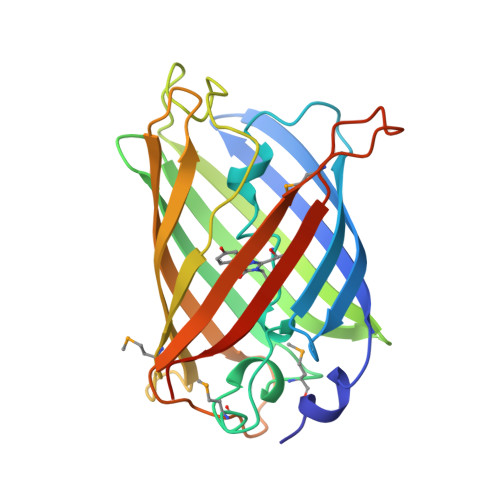


**info@rcsb.org • www.rcsb.org**

**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](http://www.wwpdb.org/), the RCSB PDBa 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.

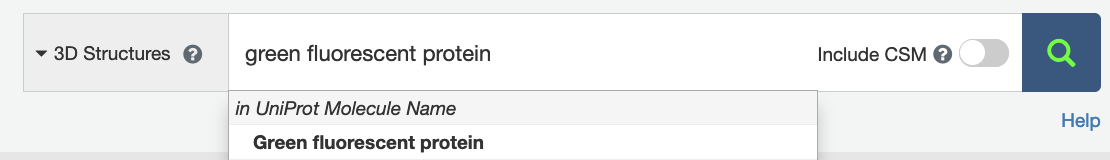
### Part I. Finding and Exploring the 3D Structure of Green Fluorescent Protein

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.

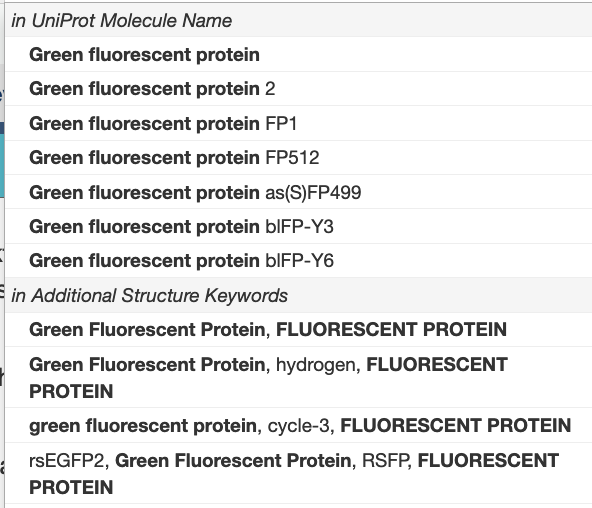
**Task 1: Find structures of green fluorescent protein at the RCSB PDB website.**

1. Go to the RCSB PDB website at <http://www.rcsb.org>

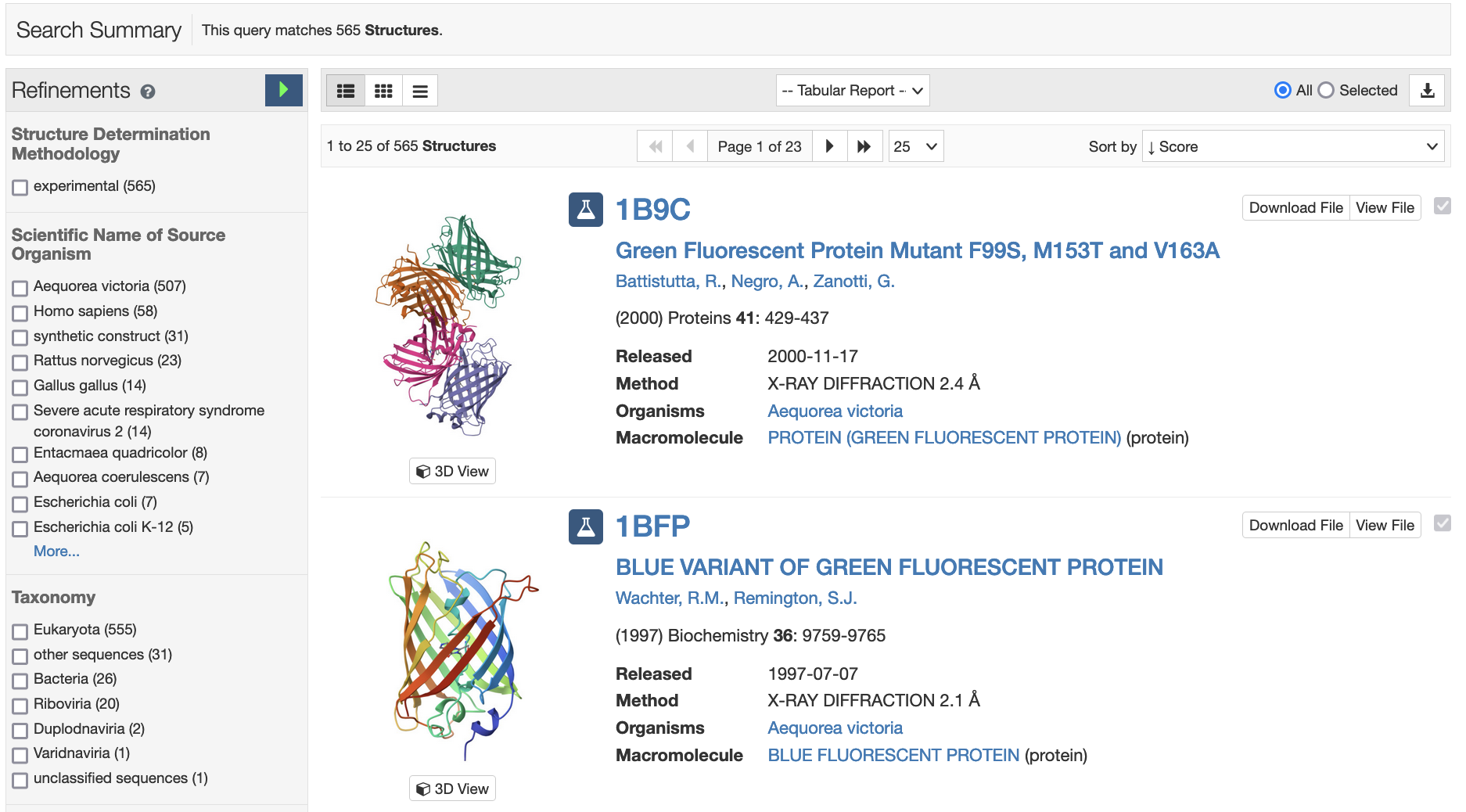
1. In the top search box type the keyword “green fluorescent protein”:



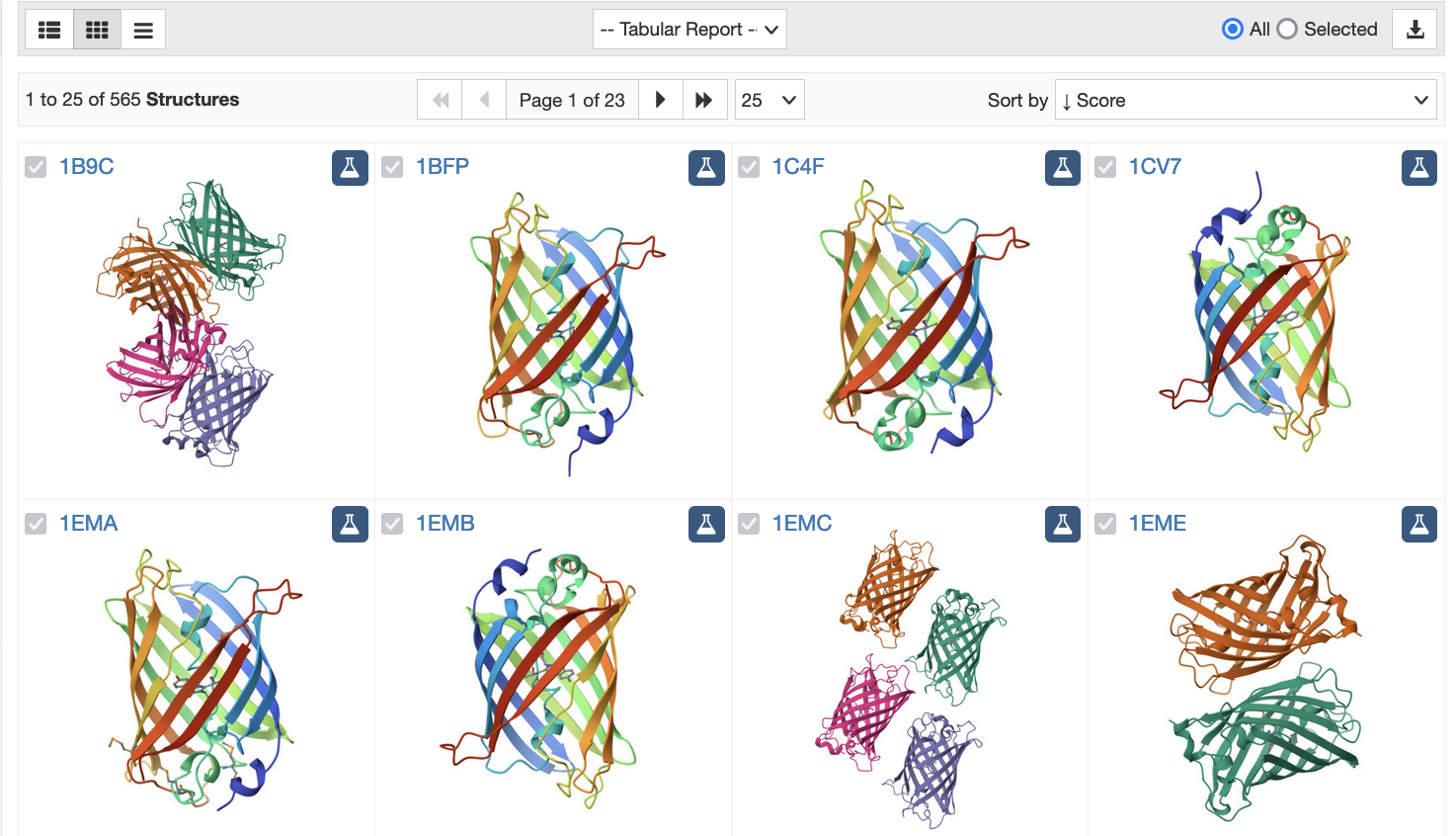
1. From the autocomplete list options that appear, select the UniProt Molecule Name “Green fluorescent protein”



1. Click on the green magnifying glass (search icon) to launch the search.
2. The result page will contain a list of proteins related to GFP.



1. You can explore all of these different structures by clicking on different examples, creating reports, or generating a gallery image.



Example image collage for some of the structures found when the search was run for **green fluorescent protein** in September 2023

**Q1.** How many green fluorescent protein structures did you find in the archive? How did you figure this out?

Ans:

Teaching Note:

* If you do not select UniProt Molecule Name in the autocomplete options and select another option or none of the options you will get a different set of results.

You can learn more about Green Fluorescent Protein in the Molecule of the Month feature on GFP at: <http://pdb101.rcsb.org/motm/42>. If you want to try building your own model of GFP, there is a paper cut-out form at: <http://pdb101.rcsb.org/learn/resource/green-fluorescent-protein-gfp-activity-page>. Or, from the PDB101 home page (<http://pdb101.rcsb.org/>), select ‘Paper Models’ from the learn menu.

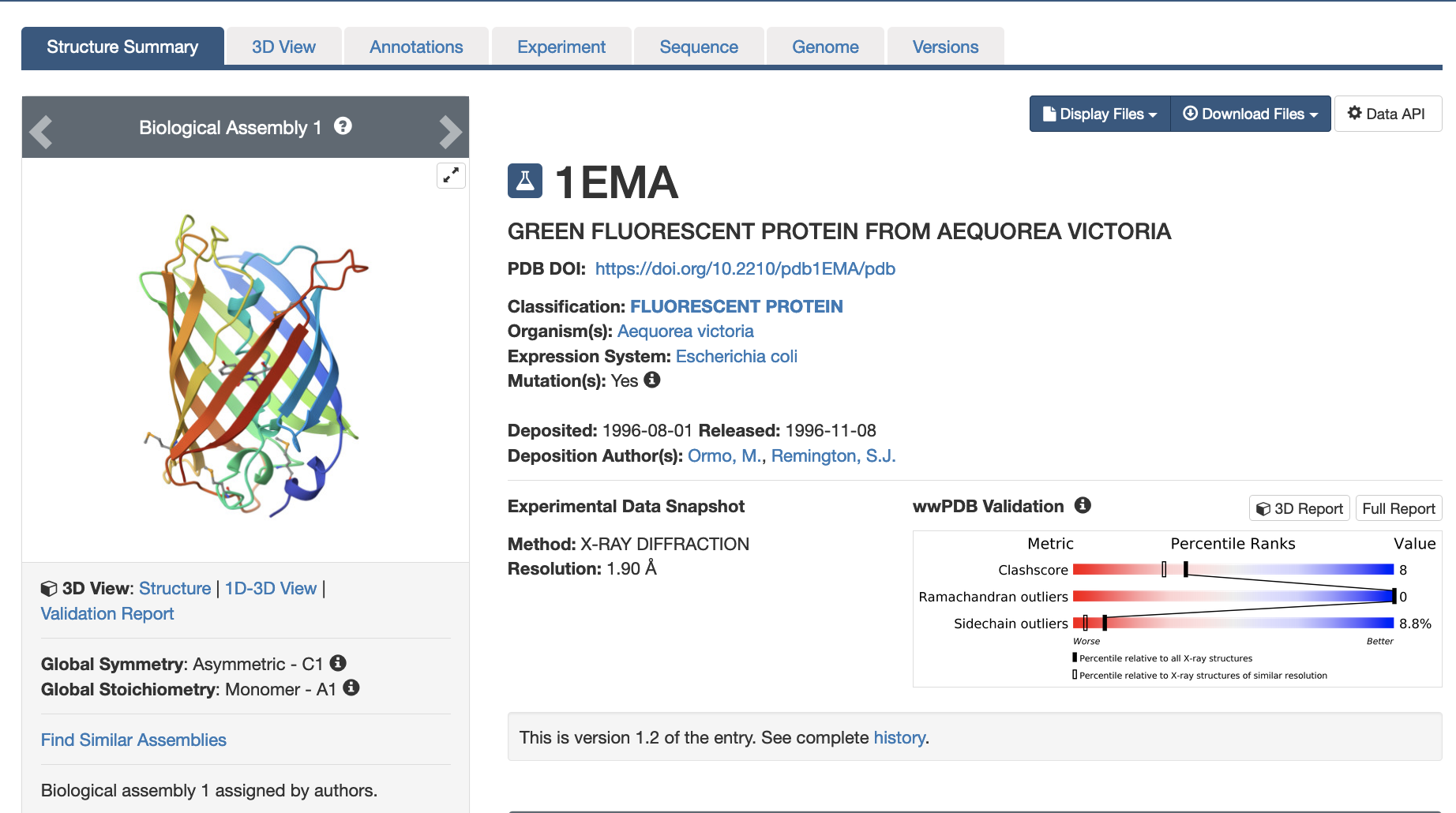
**Task 2: Explore the structure of a specific green fluorescent protein structure**

For the next part of the exercise, we will find the a green fluorescent protein that was taken from the jellyfish *Aequorea victoria* with PDB ID 1ema[[1]](#footnote-0). You can easily find this protein by entering the PDB ID 1ema in the search bar at the top of the page. This will take you to the Structure Summary page for 1ema.

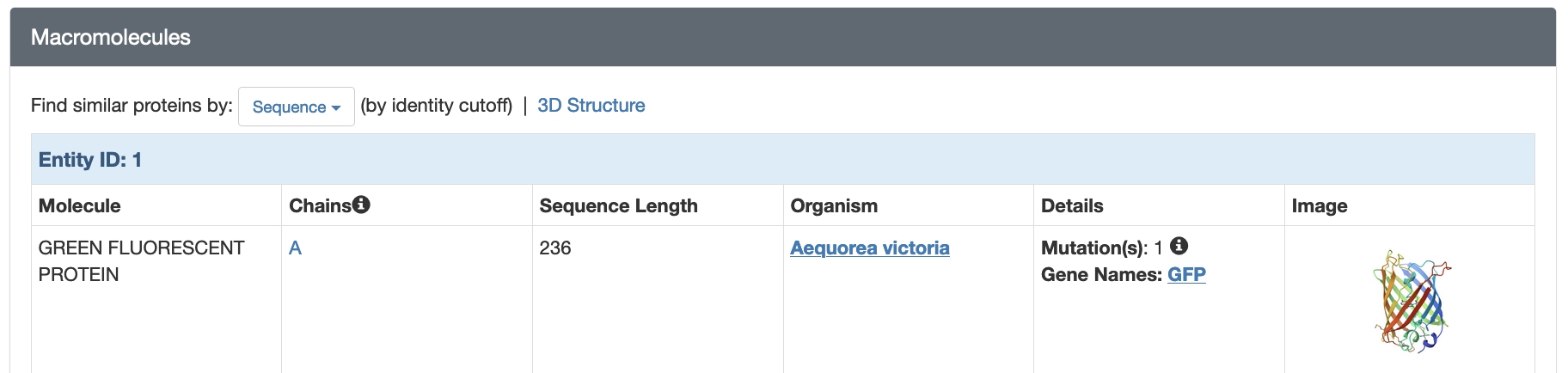


Now let’s take a look at the GFP structure in close detail.

1. You should still be on the Structure Summary Page for entry 1ema.

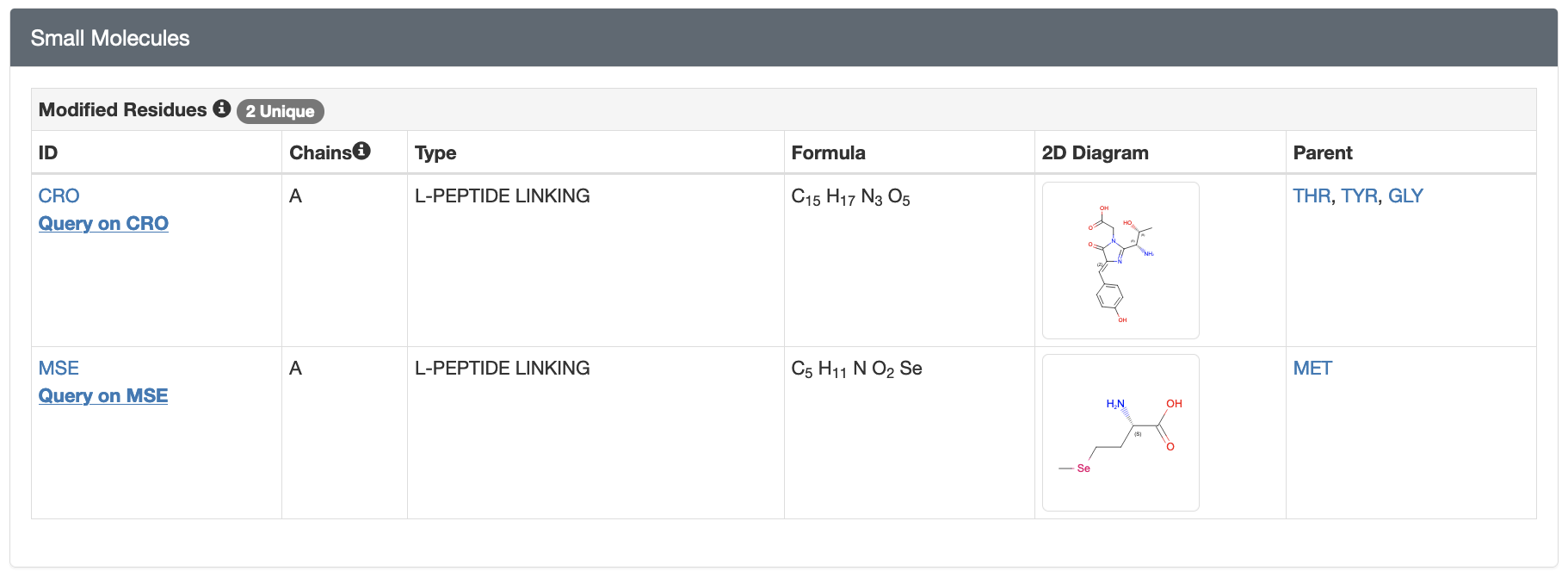


1. On the left side of the Structure Summary page, you will see a box containing an image and links to 3-D molecular viewers.
2. Review the contents of this structure by scrolling down the page. Information about components in this structure is presented in sections titled Macromolecules and Small Molecules, shown below



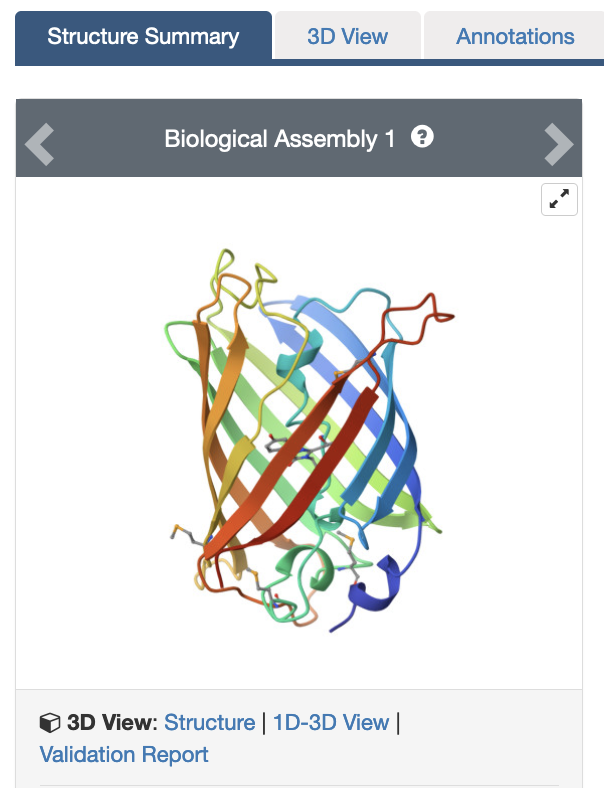
To learn more about the overall GFP protein sequence and its various annotations, click on the Sequence tab at the top of the structure summary page.

1. To learn more about the *Small Molecules* associated with the protein click on the hyperlinks and buttons shown above.

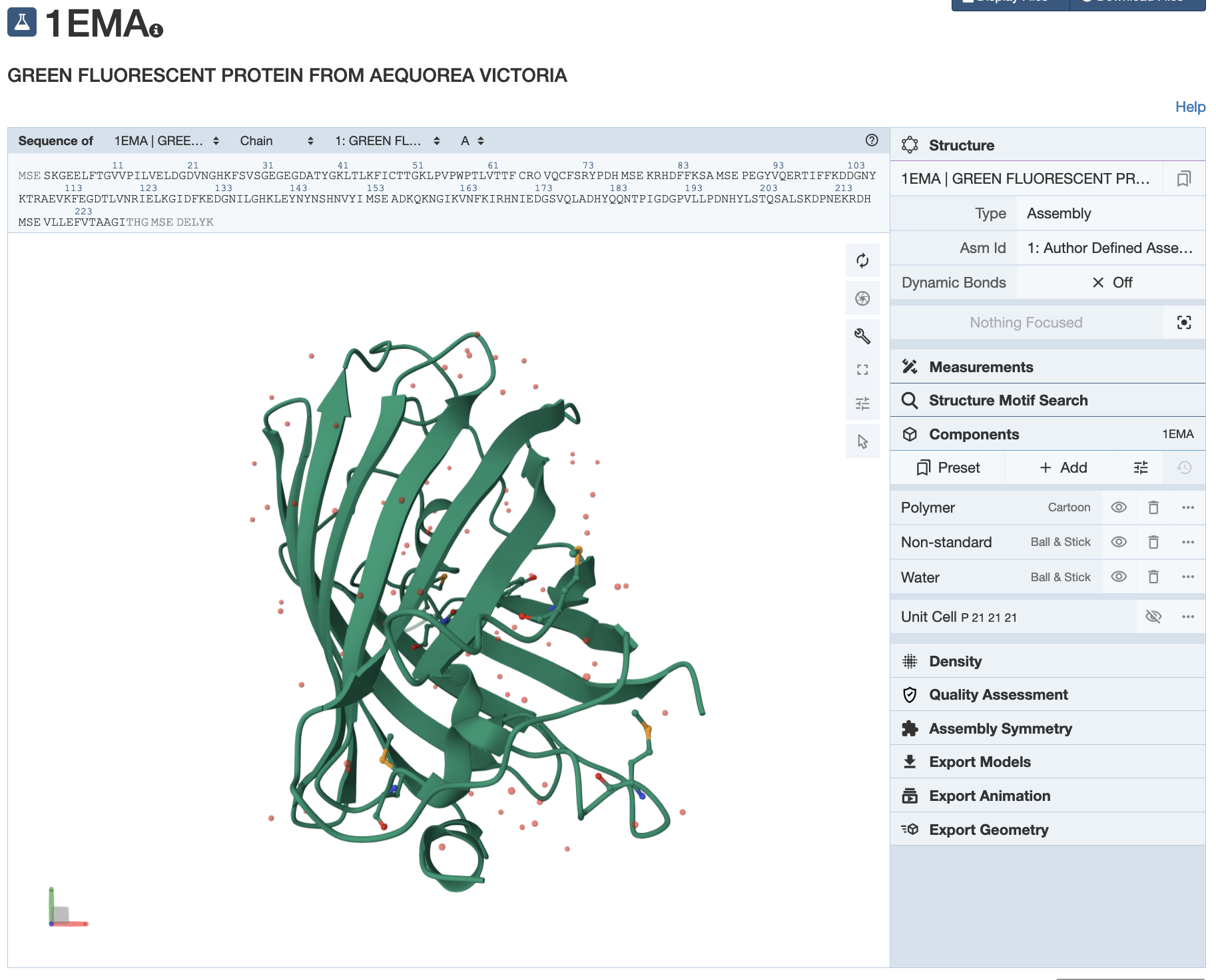


**Task3: Visualize the structure of green fluorescent protein**

1. Click on the 3D Structure tab at the top of the page or on the Structure hyperlink below the snapshot of the structure.



1. This opens the structure in Mol\*



The top of the display shows the sequence of the polymers and is called the sequence panel. The white space showing the 3D structure of the protein is called the 3D canvas and the blue panel on the right is the controls panel for the display.

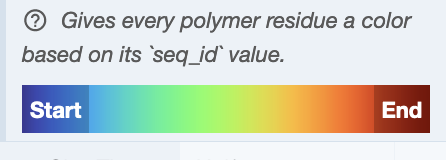
1. Color the structure by secondary structure using the following steps:
   1. In the Components section of the Controls panel, click on the 3 dots next to the Polymer to open options
   2. Select the options Set coloring >> Residue properties >> secondary structure to color the cartoon representation of the structure.

**Q2.** What secondary structural elements can you identify in the structure? List how many of these elements you see in each polymer chain. Save an image using the camera lens icon in the vertical menu on the 3D canvas and include it below to support your answer.

Ans:

1. Repeat the steps for coloring the cartoon representation of the structure but this time color by Sequence ID (instead of the Secondary structure).

In this color scheme the N-terminal residues are colored in blue while the C-terminal residues are colored in red.



**Q3.** Identify the N-terminal and C-terminal secondary structural elements. Support your answer with a suitable figure.

Ans.

**Task 4: Explore the chromophore and its interactions with the green fluorescent protein.**

To explore the business end of the green fluorescent protein (GFP) in atomic detail use the following steps.

1. Click on the ligand (CRO) displayed in ball and stick representation at the center of the GFP protein. This should zoom in on this part of the structure, center on it and display the non-covalent interaction in the neighborhood.

**Q4.** Identify 3 different amino acid residues that non-covalently interact with CRO. List their name and type of interaction they form. Support your answer with a suitable figure.

Ans.

Note:

The amino acid residue numbers can be read off from the bottom right corner of the 3D canvas. Note that in this case 2 sets of numbers are listed. The first one is the internal PDB numbering while the second one is what matches the paper in the literature. For example E222 will show up as

Part II. Gene and Protein Sequences of Green Fluorescent Protein

In this second part, we will use several online resources to do the following:

1. Find the sequence of the gene for green fluorescent protein referenced in the paper describing the structure (PDB ID 1ema).
2. Translate the mRNA into a protein sequence.
3. Compare that protein sequence with that in UniProt
4. Find mutant forms of the protein with altered function.

**Task 1: Find the sequence of the gene for GFP gene (referenced in PDB ID 1ema).**

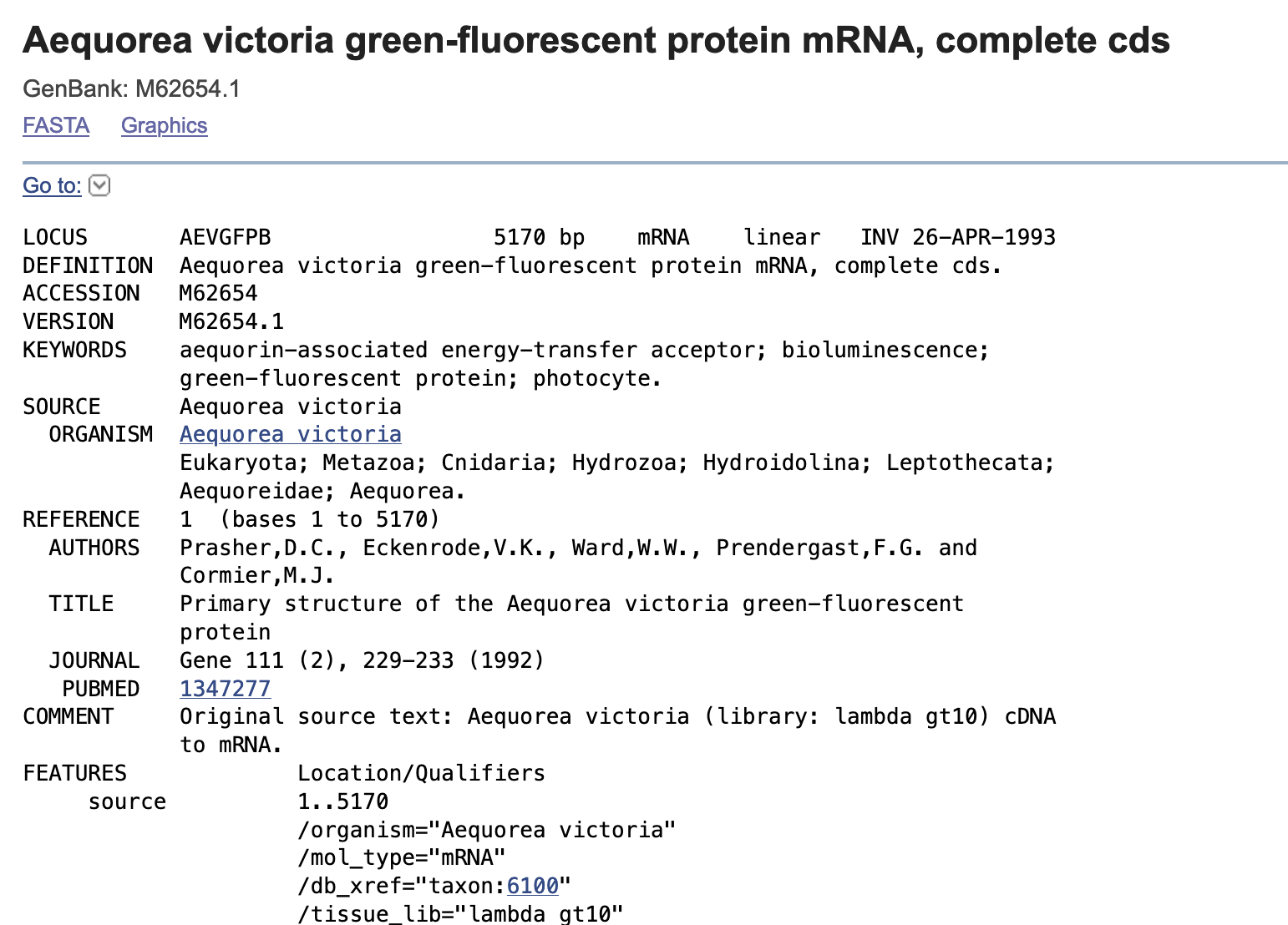
1. Point your web browser to GenBank: <https://www.ncbi.nlm.nih.gov/genbank/>

Alternatively, go to the NCBI website: <http://www.ncbi.nlm.nih.gov/> and select the *Nucleotide* database in the *search* pulldown menu to access this same page.

Teaching Notes:

You can search for the gene by typing the name of the protein “Green fluorescent protein” in the top search box. However, this is likely to yield many results so to simplify the steps here the GenBank identifier of the gene is provided here.

1. Search for the gene sequence of green fluorescent protein with the GenBank ID: M62654.1 by typing it in the top search box.
2. Open the page and explore the contents

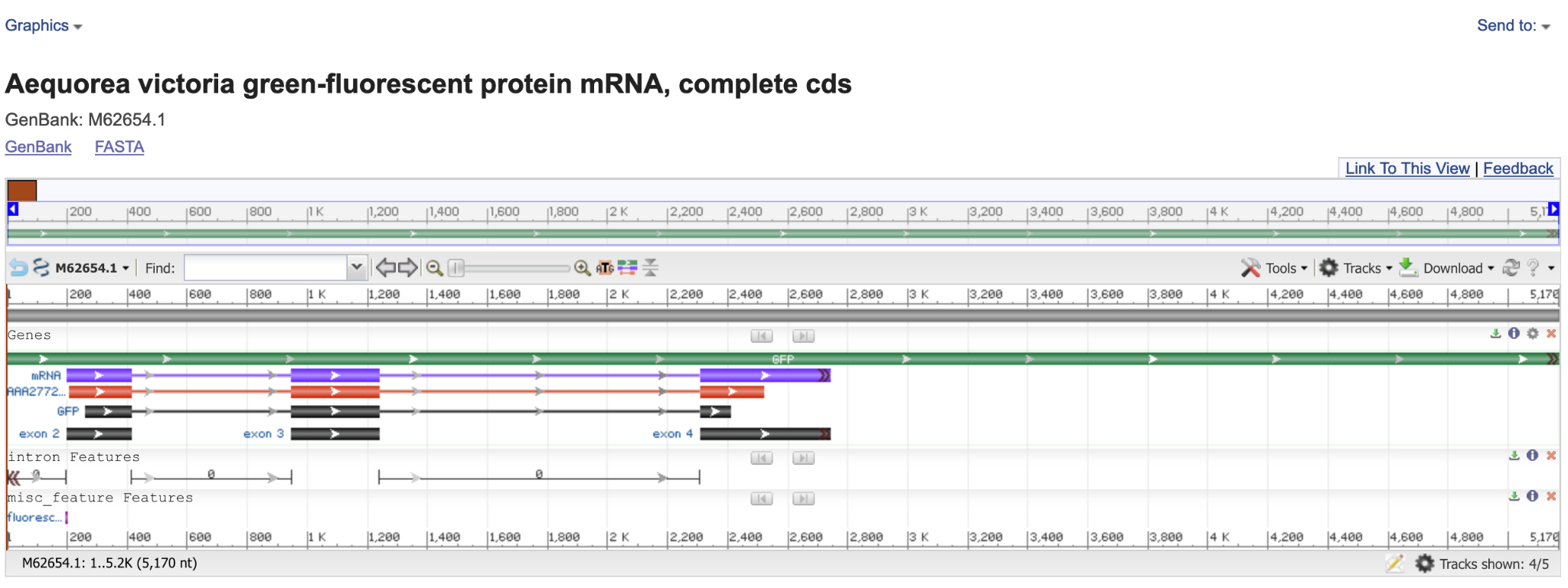


1. Now explore the features in the gene sequence and answer the following questions



Note that the coding sequence is described as a join between the regions 208 to 413, 946 to 1240, and 2308 to 2523)

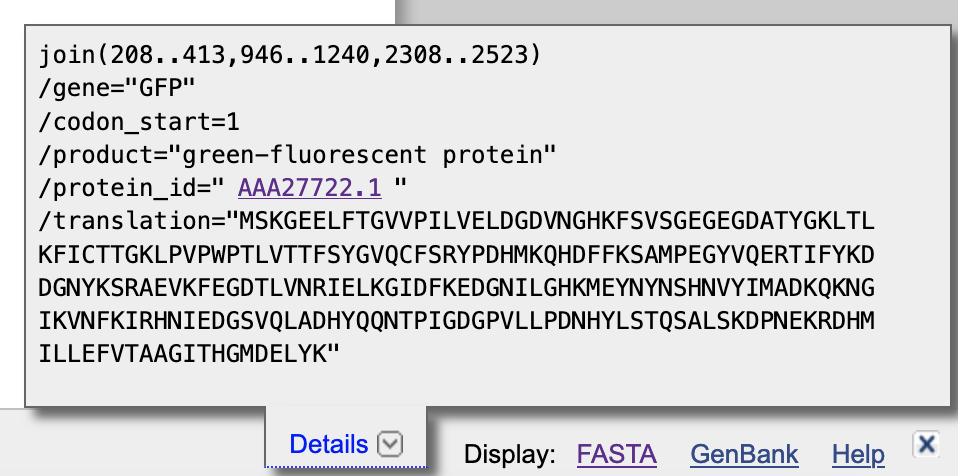
1. Examine the graphical representation of the gene by clicking on the *Graphics* hyperlink on the top of the page.



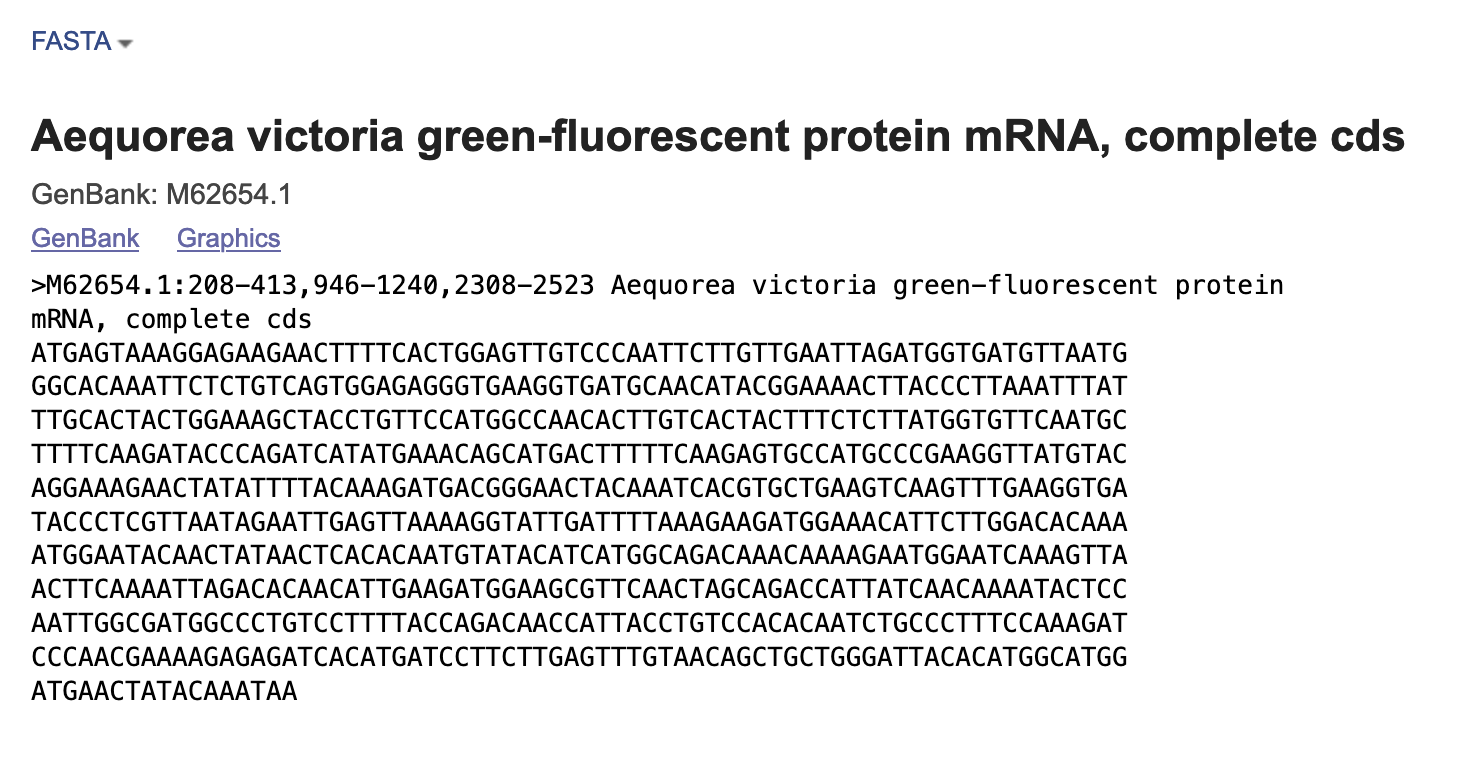
The orange/red track in the middle of the page represents the coding region of the gene. Note that there are 3 exons (denoted by thick lines) and three intron - one in the beginning (not shown) and two shown in between the exons (as a thin line).

Note: 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.

1. Save the FASTA sequence for the CDS by clicking on the FASTA hyperlink at the bottom of the page



1. This opens a page that shows the FASTA format sequence for the CDS



1. Copy and save this sequence.

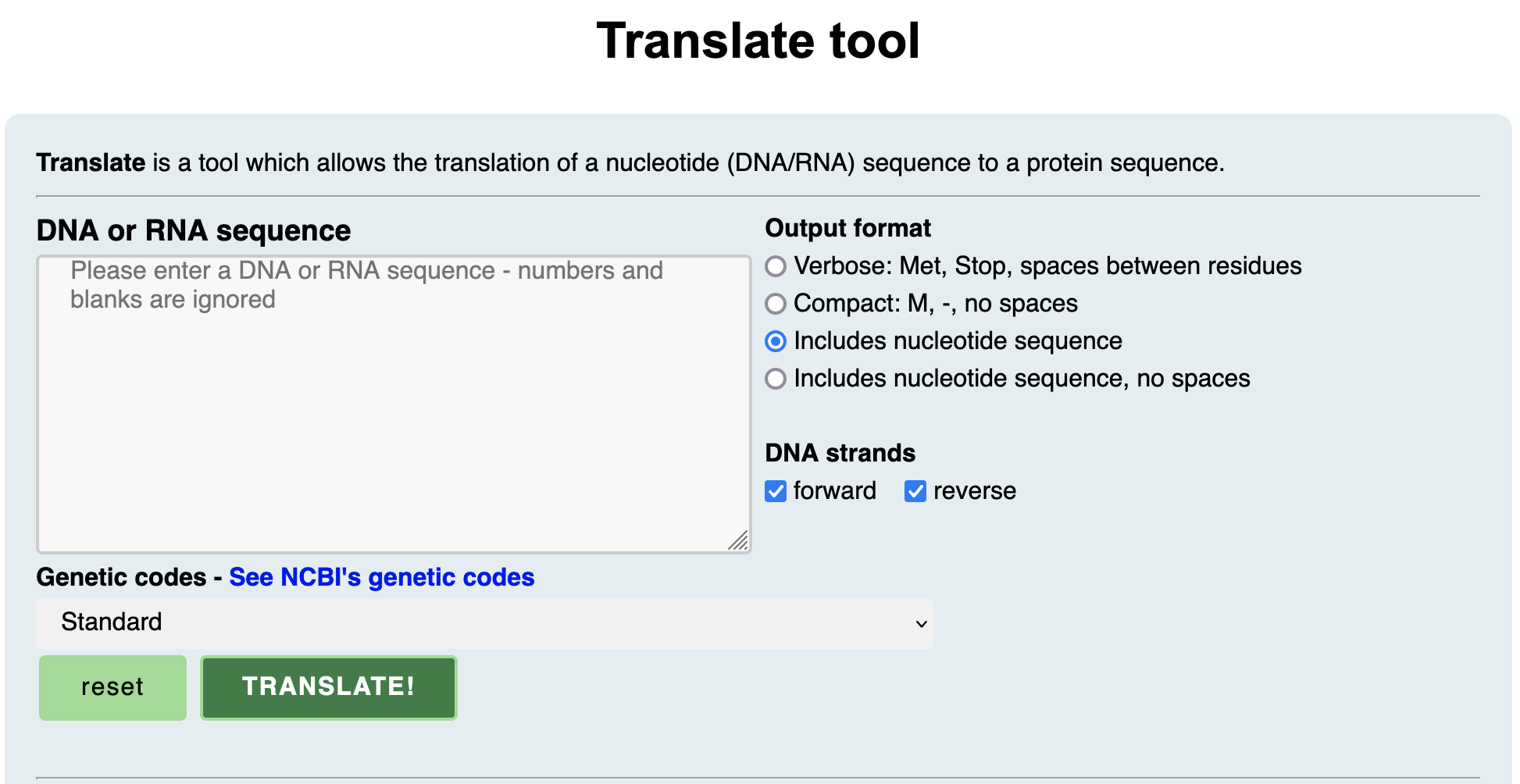
Teaching Note:

* The translated amino acid sequence for the cds is displayed in the screenshot above. In the next step we will translate the cds sequence. You can use the amino acid sequence listed above to make sure that the correct translation frame was used.
* You can also skip the translation step and use the translated amino acid sequence presented here.

**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.

1. Point your browser to Expasy Bioinformatics Resource Portal - <http://www.expasy.org/tools/>
2. Search for the *Translate* tool on the page and click on it. Paste the sequence that you selected in 7 above in the box as shown below. Click in the text window and paste your sequence using Edit->Paste from the file where you saved the sequences. Be careful that you paste only the FASTA sequence of the CDS.
3. Your browser should look like the following:



| **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*. |
| --- |

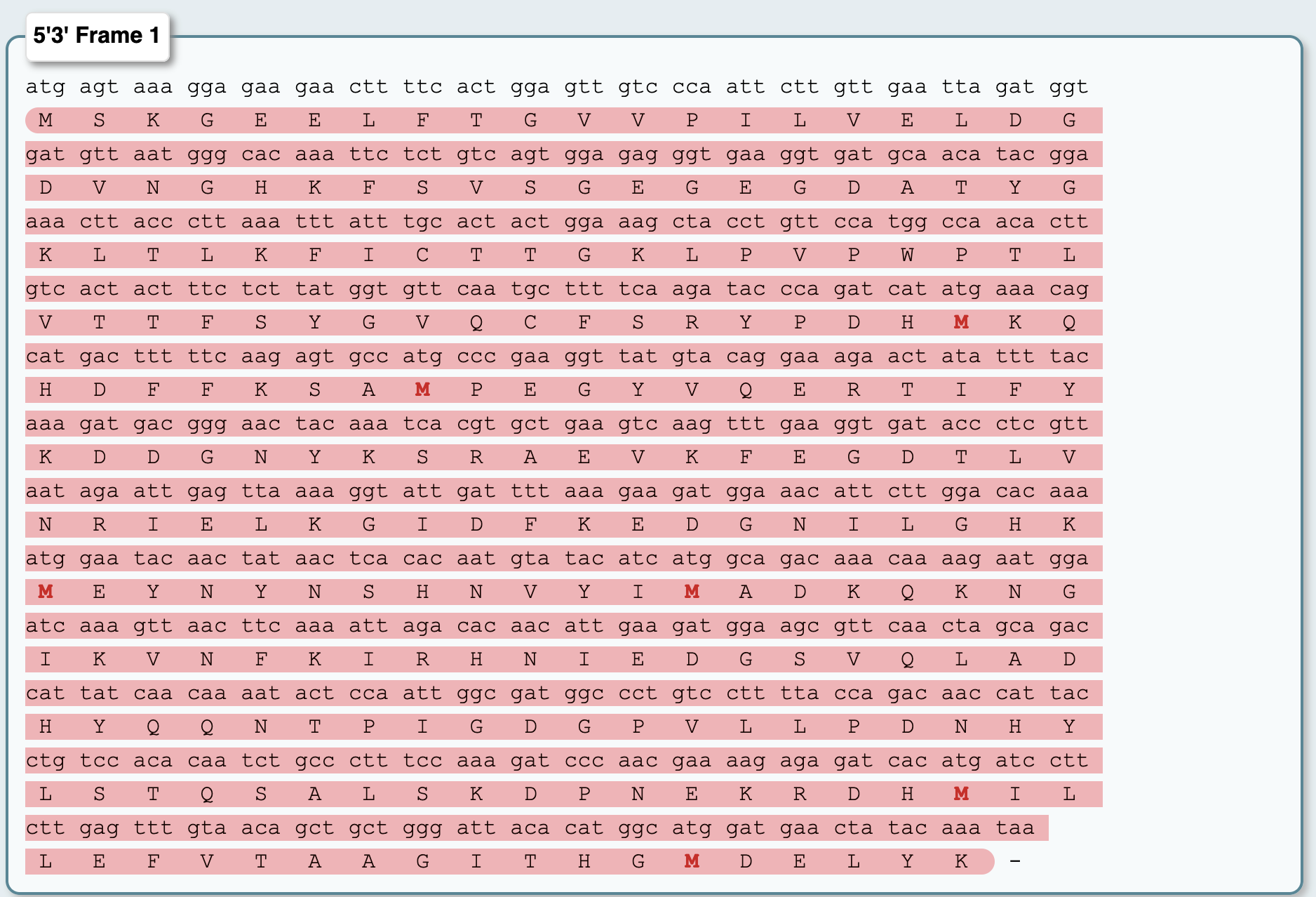
5. Change the *Output format* choice to “*Includes Nucleotide Sequence”*

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 these 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 1* appears to generate the best translation. Each line contains the DNA sequence and highlights the three-letter codon along with the corresponding amino acid.

The start codon is AUG (or in the DNA case it is ATG) and ends with UAA (or in the DNA case it is TAA).



You can also rerun this translation with the Compact: M, -, no spaces mode.

Compare this sequence with that of the translated protein sequence in GenBank. Are they the same?

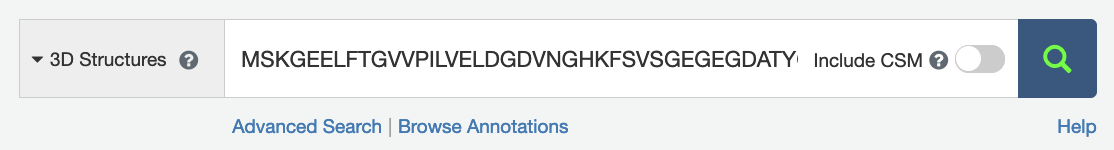
**Task 3: Find proteins with similar sequences at the PDB**

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

MSKGEELFTGVVPILVELDGDVNGHKFSVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTFSYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFYKDDGNYKSRAEVKFEGDTLVNRIELKGIDFKEDGNILGHKMEYNYNSHNVYIMADKQKNGIKVNFKIRHNIEDGSVQLADHYQQNTPIGDGPVLLPDNHYLSTQSALSKDPNEKRDHMILLEFVTAAGITHGMDELYK

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](http://www.rcsb.org)
3. Paste the sequence in the top search box and start the search.



1. 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.

**Q4.** List the PDB ID of the polymer chain that best matches the search. What is the sequence identity between this entry and the sequence in the query.

Ans.

**Task 4: Compare the structures of the best matching GFP protein with PDB ID 1bfp**[[2]](#footnote-1)

The PDB ID 1bfp is a mutant form of the protein, created by researchers, which changes the color of fluorescence to blue).

1. Open the structure summary page for this PDB entry at <https://www.rcsb.org/structure/1BFP>

**Q5.** From reading the title list the mutations present in this entry.

Ans.

1. Once you visualize the structures in Mol\*, locate these mutations in the 3D structure by clicking on them in the sequence panel (on the top of the page). Explore their interactions with neighboring residues and save images to include here.

**Q6.** For each of the mutations list 2 amino acids that form noncovalent interactions with the residue. Support your answer with suitable figures.

Ans.

**Task 5: Compare the structures of best match that you identified in Q4 and PDB ID 1bfp.**

1. Open the Pairwise Structure Alignment tool at <https://www.rcsb.org/alignment>. Add the PDB IDs of the structure in Ans 4 and 1bfp in the boxes provided on the page. Run the structure comparison with the default options.

**Q7.** How closely are these structures related? List the root mean squared deviation (rmsd) and %identity. Include a figure showing the structure superposition.

Ans.

**Q8.** What does the structure alignment result tell you about the structural and corresponding functional changes in the engineered protein.

Ans.

**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. November 2015.

This tutorial was last updated by Shuchismita Dutta in September 2023.

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 [↑](#footnote-ref-0)
2. 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 [↑](#footnote-ref-1)