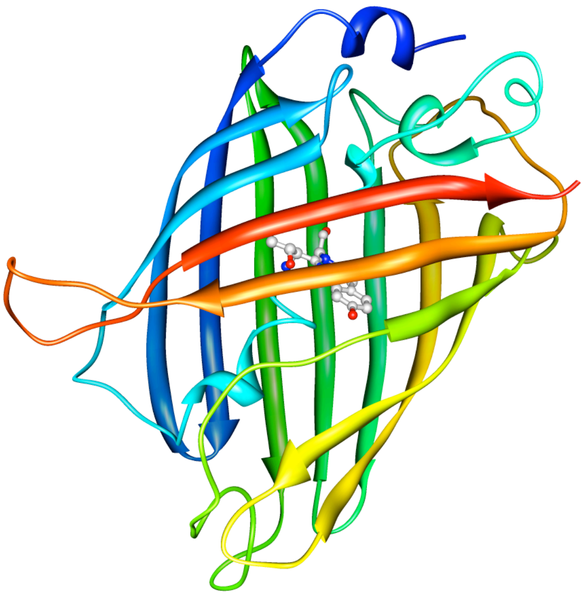
info@rcsb.org • www.rcsb.org

Bioinformatics of Green Fluorescent Protein



This bioinformatics tutorial explores the relationship between gene sequence, protein structure, and biological function in the context of the *green fluorescent protein* (GFP). In this tutorial you will:

- find protein structures using search tools on the RCSB PDB website;
- use molecular visualization tools to explore the GFP structure and function;
- find the GFP gene and view important mutations.

The PDB archive is the primary repository of experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the wwPDB, the RCSB PDB[a] curates and annotates structural data from researchers around the world. The RCSB PDB also provides a variety of tools and resources for searching, visualizing, downloading, and analyzing biomolecular structures.

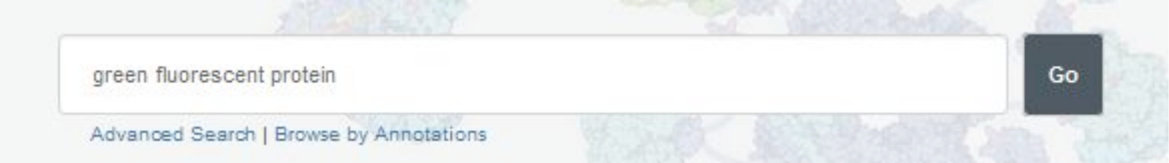
Please send any comments or questions about this tutorial to info@rcsb.org.

Part I. Finding and Exploring the 3D Structure of Green Fluorescent Protein

In this first part, we will find a structure of green fluorescent protein in the RCSB PDB, then use several tools to explore its structure and function.

Task 1: Find structures of green fluorescent protein at the RCSB PDB website.

1. Go to the RCSB PDB website at http://www.rcsb.org
2. Perform a "PDB ID or keyword" search by typing the keyword "green fluorescent protein" in the text box on the search bar at the top of the first page:



green fluorescent protein

3. Click the *Go* button.
4. The result page will contain a list of proteins related to GFP.

You can explore all of these different structures by clicking on different examples, creating reports, or generating an image collage.

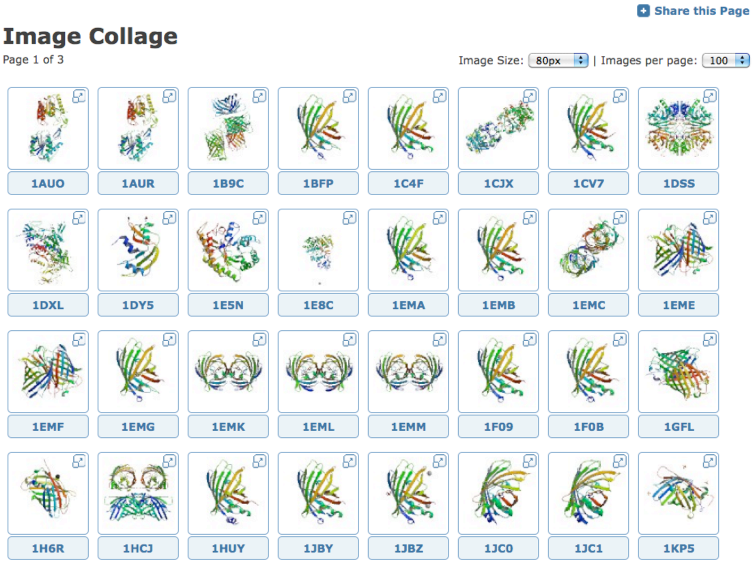


Image Collage

Page 1 of 3

Image Size: 80px | Images per page: 100

1AUO 1AUR 1B9C 1BFP 1C4F 1CJX 1CV7 1DSS

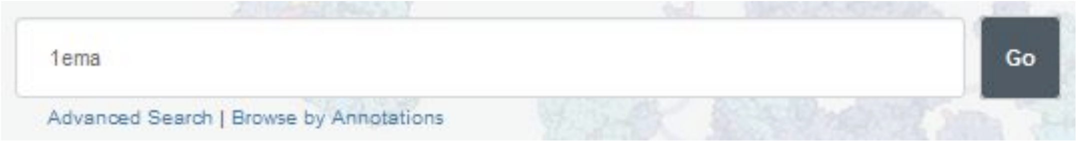
1DXL 1DY5 1E5N 1E8C 1EMA 1EMB 1EMC 1EME

1EMF 1EMG 1EMK 1EML 1EMM 1F09 1F0B 1GFL

1H6R 1HCJ 1HUY 1JBY 1JBZ 1JC0 1JC1 1KP5

Example image collage for some of the structures found when searching for **green fluorescent protein.**

For this exercise, we will use a protein that was taken from the jellyfish *Aequorea victoria* with PDB ID 1ema[1]. You can easily find this protein by entering the PDB ID 1ema in the search bar at the top of the page and clicking *'Go'*. This will take you to the Structure Summary page for 1ema.



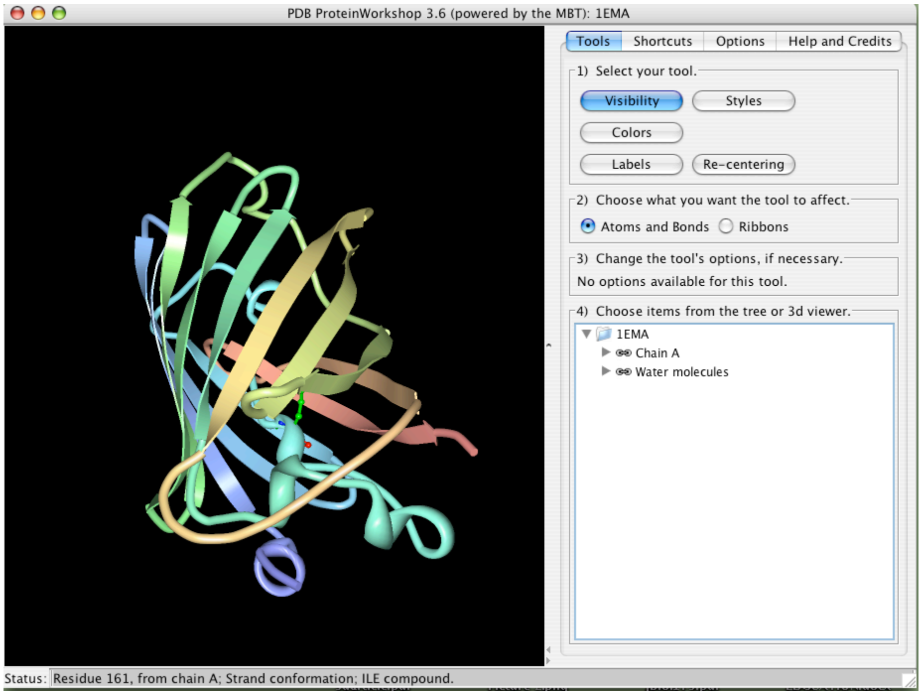
1ema

Task 2: Basic protein visualization using Protein Workshop

Now let's take a look at the GFP structure in close detail.

1. You should still be on the Structure Summary Page for entry 1emaIf you are not, simply enter the PDB ID, "1ema" into the top query bar and click the *Search* button.
2. On the right side of the Structure Summary page, you will see a box containing an image and links to 3-D molecular viewers such as Jmol, Protein Workshop etc.
3. Under the image click on the link that says *Protein Workshop*
4. This will download the viewer. In the process of doing this it will ask you if you want to *trust this application*. This is part of the Java security mechanisms; you simply accept/trust each one and click *run* when prompted.

Once the structure is loaded you should see something that looks like the following:



PDB ProteinWorkshop 3.6 (powered by the MBT): 1EMA

1) Select your tool.

2) Choose what you want the tool to affect.
● Atoms and Bonds ○ Ribbons

3) Change the tool's options, if necessary.
No options available for this tool.

4) Choose items from the tree or 3d viewer.

Status: Residue 161, from chain A; Strand conformation; ILE compound.

[1] Crystal structure of the Aequorea victoria green fluorescent protein. Ormo, M., Cubitt, A.B., Kallio, K., Gross, L.A., Tsien, R.Y., Remington, S.J. (1996) Science 273: 1392-1395

The *viewer* is to the left and the *control panels* are to the right. If you click and drag the mouse in the viewer you will see the structure rotate. You can also zoom the structure by clicking the middle button and dragging (or, shift+click with a one-button mouse), and translate the structure using the right button (or ctrl+click with a one-button mouse). Take a minute to get familiar with these controls.

Protein Workshop automatically displays a *ribbon representation* of the protein structure. This representation represents the polypeptide chain of the protein, but uses flat arrows to show beta strands and curly ribbons to show alpha helices. The chain is also colored in rainbow colors from red to blue from one end of the chain to the other, so you can follow the chain as it folds into this complex structure.

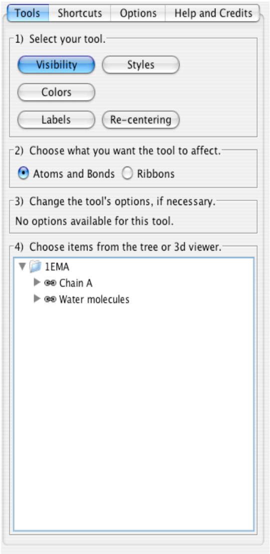
You can learn more about Green Fluorescent Protein in the Molecule of the Month feature on GFP at: http://www.rcsb.org/pdb/101/motm.do?momID=42. You can get to this from the 1ema Structure Summary Page by selecting the "Green Fluorescent Protein (GFP)" link in the 'Molecule of the Month' Box at the bottom right of the page.

If you want to try building your own model of GFP, there is a paper cut-out form at: http://www.rcsb.org/pdb/101/static101.do?p=education_discussion/educational_resources/GFP_activity.html. Or, from the RCSB PDB home page, select the PDB-101 icon at the top left and select the 'Educational Resources tab and then 'Activities/Lessons' from the menu on the left. Scroll down to 'Bioinformatics of Green Fluorescent Protein' and select 'Paper Model'.

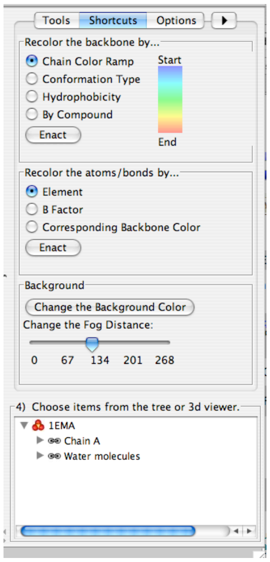
Task 3: Different ways of looking at GFP in Protein Workshop
The *control panel* of Protein Workshop is designed for quick and simple editing in a four-step process. Notice the boxes numbered 1-4 on the *Tools* panel. These are to help you go through the steps of using this tool. Other panels provide advanced options, described in more detail below.

Panels 1-3 change when the Tools, Shortcuts, or Options buttons are selected:

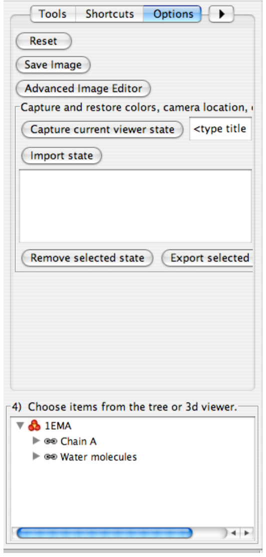
Tools Menu


Tools
Shortcuts
Options
Help and Credits
1) Select your tool.
Visibility
Styles
Colors
Labels
Re-centering
2) Choose what you want the tool to affect.
● Atoms and Bonds ○ Ribbons
3) Change the tool's options, if necessary.
No options available for this tool.
4) Choose items from the tree or 3d viewer.
1EMA
Chain A
Water molecules

Shortcuts Menu



Options Menu



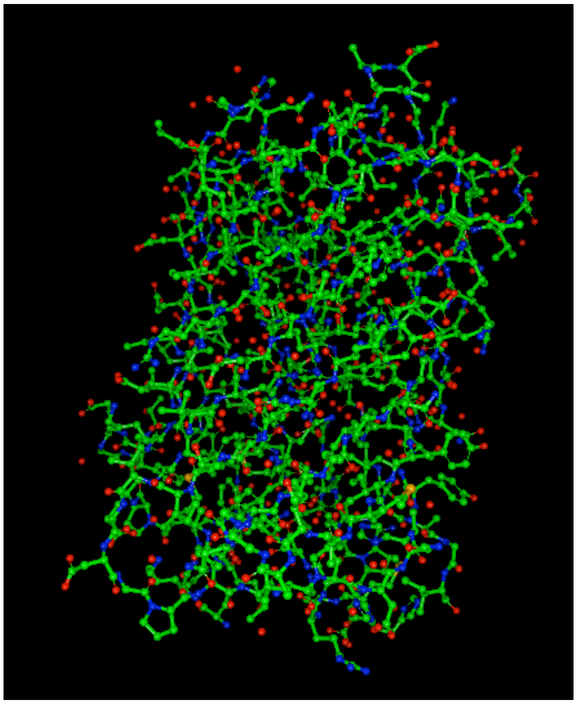
To switch to a view that shows all atoms, you need to go through several steps. First, in the *Control Panel* under the Tools menu, turn off the ribbon representation, by performing one action in each of the four boxes:

1. Click on the *Visibility* Button
2. Select *Ribbons*
3. (No options in the Visibility Tool.)
4. Click on *1ema* (the PDB ID) in the bottom tree viewer.

At this point the viewer should be blank, because you have essentially chosen to toggle OFF the ribbon representation. If you click on *1ema* again, you can toggle the ribbon back ON. With the ribbon OFF, display atoms using a ball-and-stick representation:

1. Click on the *Visibility* Button
2. Select *Atoms and Bonds* this time.
3. (No options in the Visibility Tool.)
4. Click on *1ema* (the PDB ID) in the bottom tree viewer.

You should see the following:



Each sphere represents an atom, and the lines between these atoms represent covalent bonds.

Carbon atoms are green
Nitrogen atoms are blue
Oxygen atoms are red
Sulfur atoms are yellow

To get a better look at the protein structure with all atoms and bonds shown, zoom into the protein. Hold the shift key, and click and drag the mouse downward.

As you can see, this representation shows a lot of information that makes it difficult to find specific structural features with all atoms shown. Features that you might be able to identify include:

- yellow sulfur atoms in cysteine and methionine
- single oxygen atoms in red are water molecules
- the chromophore is difficult to spot with this representation—look for a five-membered ring connected to a six-membered ring, buried in the middle of the protein. The next exercise will show you an easy way to find the chromophore.

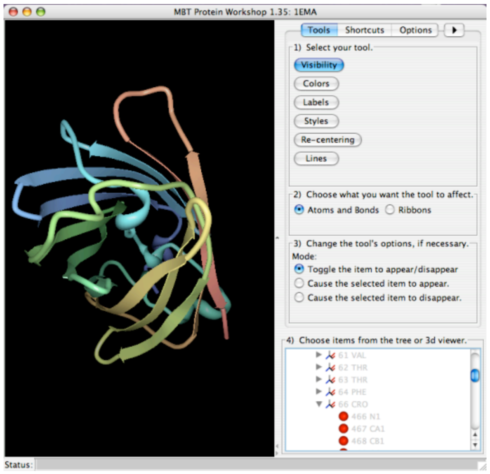
To return to the default view of the structure, we'll follow our steps in reverse.

1. Click on 1ema in the bottom tree viewer. This will make the molecule disappear.
2. Select on "Ribbons" in box #2.
3. Click on 1ema in the bottom tree viewer. This will make the molecule appear, but now in ribbon representation. You can zoom the image to show the whole protein by pressing the shift key and dragging the mouse up.

Task 4: Exploring the chromophore with a combination of ribbon view with atom view

Often, we want to explore some parts of the protein in atomic detail (such as the chromophore in GFP), but use a simple ribbon representation for the rest of the protein.

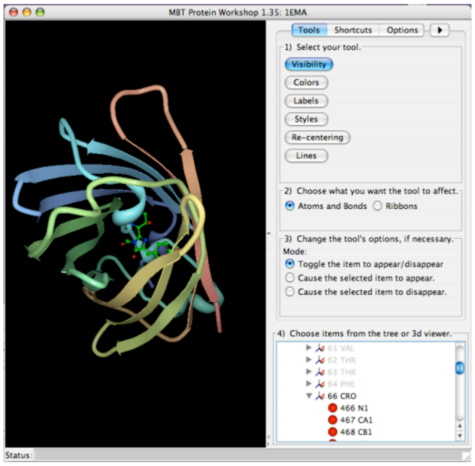
1. Still in Protein Workshop, select the Visibility tool
2. Select *Atoms and Bonds* for what you want the tool to affect – This means that when we select the chromophore, it will appear as atoms bonds.
3. (No options in the Visibility Tool)
4. In box 4, open up Chain A. Scroll down and select the position of the chromophore – CRO 66. Selecting and deselecting will cause the chromophore to appear (in atom representation) and disappear. You can rotate the structure to see how this piece fits in the overall shape of the protein.



2) Choose what you want the tool to affect.
● Atoms and Bonds ○ Ribbons

3) Change the tool's options, if necessary.
Mode:
● Toggle the item to appear/disappear
○ Cause the selected item to appear.
○ Cause the selected item to disappear.

Here, the chromopore is not shown



2) Choose what you want the tool to affect.
● Atoms and Bonds ○ Ribbons

3) Change the tool's options, if necessary.
Mode:
● Toggle the item to appear/disappear
○ Cause the selected item to appear.
○ Cause the selected item to disappear.

Here, it is shown in atoms and bonds

Task 5: Visualize hydrogen bonds in proteins

The secondary structures of proteins, like the beta sheets seen in GFP, are stabilized by hydrogen bonds. The most important hydrogen bonds in proteins are formed between N-H hydrogen atom and the C=O oxygen atom in the protein backbone—these hydrogen bonds link different portions of the chain and stabilize the folded structure. However, it can be tricky to see these bonds using many of the structures in the PDB because crystallographic structures typically do not include the hydrogen atoms. Instead, we often draw the hydrogen bond between the nitrogen atom in the N-H group (which is included in the PDB file) and the oxygen atom of the C=O group.

Protein Workshop does not display hydrogen bonds, so we'll look at hydrogen bonds in GFP using *JMol*.

1. Go again to the Structure Summary page browser page for PDB entry 1ema.
2. Underneath the image on the right, select "3D View" to launch the *JMol* viewer. Select "Allow" to run the Java program.

3. Commands can be typed in the'Scripting Options' section (to see the text box, you will have to expand it by clicking the title) beneath the image viewer to change the representation of the molecule.

4. To show the hydrogen bonds in the protein backbone:

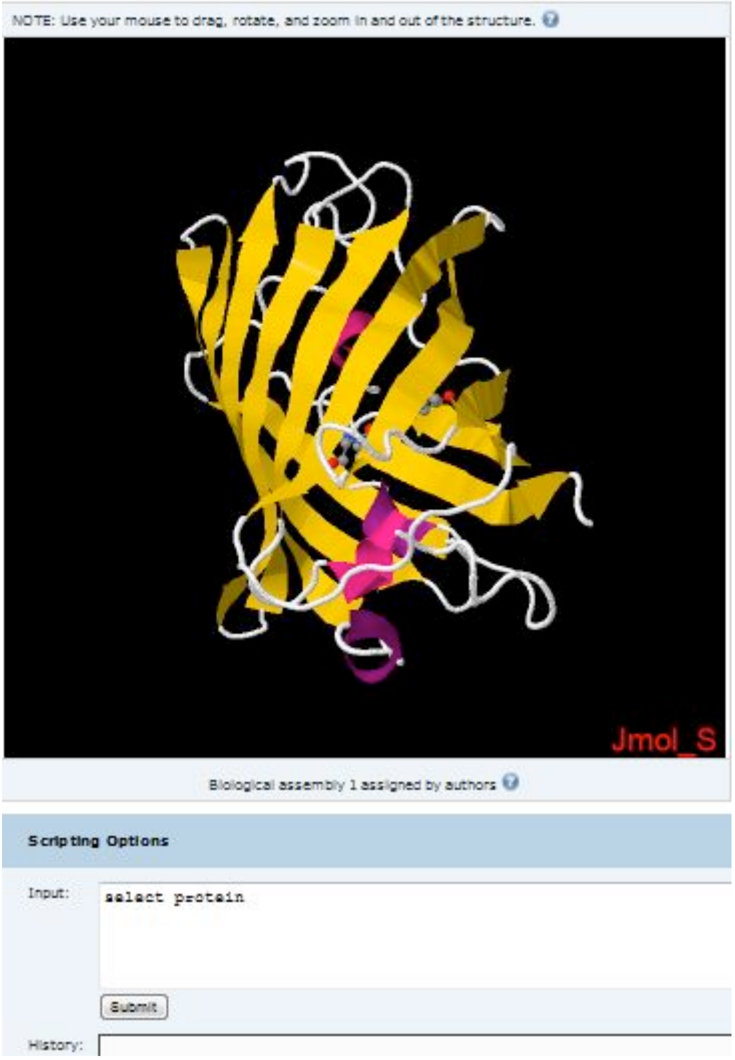
a) Type **select protein** in the 'Input' box, and select the 'Submit' button.

b) Type **cartoon off** and select the 'Submit' button.

c) Enter **select backbone** and select the 'Submit' button.

d) Enter **wireframe 100** and select the 'Submit' button.

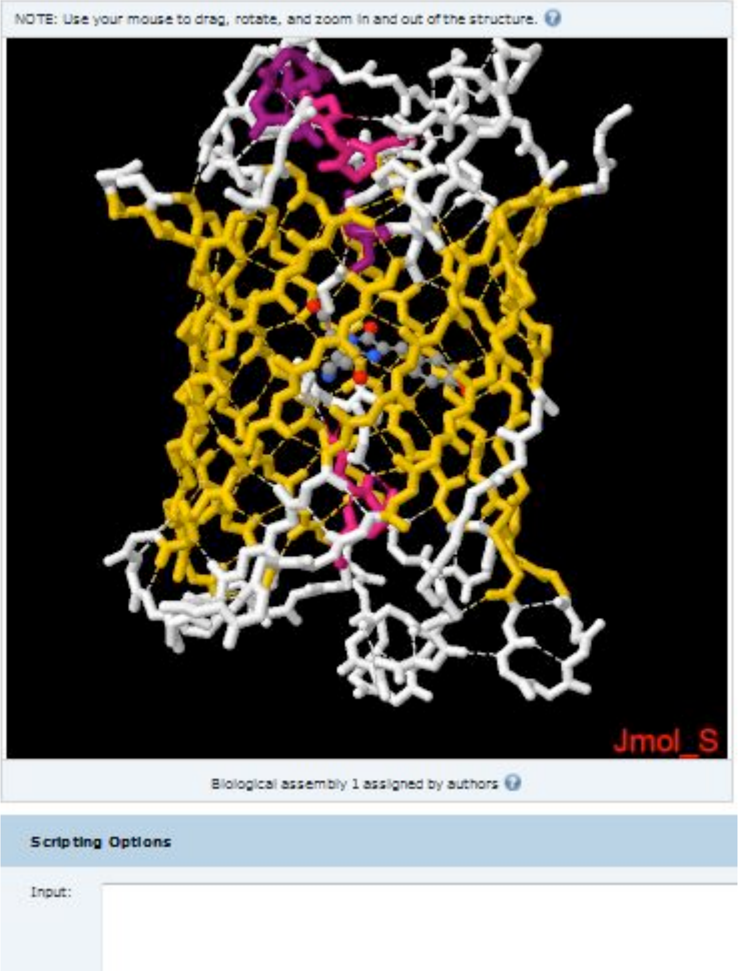
e) Enter **calculate hbonds** and select the 'Submit' button.



NOTE: Use your mouse to drag, rotate, and zoom in and out of the structure.

Biological assembly 1 assigned by authors

Input: select protein



NOTE: Use your mouse to drag, rotate, and zoom in and out of the structure.

Biological assembly 1 assigned by authors

You can rotate the GFP to see all of the hydrogen bonds.

<u>Additional Activity</u>: Hydrogen bonds are also important for stabilizing alpha helices. Try using *Jmol* to look at the hydrogen bonds in the hemoglobin structure with PDB ID 4hhb[2].

[2] The crystal structure of human deoxyhaemoglobin at 1.74 A resolution. Fermi, G., Perutz, M.F., Shaanan, B., Fourme, R. (1984) J.Mol.Biol. 175: 159-174

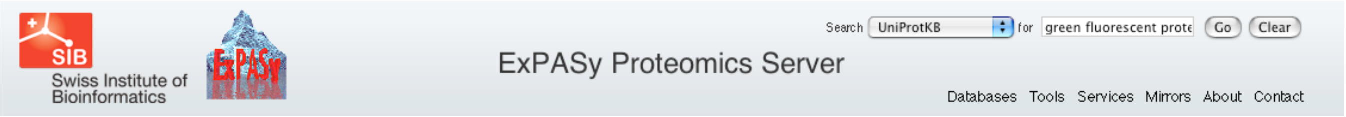
Part II. Gene and Protein Sequences of Green Fluorescent Protein

In this second part, we will use several online resources to find the sequence of the gene for green fluorescent protein, translate this into a protein sequence, and use this sequence to find mutant forms of the protein with altered function.

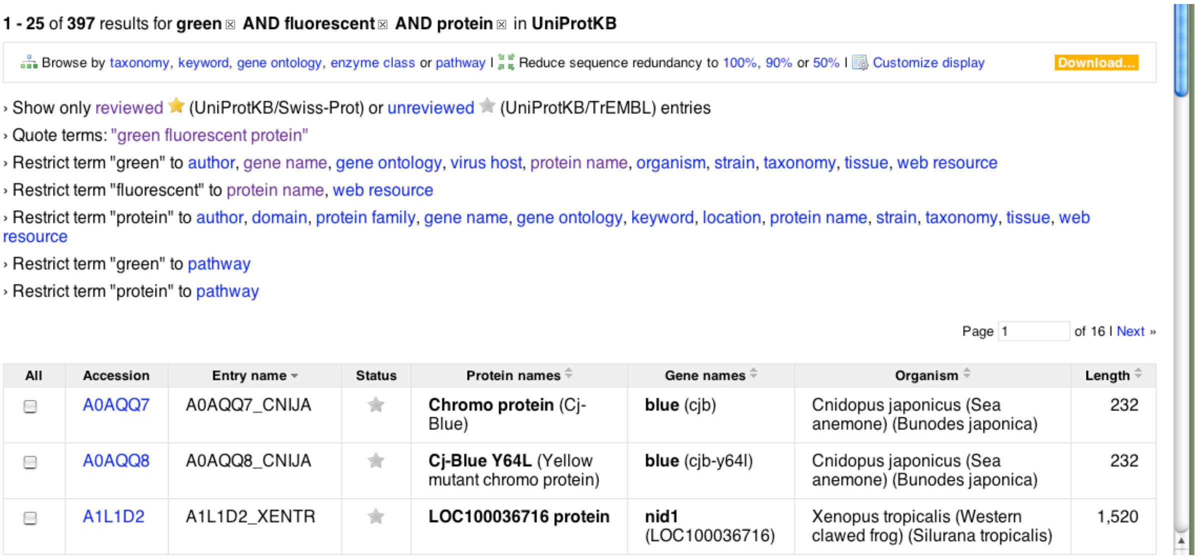
Task 1: Find the DNA sequence for the GFP gene in UniProtKB

The UniProtKB[b] database is a resource that organizes and annotates protein sequences. This database contains important information for researchers to study the relationship between protein sequence and protein function. There are two kinds of annotations available from this database. One, where staff scientists annotate the sequences manually based on published literature, while the other is an automated process using sophisticated software tools. It is generally accepted that the manually annotated entries have more reliable information, so we will use the human-annotated sequences from a section of the UniProtKB, called Swiss-Prot[c].

1. Point your web browser to the following URL: http://ca.expasy.org/
2. In the top pulldown list select *UniProtKB* and enter the text "Green fluorescent protein" in the adjacent box.


SIB
Swiss Institute of Bioinformatics
ExPASy
ExPASy Proteomics Server
Search UniProtKB for green fluorescent prote Go Clear
Databases Tools Services Mirrors About Contact

3. Click on the *Go* button.
You should see a result much like the following:



1 - 25 of **397** results for **green** ☒ **AND fluorescent** ☒ **AND protein** ☒ in **UniProtKB**

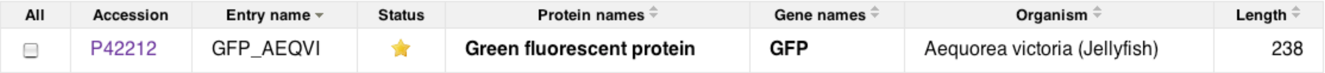
Browse by taxonomy, keyword, gene ontology, enzyme class or pathway | Reduce sequence redundancy to 100%, 90% or 50% | Customize display Download...

› Show only reviewed ★ (UniProtKB/Swiss-Prot) or unreviewed ★ (UniProtKB/TrEMBL) entries
› Quote terms: "green fluorescent protein"
› Restrict term "green" to author, gene name, gene ontology, virus host, protein name, organism, strain, taxonomy, tissue, web resource
› Restrict term "fluorescent" to protein name, web resource
› Restrict term "protein" to author, domain, protein family, gene name, gene ontology, keyword, location, protein name, strain, taxonomy, tissue, web resource
› Restrict term "green" to pathway
› Restrict term "protein" to pathway

Page 1 of 16 | Next »

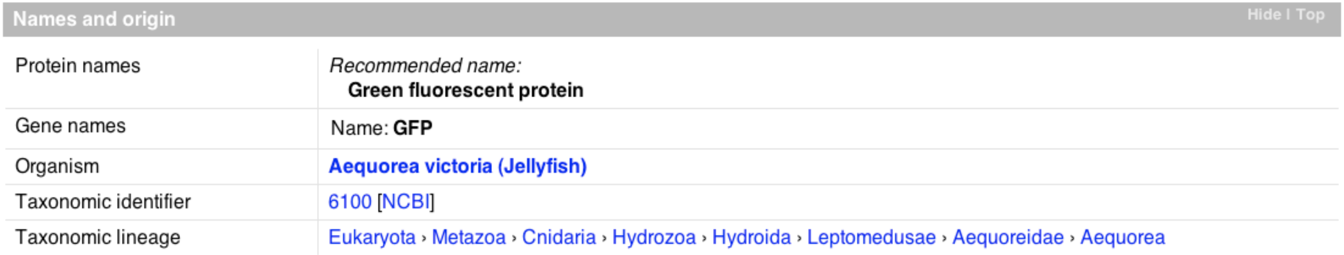
All	Accession	Entry name	Status	Protein names	Gene names	Organism	Length
□	A0AQQ7	A0AQQ7_CNIJA	★	**Chromo protein** (Cj-Blue)	**blue** (cjb)	Cnidopus japonicus (Sea anemone) (Bunodes japonica)	232
□	A0AQQ8	A0AQQ8_CNIJA	★	**Cj-Blue Y64L** (Yellow mutant chromo protein)	**blue** (cjb-y64l)	Cnidopus japonicus (Sea anemone) (Bunodes japonica)	232
□	A1L1D2	A1L1D2_XENTR	★	**LOC100036716 protein**	**nid1** (LOC100036716)	Xenopus tropicalis (Western clawed frog) (Silurana tropicalis)	1,520

4. Refine this search by first clicking on the *reviewed* entries. Then refine the search by restricting the term 'green' to **protein name**. Of the results in the *UniProtKB/Swiss-Prot* click on the one titled *GFP_AEQVI(P42212)*



All	Accession	Entry name	Status	Protein names	Gene names	Organism	Length
□	P42212	GFP_AEQVI	★	**Green fluorescent protein**	**GFP**	Aequorea victoria (Jellyfish)	238

You should now see the summary page for GFP.



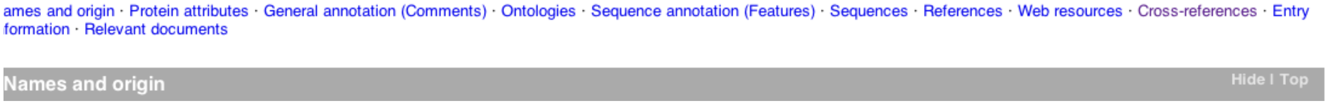
Names and origin

Protein names	*Recommended name:* **Green fluorescent protein**
Gene names	Name: **GFP**
Organism	**Aequorea victoria (Jellyfish)**
Taxonomic identifier	6100 [NCBI]
Taxonomic lineage	Eukaryota › Metazoa › Cnidaria › Hydrozoa › Hydroida › Leptomedusae › Aequoreidae › Aequorea

Shortcut: To jump to this page enter the accession id "P42212" into the search window on the site http://ca.expasy.org. Click the *Go* button.

Let's take a look at the gene sequence.

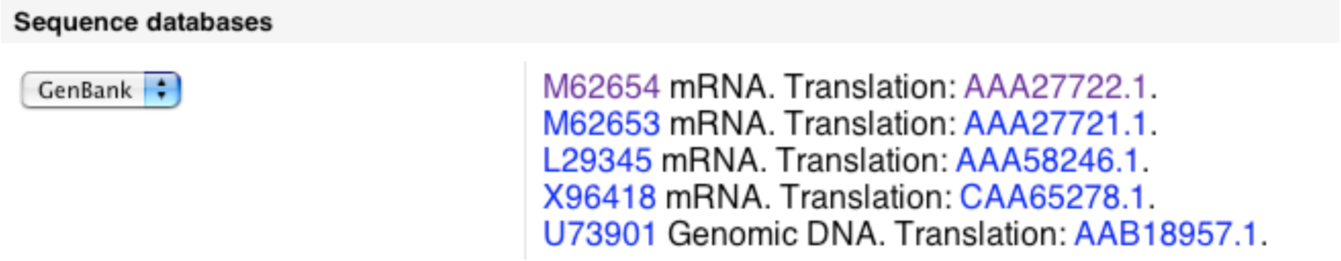
5. At the top of the page, click on the option titled *Cross-references*



ames and origin · Protein attributes · General annotation (Comments) · Ontologies · Sequence annotation (Features) · Sequences · References · Web resources · Cross-references · Entry formation · Relevant documents

Names and origin

6. Click on the radio button next to the *Genbank* option under "Sequence databases"



Sequence databases

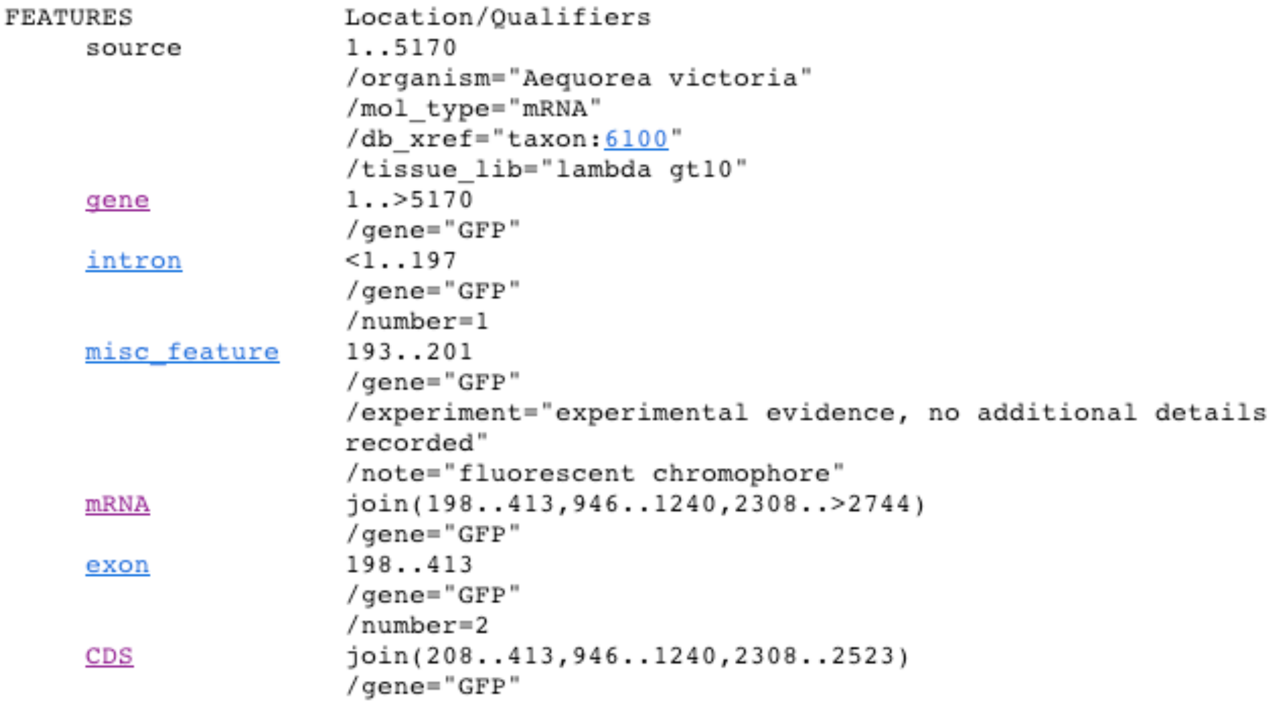
GenBank

M62654 mRNA. Translation: AAA27722.1.
M62653 mRNA. Translation: AAA27721.1.
L29345 mRNA. Translation: AAA58246.1.
X96418 mRNA. Translation: CAA65278.1.
U73901 Genomic DNA. Translation: AAB18957.1.

7. Click on the mRNA link on the top line. This will retrieve the *Genbank* entry for this molecule.
8. On the resulting page, scroll down to the bottom. This is the mRNA gene sequence that encodes the protein sequence of GFP. This sequence represents the entire genomic sequence for the gene, including the 5' untranslated region (UTR), introns and 3'UTR. The sequence represents the pre-mRNA before splicing and translation. Thus, it turns out that there is more information here than we need. Only parts of this sequence are used for the protein. Let's simplify this by looking at only these coding regions.

Shortcut: To jump to this page, go to the NCBI website: http://www.ncbi.nlm.nih.gov/ and select the *Nucleotide* database in the *search* pulldown menu. Enter the text M62654 and click the *Go* button.

9. Click on the link labeled ***mRNA*** under the *FEATURES* category.   This sequence corresponds to the mature mRNA that serves as template for translation (includes 5’ and 3’ UTRs). The mRNA differs from the **CDS (CoDing Sequence)** in that the CDS sequence only contains the region of the mRNA defined by the start and stop codons, therefore coding sequences begin with an "ATG" and end with a stop codon.



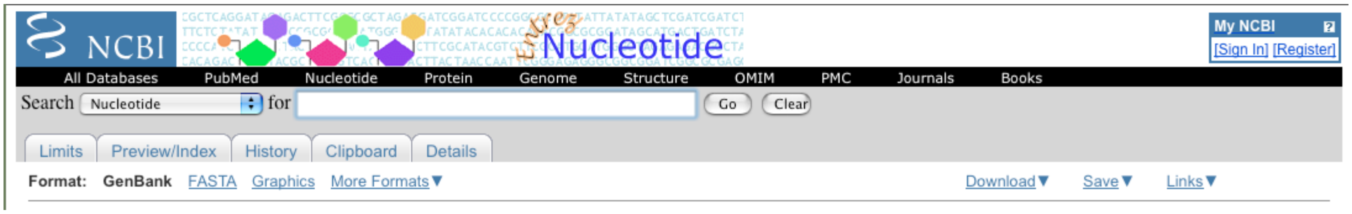
```
FEATURES             Location/Qualifiers
     source          1..5170
                     /organism="Aequorea victoria"
                     /mol_type="mRNA"
                     /db_xref="taxon:6100"
                     /tissue_lib="lambda gt10"
     gene            1..>5170
                     /gene="GFP"
     intron          <1..197
                     /gene="GFP"
                     /number=1
     misc_feature    193..201
                     /gene="GFP"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="fluorescent chromophore"
     mRNA            join(198..413,946..1240,2308..>2744)
                     /gene="GFP"
     exon            198..413
                     /gene="GFP"
                     /number=2
     CDS             join(208..413,946..1240,2308..2523)
                     /gene="GFP"
```

10. Scroll to the bottom of the page. This is the entire DNA sequence that translates into our protein GFP.

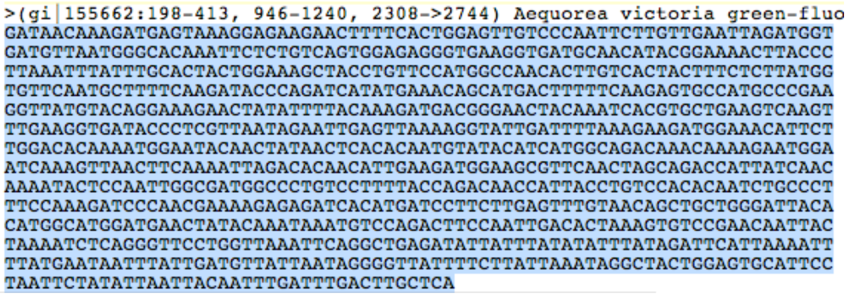
Task 2: Translation of the DNA sequence to the protein sequence

Now that we have the DNA sequence, translate it back to the protein sequence using the translate tool provided on the ExPASy website. The first thing we need to do is copy the DNA sequence to your computer clipboard. We will then paste this into the tool.

1. Scroll to the top of the sequence page and locate the menu option called *Format*. Select the FASTA format option.



In the resulting page select everything except the first line. In other words we just want to select the actual sequence. (The FASTA format always has a comment line on the top line and all subsequent lines are the sequence).



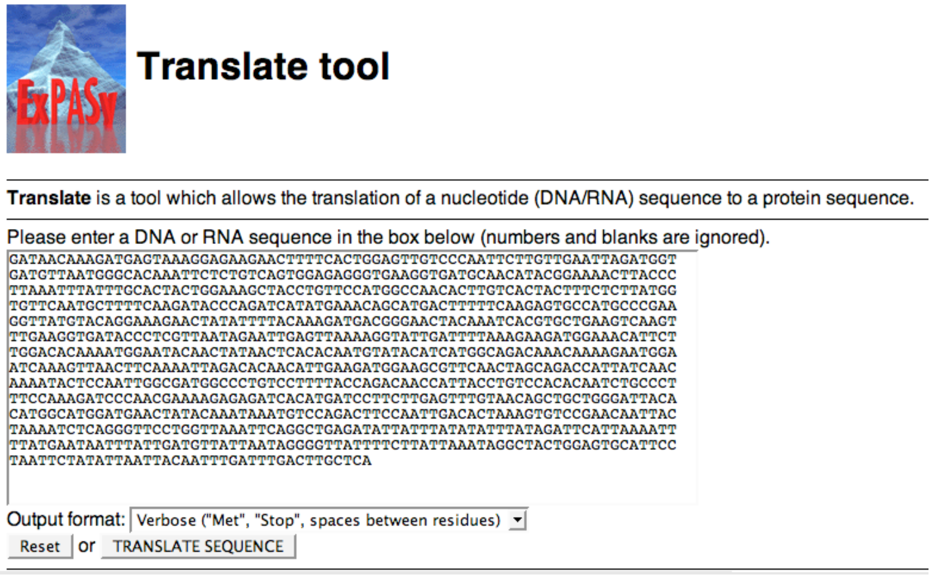
```
>(gi|155662:198-413, 946-1240, 2308->2744) Aequorea victoria green-fluo
GATAACAAAGATGAGTAAAGGAGAAGAACTTTTCACTGGAGTTGTCCCAATTCTTGTTGAATTAGATGGT
GATGTTAATGGGCACAAATTCTCTGTCAGTGGAGAGGGTGAAGGTGATGCAACATACGGAAAACTTACCC
TTAAATTTATTTGCACTACTGGAAAGCTACCTGTTCCATGGCCAACACTTGTCACTACTTTCTCTTATGG
TGTTCAATGCTTTTCAAGATACCCAGATCATATGAAACAGCATGACTTTTTCAAGAGTGCCATGCCCGAA
GGTTATGTACAGGAAAGAACTATATTTTACAAAGATGACGGGAACTACAAATCACGTGCTGAAGTCAAGT
TTGAAGGTGATACCCTCGTTAATAGAATTGAGTTAAAAGGTATTGATTTTAAAGAAGATGGAAACATTCT
TGGACACAAAATGGAATACAACTATAACTCACACAATGTATACATCATGGCAGACAAACAAAAGAATGGA
ATCAAAGTTAACTTCAAAATTAGACACAACATTGAAGATGGAAGCGTTCAACTAGCAGACCATTATCAAC
AAAATACTCCAATTGGCGATGGCCCTGTCCTTTTACCAGACAACCATTACCTGTCCACACAATCTGCCCT
TTCCAAAGATCCCAACGAAAAGAGAGATCACATGATCCTTCTTGAGTTTGTAACAGCTGCTGGGATTACA
CATGGCATGGATGAACTATACAAATAAATGTCCAGACTTCCAATTGACACTAAAGTGTCCGAACAATTAC
TAAAATCTCAGGGTTCCTGGTTAAATTCAGGCTGAGATATTATTTATATATTTATAGATTCATTAAAATT
TTATGAATAATTTATTGATGTTATTAATAGGGGTTATTTTCTTATTAAATAGGCTACTGGAGTGCATTCC
TAATTCTATATTAATTACAATTTGATTTGACTTGCTCA
```

2. Copy this to your computer clipboard using Edit->Copy from the Browser menu or for Mac users: ‘Apple’ + ‘c’, and for PC users: ‘ctrl’ + ‘c’

3. Point your web browser to the Translate tool: http://www.expasy.org/tools/dna.html

4. Click in the text window and paste your sequence using Edit->Paste from the Browser menu or for Mac users: ‘Apple’ + ‘v’, and for PC users: ‘ctrl’ + ‘v’

Your browser should look like the following:



Translate tool

Translate is a tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.

Please enter a DNA or RNA sequence in the box below (numbers and blanks are ignored).

```
GATAACAAAGATGAGTAAAGGAGAAGAACTTTTCACTGGAGTTGTCCCAATTCTTGTTGAATTAGATGGT
GATGTTAATGGGCACAAATTCTCTGTCAGTGGAGAGGGTGAAGGTGATGCAACATACGGAAAACTTACCC
TTAAATTTATTTGCACTACTGGAAAGCTACCTGTTCCATGGCCAACACTTGTCACTACTTTCTCTTATGG
TGTTCAATGCTTTTCAAGATACCCAGATCATATGAAACAGCATGACTTTTTCAAGAGTGCCATGCCCGAA
GGTTATGTACAGGAAAGAACTATATTTTACAAAGATGACGGGAACTACAAATCACGTGCTGAAGTCAAGT
TTGAAGGTGATACCCTCGTTAATAGAATTGAGTTAAAAGGTATTGATTTTAAAGAAGATGGAAACATTCT
TGGACACAAAATGGAATACAACTATAACTCACACAATGTATACATCATGGCAGACAAACAAAAGAATGGA
ATCAAAGTTAACTTCAAAATTAGACACAACATTGAAGATGGAAGCGTTCAACTAGCAGACCATTATCAAC
AAAATACTCCAATTGGCGATGGCCCTGTCCTTTTACCAGACAACCATTACCTGTCCACACAATCTGCCCT
TTCCAAAGATCCCAACGAAAAGAGAGATCACATGATCCTTCTTGAGTTTGTAACAGCTGCTGGGATTACA
CATGGCATGGATGAACTATACAAATAAATGTCCAGACTTCCAATTGACACTAAAGTGTCCGAACAATTAC
TAAAATCTCAGGGTTCCTGGTTAAATTCAGGCTGAGATATTATTTATATATTTATAGATTCATTAAAATT
TTATGAATAATTTATTGATGTTATTAATAGGGGTTATTTTCTTATTAAATAGGCTACTGGAGTGCATTCC
TAATTCTATATTAATTACAATTTGATTTGACTTGCTCA
```

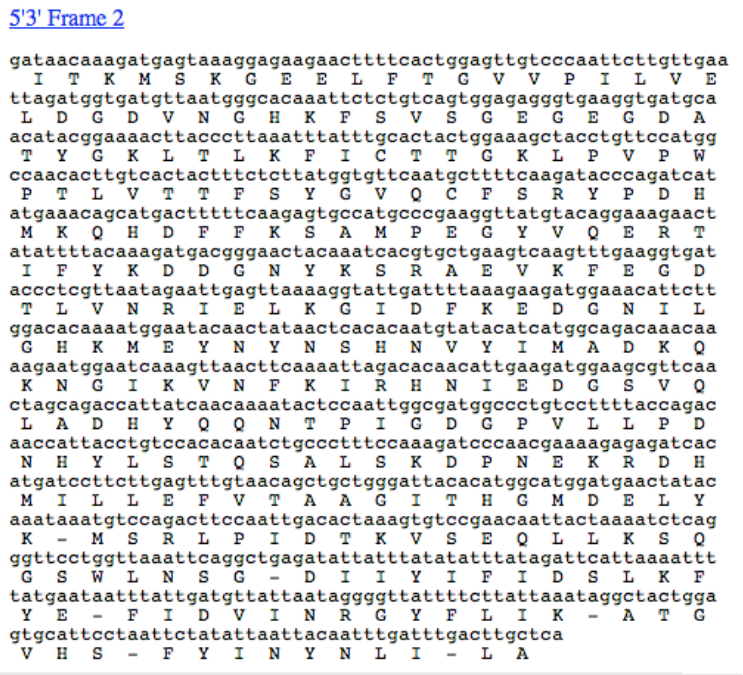
Output format: Verbose ("Met", "Stop", spaces between residues)

> **Warning:** Notice that the first line starts with "GATAAC…". If your first line starts with ">gi|1555662… ", you forgot to take out the comment line of the FASTA format. You only want to copy and paste the *sequence*.

5. At the bottom of the page, change the *Output format* choice to "*Includes Nucleotide Sequence"*
6. Now click on the *Translate Sequence* button.

The results will show six different sequences that each represent the different reading frames of DNA (three in one direction and three in the other). Only one of theses frames is the correct one used to translate the protein. Typically, the correct reading frame is the longest, uninterrupted (i.e. no internal stop codons) translation.

Notice that the *5' 3' Frame 2* appears to generate the best translation.



5'3' Frame 2

```
gataacaaagatgagtaaaggagaagaacttttcactggagttgtcccaattcttgttgaa
  I  T  K  M  S  K  G  E  E  L  F  T  G  V  V  P  I  L  V  E
ttagatggtgatgttaatgggcacaaattctctgtcagtggagagggtgaaggtgatgca
 L  D  G  D  V  N  G  H  K  F  S  V  S  G  E  G  E  G  D  A
acatacggaaaacttacccttaaatttatttgcactactggaaagctacctgttccatgg
 T  Y  G  K  L  T  L  K  F  I  C  T  T  G  K  L  P  V  P  W
ccaacacttgtcactactttctcttatggtgttcaatgcttttcaagatacccagatcat
 P  T  L  V  T  T  F  S  Y  G  V  Q  C  F  S  R  Y  P  D  H
atgaaacagcatgactttttcaagagtgccatgcccgaaggttatgtacaggaaagaact
 M  K  Q  H  D  F  F  K  S  A  M  P  E  G  Y  V  Q  E  R  T
atatttacaaagatgacgggaactacaaatcacgtgctgaagtcaagtttgaaggtgat
 I  F  Y  K  D  D  G  N  Y  K  S  R  A  E  V  K  F  E  G  D
accctcgttaatagaattgagttaaaaggtattgattttaaagaagatggaaacattctt
 T  L  V  N  R  I  E  L  K  G  I  D  F  K  E  D  G  N  I  L
ggacacaaaatggaatacaactataactcacacaatgtatacatcatggcagacaaacaa
 G  H  K  M  E  Y  N  Y  N  S  H  N  V  Y  I  M  A  D  K  Q
aagaatggaatcaaagttaacttcaaaattagacacaacattgaagatggaagcgttcaa
 K  N  G  I  K  V  N  F  K  I  R  H  N  I  E  D  G  S  V  Q
ctagcagaccattatcaacaaaatactccaattggcgatggccctgtccttttaccagac
 L  A  D  H  Y  Q  Q  N  T  P  I  G  D  G  P  V  L  L  P  D
aaccattacctgtccacacaatctgccctttccaaagatcccaacgaaaagagagatcac
 N  H  Y  L  S  T  Q  S  A  L  S  K  D  P  N  E  K  R  D  H
atgatccttcttgagtttgtaacagctgctgggattacacatggcatggatgaactatac
 M  I  L  L  E  F  V  T  A  A  G  I  T  H  G  M  D  E  L  Y
aaataaatgtccagacttccaattgacactaaagtgtccgaacaattactaaaatctcag
 K  -  M  S  R  L  P  I  D  T  K  V  S  E  Q  L  L  K  S  Q
ggttcctggttaaattcaggctgagatattatttatatatttatagattcattaaaattt
 G  S  W  L  N  S  G  -  D  I  I  Y  I  F  I  D  S  L  K  F
tatgaataatttattgatgttattaataggggttattttcttattaaataggctactgga
 Y  E  -  F  I  D  V  I  N  R  G  Y  F  L  I  K  -  A  T  G
gtgcattcctaattctatattaattacaatttgatttgacttgctca
 V  H  S  -  F  Y  I  N  Y  N  L  I  -  L  A
```

This shows how the protein is translated. Each line contains the DNA sequence and highlights the three-letter codon along with the corresponding amino acid. If we look at the sequences labeled *5'3' Frame 2* we will see that 'g' is not used, 'ata' translates to 'I', 'aca translates to 'T', aag translates to 'K' and 'atg' translates to 'M'.

The start codon is AUG (or in the DNA case it is ATG). This means that the process of translation, where the mRNA sequence is converted into a protein sequence, requires this three-letter code in order to start. This start codon occurs pretty early in our sequence (11th from the beginning), so this is probably a good result.

7. Click the link titled *5'3' Frame 2*.

The result page highlights the Methionine residues, or the starting point of the protein sequence. This corresponds to the ATG discussed previously.

Click the first 'M' in the sequence.

The resulting page will have a sequence that looks a lot like the protein sequence we started with. How can we tell if this sequence is the same as the one we had before?

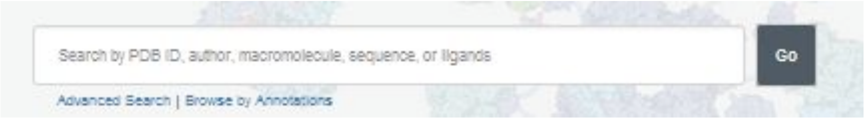
Task 3: Find proteins with similar sequences at the PDB

From the result page in the translation tool click the link that says *Fasta format* (highlighted in blue letters). You should get the following result:

```
>virt|VIRT14468|VIRT_14468 Translation of nucleotide sequence generated on
MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTL
VTTFSYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFYKDDGNYKSRAEVKFEGDTLV
NRIELKGIDFKEDGNILGHKMEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLAD
HYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMILLEFVTAAGITHGMDELYK
```

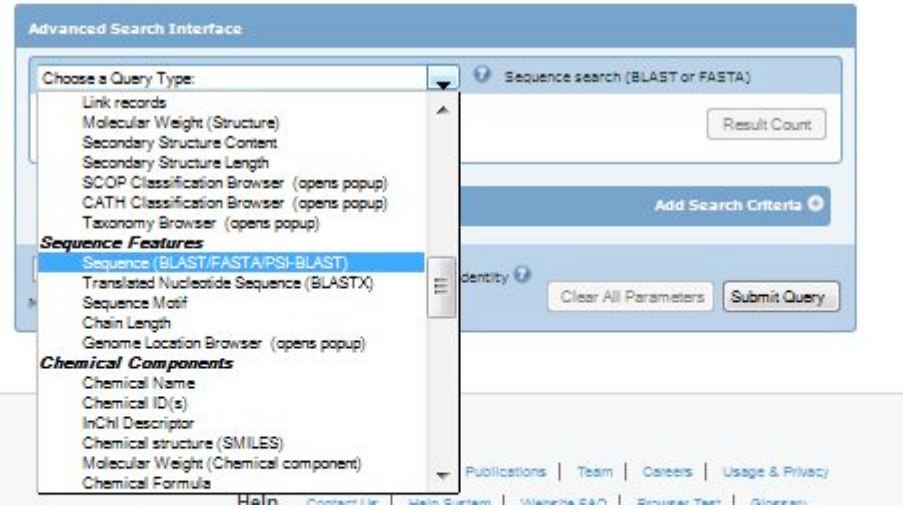
Now let’s use this protein sequence and compare it with sequences in the PDB. If we find sequences that are the same, it means that a researcher somewhere in the world solved the 3D structure of this protein.

1. Copy the sequence (remembering not to copy the first line)
2. Go to the RCSB PDB website: www.rcsb.org
3. Below the search bar, select the *Advanced Search* link

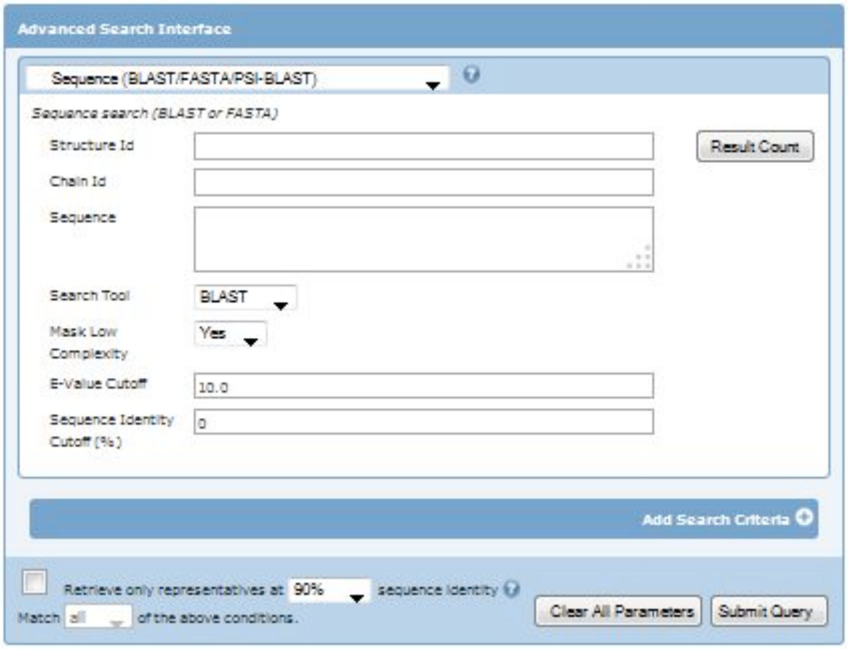

Search by PDB ID, author, macromolecule, sequence, or ligands
Go
Advanced Search | Browse by Annotations

(You can also get to the *Advanced Search* from the *Search* menu at the top of the page.)

4. In the advanced search window click on the *Choose a Query Type* and select the option labeled, *Sequence (Blast/Fasta)*


Advanced Search Interface
Choose a Query Type:
Link records
Molecular Weight (Structure)
Secondary Structure Content
Secondary Structure Length
SCOP Classification Browser (opens popup)
CATH Classification Browser (opens popup)
Taxonomy Browser (opens popup)
Sequence Features
Sequence (BLAST/FASTA/PSI-BLAST)
Translated Nucleotide Sequence (BLASTX)
Sequence Motif
Chain Length
Genome Location Browser (opens popup)
Chemical Components
Chemical Name
Chemical ID(s)
InChI Descriptor
Chemical structure (SMILES)
Molecular Weight (Chemical component)
Chemical Formula
Sequence search (BLAST or FASTA)
Result Count
Add Search Criteria
dentity
Clear All Parameters
Submit Query
Publications | Team | Careers | Usage & Privacy

This will cause the user interface to change to allow you to enter parameters for perform this search.



Advanced Search Interface

Sequence (BLAST/FASTA/PSI-BLAST)

Sequence search (BLAST or FASTA)

Structure Id:

Chain Id:

Sequence:

Search Tool: BLAST

Mask Low Complexity: Yes

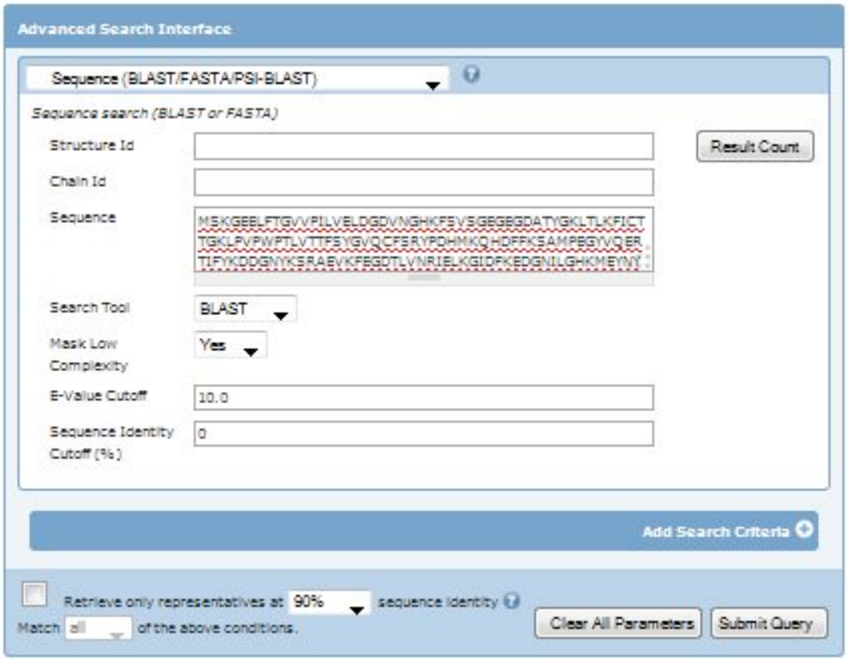
E-Value Cutoff: 10.0

Sequence Identity Cutoff (%): 0

□ Retrieve only representatives at 90% sequence identity

Match all of the above conditions.

5. Click inside the box titled *Sequence* and paste your sequence from the translation tool. It should look like the following:



Advanced Search Interface

Sequence (BLAST/FASTA/PSI-BLAST)

Sequence search (BLAST or FASTA)

Structure Id:

Chain Id:

Sequence:

```
MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICT
TGKLPVPWPTLVTTFSYGVQCFSRYPDHMKQHDFFKSAMPEGYVQER
TIFYKDDGNYKSRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKMEYNY
```

Search Tool: BLAST

Mask Low Complexity: Yes

E-Value Cutoff: 10.0

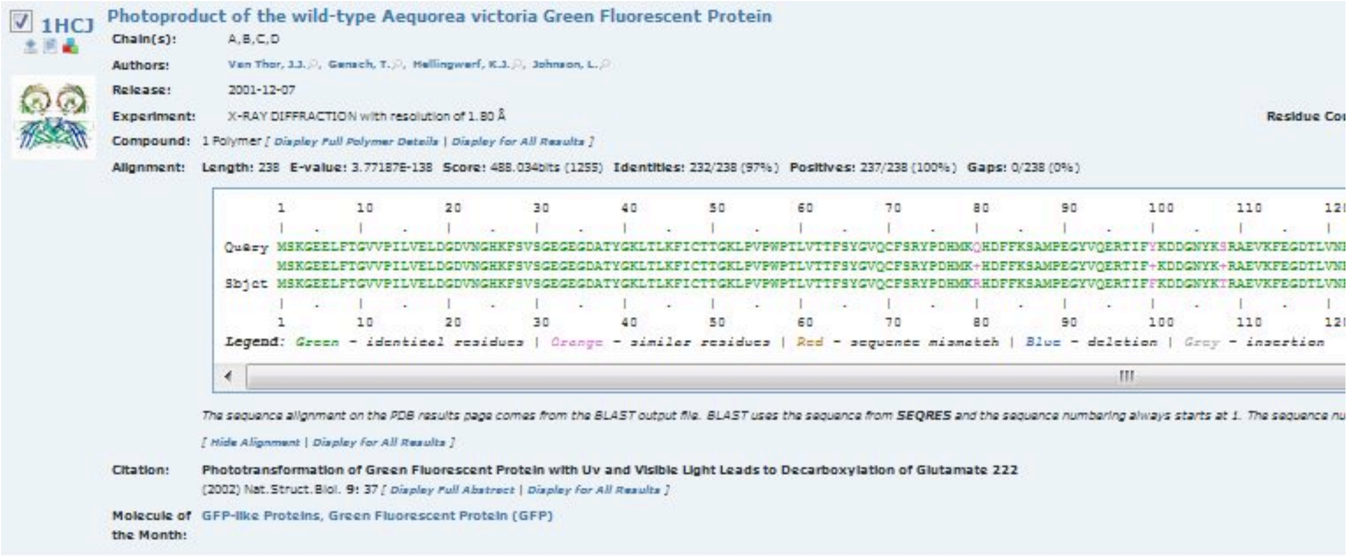
Sequence Identity Cutoff (%): 0

□ Retrieve only representatives at 90% sequence identity

Match all of the above conditions.

6. Now click on the *Submit Query* button on the bottom right corner.

The results page will have a list of proteins in the PDB that closely match the sequence you entered. You can see the similarity by looking at the sequence alignment viewer for each structure:



☑ 1HCJ **Photoproduct of the wild-type Aequorea victoria Green Fluorescent Protein**

Chain(s): A,B,C,D

Authors: Van Thor, J.J., Gensch, T., Hellingwerf, K.J., Johnson, L.

Release: 2001-12-07

Experiment: X-RAY DIFFRACTION with resolution of 1.80 Å

Residue Co

Compound: 1 Polymer *[Display Full Polymer Details | Display for All Results]*

Alignment: **Length:** 238 **E-value:** 3.77187E-138 **Score:** 488.034bits (1255) **Identities:** 232/238 (97%) **Positives:** 237/238 (100%) **Gaps:** 0/238 (0%)

```
      1        10        20        30        40        50        60        70        80        90        100       110       12
      |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |
Query MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTFSYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFYKDDGNYKSRAEVKFEGDTLVN
      MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTFSYGVQCFSRYPDHMK+HDFFKSAMPEGYVQERTIF+KDDGNYK+RAEVKFEGDTLVN
Sbjct MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTFSYGVQCFSRYPDHMKRHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEVKFEGDTLVN
      |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |    .    |
      1        10        20        30        40        50        60        70        80        90        100       110       12
Legend: Green - identical residues | Orange - similar residues | Red - sequence mismatch | Blue - deletion | Grey - insertion
```

The sequence alignment on the PDB results page comes from the BLAST output file. BLAST uses the sequence from ***SEQRES*** *and the sequence numbering always starts at 1. The sequence nu*

[Hide Alignment | Display for All Results]

Citation: **Phototransformation of Green Fluorescent Protein with Uv and Visible Light Leads to Decarboxylation of Glutamate 222**
(2002) Nat.Struct.Biol. **9:** 37 *[Display Full Abstract | Display for All Results]*

Molecule of the Month: GFP-like Proteins, Green Fluorescent Protein (GFP)

You will need to select 'Display Full Alignment' in the 'Alignment' section of the entry to see this view. You can also select 'Display for All Results' to see this view for all resulting entries.

As you can see, there are many structures in the PDB that contain similar sequences. These include proteins that are very similar to GFP and mutant forms.

Look at the entry for 1hcj[3]. Scroll through the sequence alignment and notice that only five amino acids in the entire sequence that are different (highlighted in orange). Also notice that these changes are conservative—the amino acids in the two forms are similar, such as a change from hydrophobic isoleucine to hydrophobic valine. This makes sense since entry 1hcj is a variant of GFP with the same properties as the GFP used for the DNA sequence.

Now browse down the list until you find 1bfp[4](it is probably on the second page of entries). This is a mutant form created by researchers, which changes the color of fluorescence to blue. Notice, however, that it only takes a few small mutations to do this. In the next task, we'll look at these mutations.

[3] Structural analyses of DNA recognition by the AML1/Runx-1 Runt domain and its allosteric control by CBFbeta. Tahirov, T.H., Inoue-Bungo, T., Morii, H., Fujikawa, A., Sasaki, M., Kimura, K., Shiina, M., Sato, K., Kumasaka, T., Yamamoto, M., Ishii, S., Ogata, K. (2001) Cell(Cambridge,Mass.) 104: 755

[4] Crystal structure and photodynamic behavior of the blue emission variant Y66H/Y145F of green fluorescent protein. Wachter, R.M., King, B.A., Heim, R., Kallio, K., Tsien, R.Y., Boxer, S.G., Remington, S.J. (1997) Biochemistry 36: 9759-9765

Task 4: Visualize genetic mutations in 3D

The following table shows the point mutations that are necessary for the different colors of green fluorescent protein:

Green Fluorescent	No mutation
Yellow Fluorescent	S65G, S72A, T203F
Cyan Fluorescent	Y66W
Blue Fluorescent	Y66H, Y145F

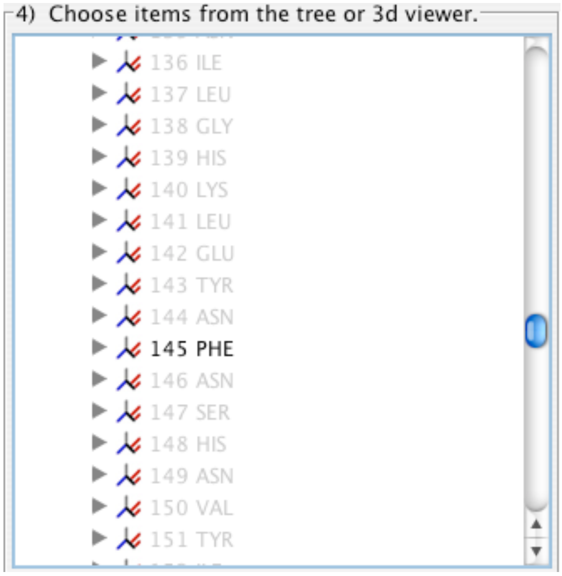
where "S65G" is a shorthand notation that says that position 65 on the protein sequence was changed from serine to glycine, etc.

1. In the structural alignment for entry 1bf3 (found in task 3), find the mutations Y66H and Y145F.

To visualize the point mutations for the blue fluorescent protein found in task 3, we can use Protein Workshop:

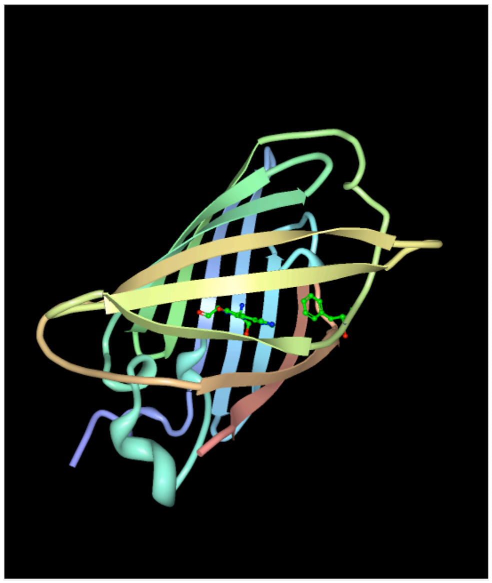
2. Click on the image or the title for entry 1bfp
3. Click on the link: *Protein Workshop* located under the structure image.
4. On the control panel on the right follow the 1-2-3-4 step process as described earlier for 1ema:
   Select *Visibility*
   Select *Atoms and Bonds*
   Select the amino acid labeled 145 PHE
5. If you do a similar thing for 66 HIS, you'll see that it is already turned on. Toggle it on and off by clicking on the label, and notice that it is part of the chromophore.

You should see something like the following in the viewer:



4) Choose items from the tree or 3d viewer.

136 ILE
137 LEU
138 GLY
139 HIS
140 LYS
141 LEU
142 GLU
143 TYR
144 ASN
145 PHE
146 ASN
147 SER
148 HIS
149 ASN
150 VAL
151 TYR



Notice that the modified amino acid, phenylalanine 145, is close to the chromophore.

You can learn more about proteins that are like Green Fluorescent Protein at:
http://www.rcsb.org/pdb/101/motm.do?momID=174

Appendix I

Letter codes for amino acids in a protein chain:

A	Alanine	Ala
C	Cysteine	Cys
D	Aspartic Acid	Asp
E	Glutamic Acid	Glu
F	Phenylalanine	Phe
G	Glycine	Gly
H	Histidine	His
I	Isoleucine	Ile
K	Lysine	Lys
L	Leucine	Leu
M	Methionine	Met
N	Asparagine	Asn
P	Proline	Pro
Q	Glutamine	Gln
R	Arginine	Arg
S	Serine	Ser
T	Threonine	Thr
V	Valine	Val
W	Tryptophan	Trp
Y	Tyrosine	Tyr

References

(a) Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. (2000) The Protein Data Bank. *Nucleic Acids Research* 28: 235-242.
(b) Bairoch A., Apweiler R., Wu C.H., Barker W.C., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin M.J., Natale D.A., O'Donovan C., Redaschi N., Yeh L.S. (2005) The Universal Protein Resource (UniProt) *Nucleic Acids Res*. 33: D154-159.
(c) Gasteiger E., Gattiker A., Hoogland C., Ivanyi I., Appel R.D., Bairoch A. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res*. 31:3784-3788.

Acknowledgements

This tutorial was written by Jeff Milton, Monica Sekharan, Christine Zardecki, Rachel Kramer Green, Shuchismita Dutta, and David Goodsell for the RCSB PDB. Last update, January 20, 2015.