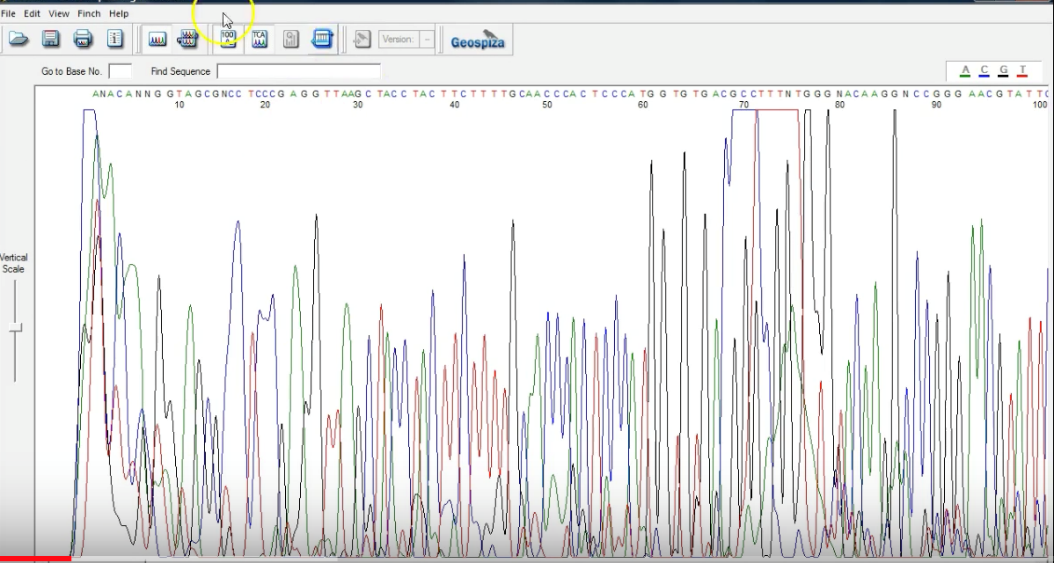
# Exercise 20: Bioinformatics

## introduction

Sequenced DNA yields a data file will show the signal strength for each of the four nucleotides at each nucleotide position of your sequenced DNA. So, how does this tell you the identity of your organism? The final steps involve using computer programs to analyze the data and then compare your sequence with a large database containing known DNA sequences. Learning about biology by using or searching large databases of biological data is often called **bioinformatics**. This is a rapidly growing area in science as it is at the intersection of biology, applied mathematics and computer science.

The first step is to look at your DNA sequence. You will use a program called FinchTV to visualize your DNA sequence. You download this program, and open your file. You will see something like this:



Each colored line represents one nucleotide. In this image, the red line shows thymine, the green adenine, the black guanine, and the blue cytosine. Ideally, at each position, one of these lines will be high forming a peak while the other lines are low. This is great and means that there is no ambiguity in that part of your sequence. Sometimes, lots of repetition of a certain nucleotide can complicate interpretation, or the signal from one may carry over to complicate the next one. The program will label places where it cannot determine the identity of the nucleotide from the data with an “N.” That is ok.

Often the beginning and ends of your sequence will be “messy.” You will need to examine the sequence data visually, and trim away these ends. Often 20-50 base pairs on each end will not be useable.

From here, you can copy the data sequence and paste it into a website/database maintained by the **NCBI** (National Center for Biotechnology Information). Conducting a search called a **BLASTn** analysis. This analysis will compare your **n**ucleotide (that is what the ‘n’ at the end stands for) sequence to those in the database. This will likely provide you with an identification of your organism.

## LAB OVERVIEW

Your instructor will provide you with a file containing the data from your DNA sequencing reactions. You will need to download a program called “FinchTV” from here: <https://digitalworldbiology.com/FinchTV>. You will use this program to decide where to trim off the messy ends of your sequence, if needed. You copy the trimmed sequence from here and paste it into the nucleotide BLAST search engine here: [www.ncbi.nlm.nih.gov/BLAST](http://www.ncbi.nlm.nih.gov/BLAST). ***THIS LAB NEEDS TO BE DONE TWICE:*** once with the forward sequence and once with the reverse. Ideally, the two should yield very similar results.

Go through this tutorial on how to use the BLASTn tool: <https://digitalworldbiology.com/blast>.

**Using Finch TV:**

1. Download FinchTV from <https://digitalworldbiology.com/FinchTV>, if it is not already on your computer, and open the program. You can also use a program called “4 Peaks” (for Mac) <http://www.mekentosj.com/science/4peaks> but these instructrions are for FinchTV.
2. Drag our DNA sequence file into the program or open your sequence using FinchTV.
3. Use the Vertical Scale adjustment on the left side of the program window to adjust the peak height. It is important to see the entire peak. Peak height corresponds to the intensity of the signal.
4. Use Horizontal scroll/sizing bars to look over your data. Good sequence is defined by tall distinct peaks that have little overlap. Hopefully most of your sequence will resemble this description.

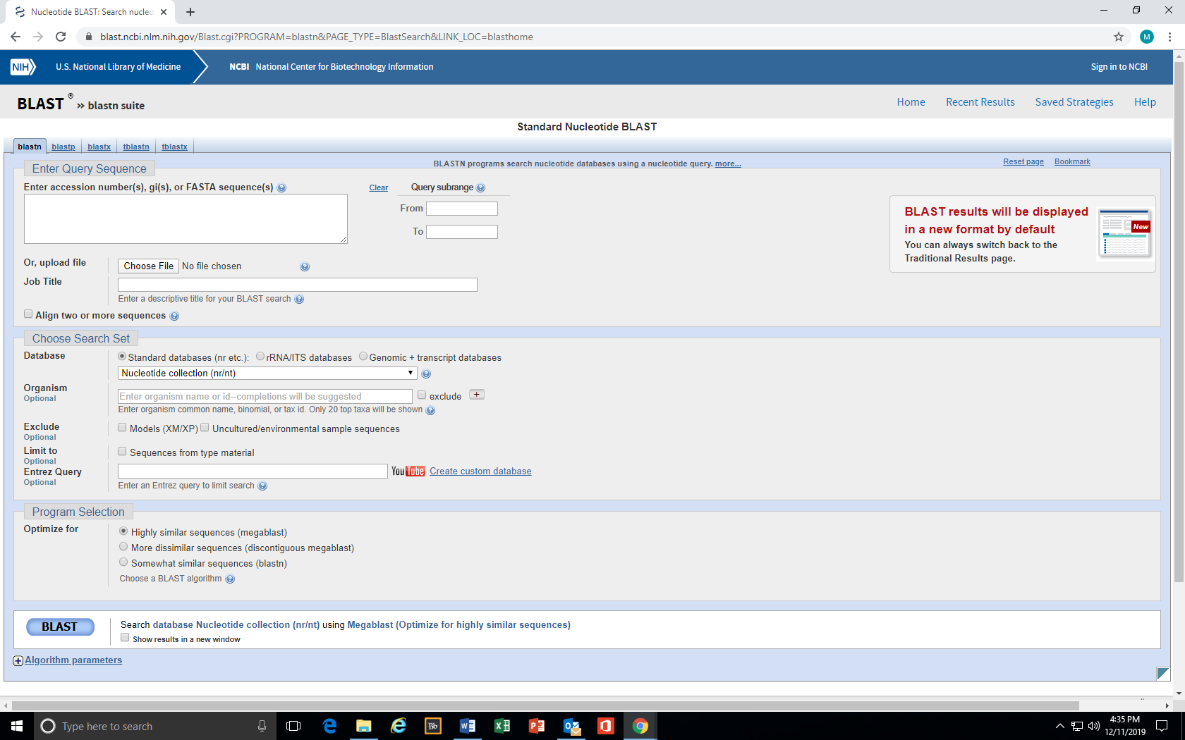
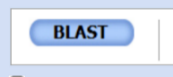
Color key for looking at chromatograms:

Guanine (G)– Black Thymine (T) –Red

Cytosine (C) –Blue Adenine (A) –Green

1. Clicking on “Wrapped View” will allow you to see the entire sequence in one screen. This can be handy.
2. Examine your sequence. Both ends will have 20-50 base pairs of messy sequence before the chromatograms start to have nice peaks. You will want to make a call where the sequence stops being messy and becomes ‘pretty’ and the trim off the messy ends. To do this, select the bases to be included in your search by holding down the mouse button and dragging from the start of the sequence you want to include in your search to the end of the sequence you want to include.
3. Select from the Edit dropdown menu “BLAST Sequence,” then mouse over and click “Nucleotide, BLASTn.”
4. You will be redirected to National Library Medicine BLAST website to complete your search.

## BLAST SEARCH:

1. The page you are redirected to should resemble the one below:
2. Under “Database” click rRNA/ITS databases
3. To start the search process click on:
4. Wait. The screen will start auto-updating, and it may take a minute or two with a good internet connection. It will ultimately come up with a page that will have a “graphic summary” and if you scroll down, a list of “Sequences producing significant alignments” and further down, it will have each match showing which base pairs match your sequence. **THIS page is the information you want.** *If you make an account, you can save your search and come back to it later!*
5. This <https://digitalworldbiology.com/BLAST/slide7.html> and <https://digitalworldbiology.com/BLAST/slide8.html> explain the meaning of the different columns. You will be most interested in the description of the sequence, the % Identity, and the E value.