Fluorescent Protein Discovery Project

Analysis of *RFP* Sequence Data

Previously, bacterial colonies expressing new fluorescent proteins were isolated and colony PCR was used to amplify *RFP*. This week we will compare the DNA sequence from these promising new *RFP* genes to the original un-mutagenized *RFP* and investigate which mutations are potentially responsible for the shift in fluorescence.

Week 13 +

- Streak single colonies showing changes in fluorescence
- Amplify selected *RFP* through colony PCR
- Analyze DNA sequences

Sanger Sequencing

DNA sequencing is any process that can determine and report the order of nucleotides in a particular DNA strand. Genewiz, the company that will sequence our *RFP* genes, uses a technique called Sanger sequencing. This is a first-generation method that was once the gold standard of sequencing techniques, and remains one of the most reliable and widely used today. It has been unseated to some degree by Next Generation Sequencing (NGS), which is able to more accurately and efficiently handle long DNA sequences, but it is still widely used for quick short DNA fragments.

Sanger sequencing resembles PCR in many ways. The first steps involves mixing the DNA template with primers that flank the region to be sequenced, dNTPs, and DNA polymerase.

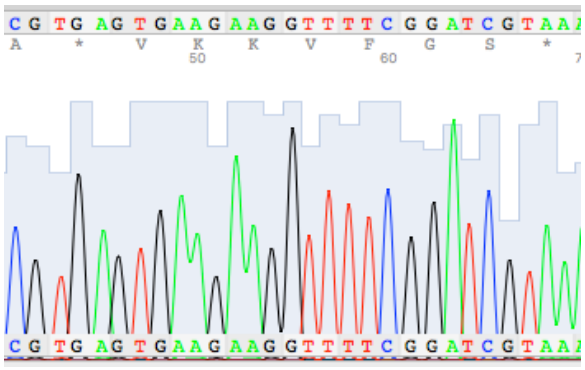
The novel part is the addition of modified dNTPs, called ddNTPs (**DI-**deoxynucleosidetriphosphates), which have had the hydroxyl group removed from the 3' end, making them ineffective at forming phosphodiester bonds with the next nucleotide in the sequence. In essence, this "gums up" the extension process and terminates elongation, as DNA polymerase is functionally unable to add additional nucleotides. A fluorescent dye it then used to label each ddNTP allowing identification of the base responsible for halting the extension process.

The result of the chain termination process is a series of sequence fragments of different lengths, terminated at the 3' end by a tagged ddNTP, which can be read by a fluorescence detector. The fragments are fed into a capillary electrophoresis system consisting of a thin tube filled with a gel polymer environment. Much like gel electrophoresis, a current is applied and the DNA fragments separate themselves out by size as they move towards the positive cathode.

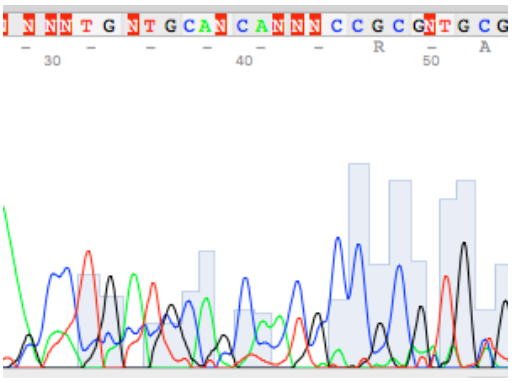
At the end of the capillary system, a laser excites the fluorescently labeled base in the order that the fragments exit, creating the sequence of interest.

The output of this sequence analysis is an electropherogram, showing the level of absorbance as each tagged nucleotide passed the detector, as well as bars for the confidence in that base call. A good electropherogram will have clean peaks with little to no "noise" in the calls, and few missing bases. A less effective result will include static-like noise, missing base calls, and an overall lack of confidence in sequence integrity.

Example of good electropherogram



Example of bad electropherogram

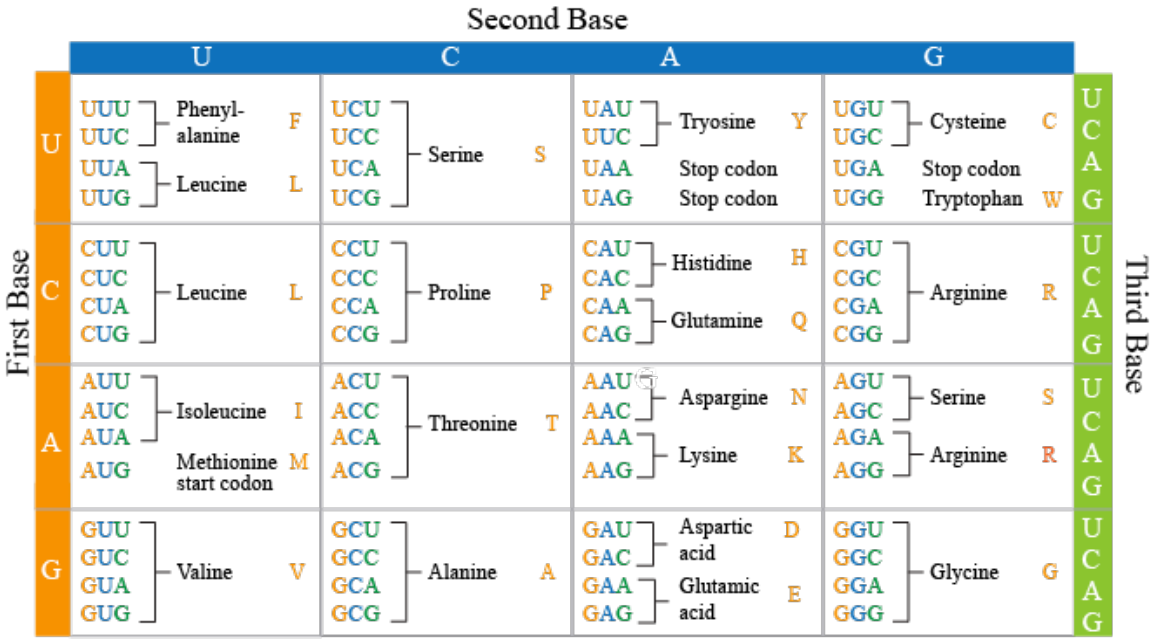


This bad example is typical of a low-confidence sequence result. The muddying of the noise makes it impossible to distinguish the called bases with any confidence. Note that when the

software is unable to determine which base should be placed in a position, it fills the gap with an N. Some other common problems with Sanger sequence results include miscalls related to gaps between nucleotides (i.e. G-A dinucleotides, which have a longer gap), double peaks, and loss of resolution later in the gel as the fragments get longer.

The Central Dogma – DNA to RNA to Amino Acid sequence

When changes occur in the DNA sequence of a protein-coding region, it may or may not result in changes to the amino acid sequence. Recall that DNA is transcribed into RNA, where thymine is replaced by the base uracil, before moving to the cytoplasm of the cell. Each set of three RNA bases creates a codon that signals the ribosome to incorporate a different amino acid into the protein chain during translation.



Changes in the DNA sequence which do not change the amino acid incorporated, are called **synonymous substitutions.** Typically, these are bases in the 3rd spot of the codon, which are interchangeable to a degree. Conversely, a change to the 2nd base in a codon will almost always change the amino acid. This is known as a **non-synonymous substitution.**

Insertion of a different amino acid could have little to no effect on a protein, or it could completely change or stop its functionality. Consider the situation where a protein consisting of 100 amino acids has a G to A mutation changing UGG tryptophan to UGA, a stop codon. This mutation will mechanistically cleaved off the remaining 94 amino acids in the protein sequence, rendering it non-functional. Even without the inclusion of stop codons, different amino acids

have different side chains which will interact differently with each other and with the environment. Side chain interactions are critical for structural conformation of proteins including the exposing/hiding of active sites. Again, a single significant amino acid chain change can completely altered the function or fortitude of a protein. However, due to the redundancy of codons that encode specific amino acids, a single nucleotide difference might not change anything at all.

Red fluorescent protein secondary and tertiary structure.
http://www.rcsb.org/pdb/explore/jmol.do?structureId=2VAD&bionumber=1

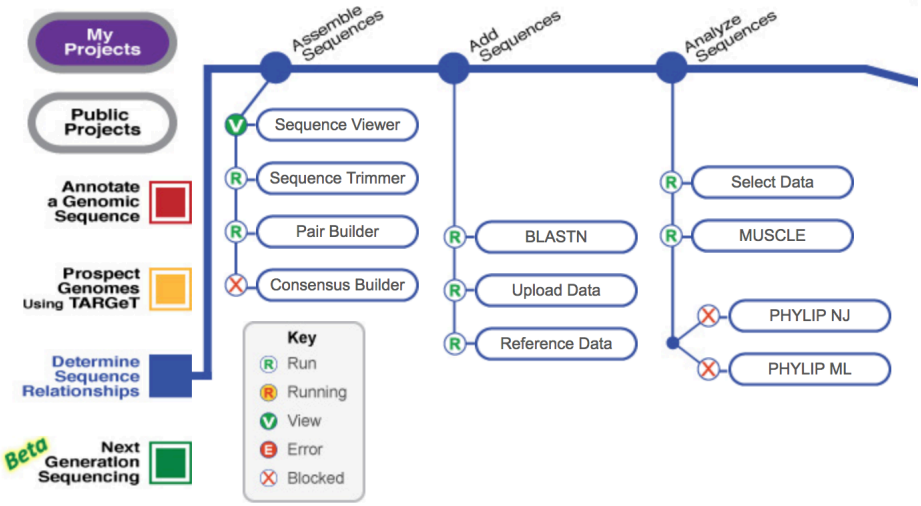
Sequence Analysis Guide

Note that there are lots of ways to analyze sequence data, and many research groups have created software to work with this type of data. Below is one curated route that will get you started to address your research question. If you have a research question that these techniques can't address, ask your TA for help to explore other bioinformatics algorithms.

1. **Reading the data.** Sequence data is available on the course website as two filetypes:
    a. sequence files in .ab1 format with interactive electropherograms (the .ab1 filetype is uploadable into other application for visualization and analysis)
    b. electropherograms in a PDF (created from the .ab1 files using the Mac platform software 4Peaks at https://nucleobytes.com/4peaks/index.html. It is not necessary for you to download this software for this course.)

2. **Assess data quality.** Using the electropherogram PDFs, describe the quality of the sequences: how much noise is present? Are there any portions of your sequence missing or that appear to be mis-called? How high is your confidence in the sequence calls? Are there regions of the sequence of lower quality?

3. **Align forward and reverse sequences for identical samples.** Navigate to DNA Subway (https://dnasubway.cyverse.org/) and create a free account. On their homepage, select Determine Sequence Relationships. Select project type "DNA" and choose files to upload ABI1 trace files. Use Ctrl+ or Alt+ mouse click to select all files you wish to compare and upload (we recommended a maximum of 12 files, or 6 unique samples, for your first use). Name your project and continue. The screen capture below shows your options in DNA Subway, which is a GUI that can push your data out to different sequence analysis sites. We'll use DNA Subway to **Assemble Sequences**, including view, trim, pair, and build consensus FASTA files.



a. For each forward/reverse pair follow the instructions below to view, align, pair, and build a consensus. Once you have a satisfactory consensus sequence, select it with your mouse and save it as a txt file in notepad or text edit.
   i. Click Sequence Viewer and compare the sequences to your electropherogram PDFs. You'll see N's for ambiguous calls, etc. Note if any sequences are missing. No action is needed at this step, so after viewing you may close the window.
   ii. Click Sequence Trimmer. This application will trim off the low quality sequence calls at either end of your sequences, but remove nothing from the middle of the sequence. Once trimmed, you'll see a new version of your sequences with the statement "Your sequences have been trimmer." Visually inspect, then close the window.
   iii. Click Pair Builder. For each reverse sequence, select the hyperlinked "F" to convert it to an "R." Check the boxes next to two sequences you want to pair and response Yes to the "Pair them?" dialog box. Repeat for every pair, then Save your pairs.
   iv. Click Consensus Builder. Wait until the samples has processed (R converts to C), when click again to view. For each sample, view the consensus, decide whether to trim or not, then select and copy the name and consensus sequences into a text editor, formatted as a FASTA file like this:

```
>SEQUENCE_1Name
AGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTC
>SEQUENCE_2Name
AGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGT
```

   v. At this point, add each consensus to the class google sheet using the link shared in class.

4. **Check for similar sequences in the NCBI database.** Using the BLAST online interface (https://blast.ncbi.nlm.nih.gov/Blast.cgi), select Nucleotide BLAST, and paste your FASTA sequence(s) into the box and click BLAST. The algorithm will search a large database of sequence information and return a list of the most similar. This is an excellent way to check for contamination and identify off target PCR amplification.

5. **Sequence Similarity.** What sequences in the database are most similar? Does this align with your expectations? How are you assessing similarity?

6. **Mutagenesis changes to gene of interest.** Compare the sequence of your original and mutated *RFP* genes. On the BLAST homepage, locate the "Global Align" tool.
   a. Set your ancestral gene sequence as the reference, and a derived gene as the Query.

   b. Repeat this for each derived sample from that ancestor.

c. How do the sequences compare? What hypotheses can you form based on the differences between the two? Are they more or less similar than you expected? What kind of mutation occurred, if any?

d. Make a table of DNA changes for your notebook, and report any other important evidence for your research question.

7. **Mutagenesis impacts on the protein sequence.** We will use the program ExPASy to determine the most likely AA sequence given our DNA sequence data.

   a. Navigate to web.expasy.org/translate/, paste a nucleotide sequence, and click translate. The program will provide several options with different codon starting locations. The best option is the one with the largest open reading frame, that is, the largest number of amino acids (AAs) found between a start and stop codon.

   b. Keep these amino acid sequences for your records and navigate back to the BLAST Global Align program. Alternatively, you can use MUSCLE (https://www.ebi.ac.uk/Tools/msa/muscle/). Click the protein tab and set your original RFP sequence as the reference, and your mutant gene as the Query.

   c. Make a table of AA changes. Are some more likely than others to have contributed to functional differences between the proteins? Were there more or fewer AA changes than nucleotide changes between the sequences? Why do you think these numbers are different?

8. **Interpreting your results.** Discuss your results with your partner, peers, and TAs. Do you have the information you want to address your research question or hypothesis? Consider other comparisons you want or need to make and identify the data and applications you'll need. Then make those comparisons.