sites present in the plasmid for many different restriction enzymes. We are interested in determining where the restriction sites are for the enzymes we actually used in the lab. The way to do this is to click on “Custom digest” near the bottom of the page on the left. Find the enzyme your group used in the big table and select the checkbox next to it. If you used more than one enzyme, only select one enzyme at a time. Then click on “Digest” at the bottom of the page. The results page shows a map of the plasmid with the restriction sites marked. To find the precise locations of the restriction sites, click on “Fragments” or you can even view what a gel of your digest would look like by clicking on “View gel.”

3. What enzyme(s) did your group use?

HINDIII

BAMHI

4. Where are the restriction sites for this enzyme in the pGLO plasmid? What fragment sizes should you see on the gel you ran?

The restriction sites for HINDIII are 4722 bp and 649 bp. The coordinates are 2115-1465 and 1466-2114.

The restriction sites for BAMHI are 4526 bp, 625 bp, and 220 bp. The coordinates are 2131-1285, 1286-1910 and 1911-2130.

**Part II – Comparing the GFP Gene From Jellyfish with the pGLO Version**

Go to Uniprot (<http://www.uniprot.org/>) and enter “green fluorescent protein” in the search box at the top of the page and click Search. On the results page, you should see the protein we used from jellyfish. Go to the page with information on the GFP protein by clicking on the accession number. Browse through this page to get an idea of the different types of information available at Uniprot and to answer the following questions:

5. How many amino acids are in the GFP sequence?

There are 238 amino acids in the GFP sequence.

6. What is the function of GFP?

GFP is an energy-transfer acceptor. It transfers energy in order to produe a fluorescent light.

Scroll down the page to the section titled “Cross References” and find the “Sequence Databases” section. In the drop down box that says EMBL, select GenBank then click on the first mRNA sequence identifier (it should be M62654). This takes you to NCBI GenBank for the page with information on the mRNA sequence that codes for GFP protein.

7. How many base pairs are in the mRNA sequence?

There are 5170 bp in the mRNA sequence.

8. How many amino acids could this sequence code for? How does this compare to the number of amino acids in the GFP protein?

1723 amino acids could be coded for. There are way more amino acids in this sequence than in the GFP protein.

9. How many exons are in the GFP mRNA? How many introns?

There are 3 exons in the GFP mRNA and 3 introns.

To see how the pGLO plasmid compare to the GFP mRNA, we can align the two sequences and see where they are similar. Go to NCBI BLAST (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) and scroll to the bottom of the page. Click on the link for “Align two (or more) sequences using BLAST (bl2seq).” The bl2seq program will compare two sequences and tell you where they are similar to each other. In the top box, enter the accession number for the GFP mRNA (M62654) and enter the accession number for the pGLO plasmid (U62637) in the bottom box. Then click on BLAST at the bottom of the page.

10. How many regions of similarity are there between the GFP mRNA and the pGLO plasmid? How do these regions compare to the exons and introns in the GFP mRNA?

There are 3 regions of similarity between the GFP mRNA and the pGLO plasmid. There are similarities between the exons in the GFP mRNA and the regions of similarity this makes sense since the exons are expressed.

**Part III – GFP Structure**

Go to PDB (<http://www.pdb.org/>) and enter 1EMA in the search box at the top of the page, then click “Search.” The results page that comes up shows a summary of information about the GFP protein and its structure. On the right side of the page, under the pretty picture, click on “Protein Workshop.” This will download a molecular viewer and display the GFP protein in an interactive window. The first thing to notice is the overall layout. The viewer is to the left and the control panels are to the right. If you click and drag in the viewer, the structure will rotate.

Next, look at the layout of the control panel. The application was designed for quick editing in a step-by-step process. Notice the numbers on the panel, 1, 2, 3, 4. To change things in the viewer, you usually have to select an option in all four areas.

The default display in Protein Workshop is a representation of the protein called “Ribbons” which is a cartoon style drawing. They look pretty and can simplify the protein structure. To see how much they simplify the structure, let’s turn them off and view all of the atoms and bonds in the protein structure.

First, click on the “Visibility” button in panel 1. Then select “Ribbons” in panel 2, and click 1EMA in panel 4. The viewer display should now be blank.

Second, turn on all atoms and bonds by clicking on the “Visibility” button in panel 1, select “Atoms and Bonds” in panel 2, and click on “1EMA” in panel 4. You should now see all of the atoms and bonds in the GFP protein. Each dot is an atom and the lines between them are bonds. The atoms are color coded by element, with green for carbon, blue for nitrogen, and red for oxygen. Be sure to spend some time rotating the protein around to see it from all angles.

So far, our visualizations have looked like lines with lots of space between them making it appear as if GFP has lots of holes and maybe even a tunnel down the middle. By viewing proteins using spacefilling models (also called CPK models) you can visual how much space each atom takes up and get a better idea of what the protein physically looks like.

To change to the spacefilling representation, click on the “Styles” button in panel 1, select “Atoms and Bonds” in panel 2, change Radius of Atoms to CPK in panel 3, and click on “1EMA” in panel 4.

Unless you are interested in something on the surface of the protein the spacefilling model may not be the best way to look at the protein. One thing that often helps visualize a particular part of a protein is to show it in a different representation than the rest of the molecule. GFP contains a chromophore, which is the functional region of the protein that fluorescesces green. Let’s show the entire protein in ribbons and the chromophore in spacefill to highlight it.

First, click on the “Visibility” button in panel 1, select “Atoms and Bonds” in panel 2, then click “1EMA” in panel 4. This should turn off the spacefilling representation.

Second, click on the “Visibility” button in panel 1, select “Ribbons” in panel 2, then click “1EMA” in panel 4. The ribbon view should be back on now.

Finally, click on the “Visibility” button in panel 1, select “Atoms and Bonds” in panel 2, then click on the + before 1EMA in panel 4. This will expand the view in panel four to show Chain A and Water molecules. Next, click on the + next to Chain A to expand the view in panel 4 to show all of the amino acids in the GFP protein. Scroll down in panel 4 until you see “66 CRO” and click on it.

The chromophore should now be displayed in spacefill and the rest of the protein should be in ribbons.

11. Show your image to the instructor or TA. Where in the protein is the chromophore located?

The chromophore is located in the middle of the protein.