LESSON 9
HANDOUT

Analyzing A DNA Sequence Chromatogram

Student Researcher Background:
DNA Analysis and FinchTV

DNA sequence data can be used to answer many types of questions. Because DNA sequences differ somewhat between species and between individuals within a species, DNA sequences are widely used for identification. In this activity, you will use bioinformatics programs to work with DNA sequences and identify the origin of a DNA sample.

Aim: Today, your job as a researcher is to:

1. Edit and trim the DNA sequence by using quality data from the chromatograms.
2. Translate the sequence to check for stop codons.
3. Use BLAST to identify the origin of the DNA sequence.
4. Use BOLD to confirm the identification of the species (or genus) and place the sample in a phylogenetic tree.

Discrepancy: A discrepancy in DNA sequencing is a point where the sequences from different samples or DNA strands disagree.

Quality values: A quality value is a number that is used to assess the accuracy of each base in a DNA sequence. Quality values can be used to help guide decisions about the **discrepancies** between different sequences. For more on quality values, see *Part II*.

Instructions: Write your answers to the questions in your lab notebook or on a separate sheet of paper, as instructed by your teacher.

PART I: Learning to Work with Sequences

Student Researcher Background:
Using FinchTV for DNA Analysis

FinchTV is a designed to allow researchers to view DNA sequence files like the **chromatograms** you are using here. In a chromatogram file, the signal intensities are presented in a graph with the four bases, each identified by different colors. Like many sequence analysis programs, FinchTV uses green for adenine, red for thymine, black for guanine, and blue for cytosine, as seen in the "DNA Sequencing Key" below.

DNA Sequencing Key

Guanine (G) = Black	**Thymine (T) = Red**
Cytosine (C) = Blue	**Adenine (A) = Green**

A. Getting Familiar with FinchTV

1. If it is not already open, open your ***DNA sequence*** chromatogram file (sequence files with the ".ab1" file extension) in FinchTV.

2. Use the ***Vertical Scale*** adjustment on the left side of the program window to adjust the peak height, as shown in ***Figure 1***. It is important for you, the researcher, to be able to clearly see the DNA sequence peaks. The height of a peak corresponds to the relative concentration of that base, at that position in the sequence. The height should be high enough for you to see clearly, but not so high that the background or "noise" peaks at the bottom of the chromatogram (***black arrow***) overwhelm your sequence data (***white arrow***).

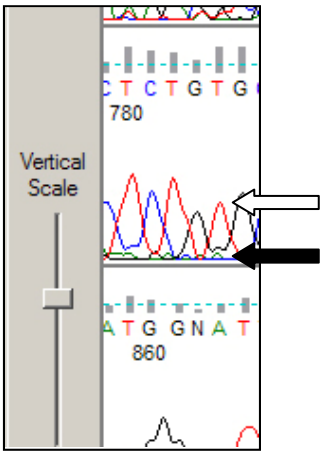

Figure 1: Vertical Scale. Source: FinchTV.

3. Click the ***Wrapped View*** icon to view the entire sequence in one screen.

4. Click the ***Base Position Numbers*** icon to view the base position numbers throughout the sequence.

5. Click the ***Base Calls*** icon to view the base calls (i.e., what the computer program interprets the sequence to be).

6. Click the ***Quality*** icon to display the quality bar graph above each DNA sequence peak. When evaluating data, it is important to look not only at what the data is, but whether or not the data is high quality. The **quality value** for DNA sequences is expressed as the "Q" value ("Q" for "Quality").

Quality values: A quality value is a number that is used to assess the accuracy of each base in a DNA sequence. Quality values represent the ability of the base calling software to identify the base at a given position and are calculated by taking the $\log_{10}$ of the error probability and multiplying it by -10.

- A base with a quality value of 10 has a one in ten chance of being misidentified.
- Bases with quality values of 20, 30, and 40, have error probabilities of one in 100, one in 1,000, and one in 10,000, respectively.

Many databases ask that submitted DNA sequences have an average quality value close to 30 or higher. Quality values can be used to help guide decisions about the **discrepancies** between different sequences, as you will do below.

B. Viewing information for a specific base

7. With the ***quality values*** displayed for your sequence, select a base by clicking it with your mouse. The selected base will be highlighted, as seen in ***Figure 2***.

8. The one letter abbreviation for that base will appear in the lower left corner, along with the sequence position and the ***quality value*** (if available). In ***Figure 2***, the selected base is a T (thymine) located at position 70 in the sequence and has a quality value of 17, which is generally accepted to be low quality.

9. Experiment by clicking on a number of different bases in your sequence. Answer these questions in your lab notebook or on another sheet of paper:

   What is the highest Quality Value you see?

   What is then lowest Quality Value you see?

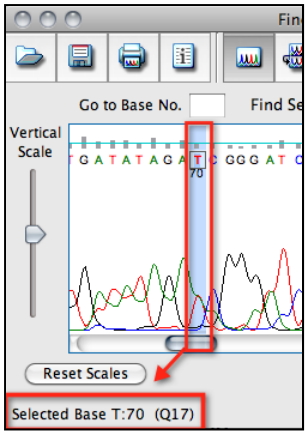

Figure 2: Quality Values. Source: FinchTV.

C. Finding a base or sequence in FinchTV

10. To find a specific base, enter the position number for that base in the ***Go to Base No***. window and click the ***Return*** or ***Enter*** key on your keyboard. The requested base will appear at the beginning of the sequence window (see ***Figure 3***).

11. Experiment by selecting a base number in your sequence.

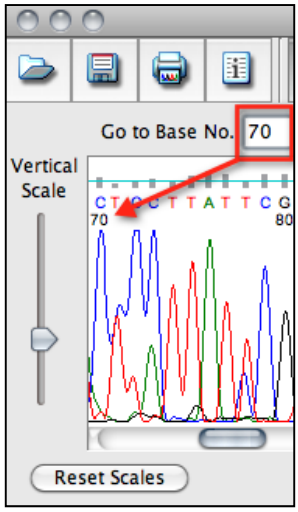

Figure 3: Finding Specific Bases. Source: FinchTV.

12. Another way to find a specific base in FinchTV is to ***enter a sequence that is located near or contains your base***. In ***Figure 4***, the sequence GGTCAA was typed in the ***Find Sequence*** window and the Return key pressed. FinchTV located the sequence and highlighted it in blue.

13. Experiment with your sequence by trying to locate the sequence "GGTCAA." ***Is that sequence present in your DNA sequence data?*** Record your answer in your lab notebook or on a separate sheet of paper.

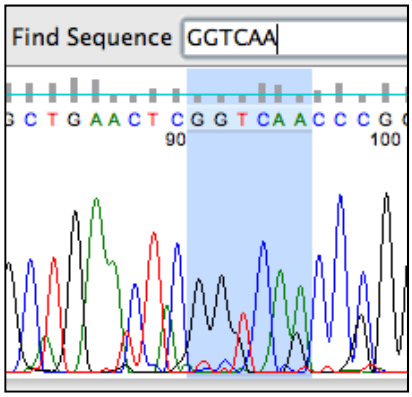

Figure 4: Finding a Specific Sequence. Source: FinchTV.

PART II: Edit and Trim the DNA Chromatogram File

Now it is time to update your DNA sequence file using the Quality scores provided for your sequence.

14. Find the file that contains your sequence chromatogram (sequence with the ".ab1" extension).

15. Make a copy of the file that contains your sequence and rename the copy so that the new file name begins with word "Edit." It always a good idea to save the original, unedited data file in case you need to go back and review it.

16. For each position that will be edited:

    a. To ***change*** a base in FinchTV, click that position and type the letter for the new base.

    b. To ***delete*** a base, select that base and click the delete key.

    c. To ***insert*** a base, click the position in the sequence, right click, choose ***Insert before base***, and enter the letter of the new base.

17. Save your edited file.

18. Chromatograms often contain low quality sequences at the 5' and 3' ends that are removed by trimming (deleting the bases). Trim your sequences by selecting the bases to be trimmed and clicking the delete key.

    a. Trim bases from the 5' end until the last 20 bases contain fewer than 3 bases with quality values below 10.

    b. Trim bases from the 3' end until the last 20 bases contain less than 3 bases with quality values below 10.

19. Save your edited DNA chromatogram file (which will include the ".ab1" extension).

PART III: Perform a blastn to Identify Your DNA Sequence

Nucleotide BLAST, or **BLASTn**, is a tool commonly used for DNA Sequence identification.

20. Open your edited DNA chromatogram file (if it is not already open).

21. Open the FinchTV ***Edit*** menu and choose ***BLAST Sequence,*** and then select ***Nucleotide, BLASTn*** (***Figure 5***). This will open BLASTn at the NCBI and paste your sequence in the query box. **Note:** If your sequence does not appear in the query box (as seen in ***Figure 6***), go back to FinchTV and select your DNA sequence first by going to the ***Edit*** menu and choosing ***Select All.***

22. From the ***Choose Search Set*** menu, select ***Nucleotide collection (nr/nt)*** (black box, ***Figure 6***). **Note:** If your BLAST search returns only human sequences, you may have forgotten to change the default database from the Human Genome.

23. Click ***BLAST***.

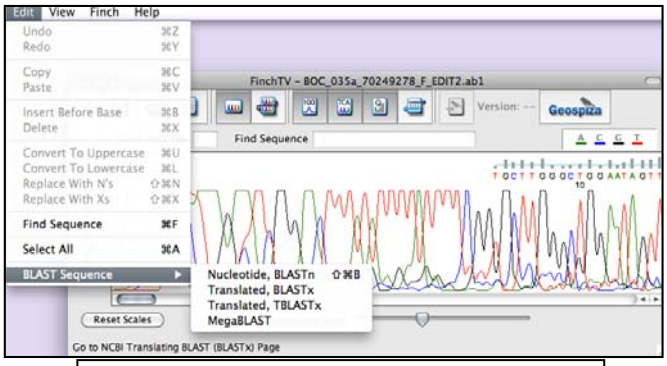

Figure 5: Choosing blastx from FinchTV. Source: FinchTV.

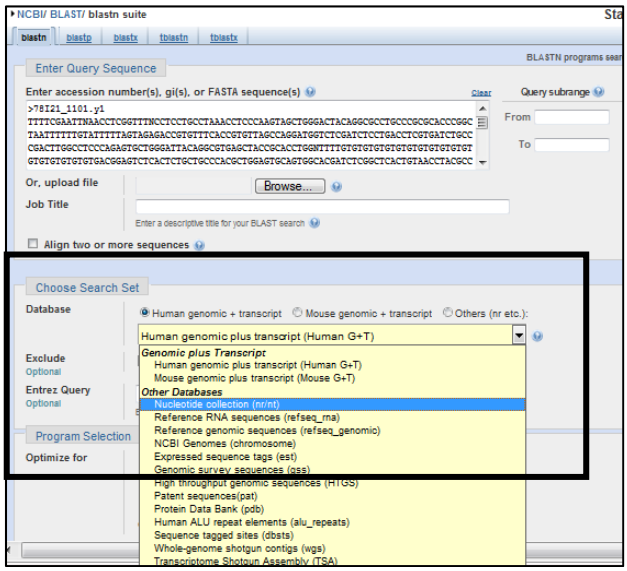

Figure 6: Using BLASTn to identify your DNA sequence. Source: NCBI BLASTn.

PART IV: Identify Your Sequence and Place it in a Phylogenetic Tree by Comparing it with Sequences in the BOLD Database

Scientists use many databases to identify DNA sequences, including the NCBI Nucleotide Database (using BLAST) and the BOLD Database (which also uses the BLAST algorithm to compare your sequence to other sequences in the BOLD Database). Sometimes the results from one database confirm the results from the other database, and sometimes the results from the NCBI Nucleotide are inconclusive, making it helpful to determine which species are most related to the species from which your DNA sequence came.

24. Open your edited DNA chromatogram file (if it is not already open).

25. View the DNA sequence by clicking on the ***Chomatogram Info*** icon.

26. Select the sequence and copy it.

27. Go to the BOLD database at http://www.barcodinglife.com. BOLD uses BLAST to compare sequences you enter to a database of sequences that meet the internationally agreed upon criteria for DNA barcoding.

28. Choose ***Identification*** from the menu near the top of the homepage.

29. Select ***All Barcode Records on BOLD*** from the ***Search Databases*** menu (black box, ***Figure 9***).

30. Paste your sequence in the text area labeled ***Enter sequence in fasta format*** (black arrow, ***Figure 7***) and click ***Submit***.

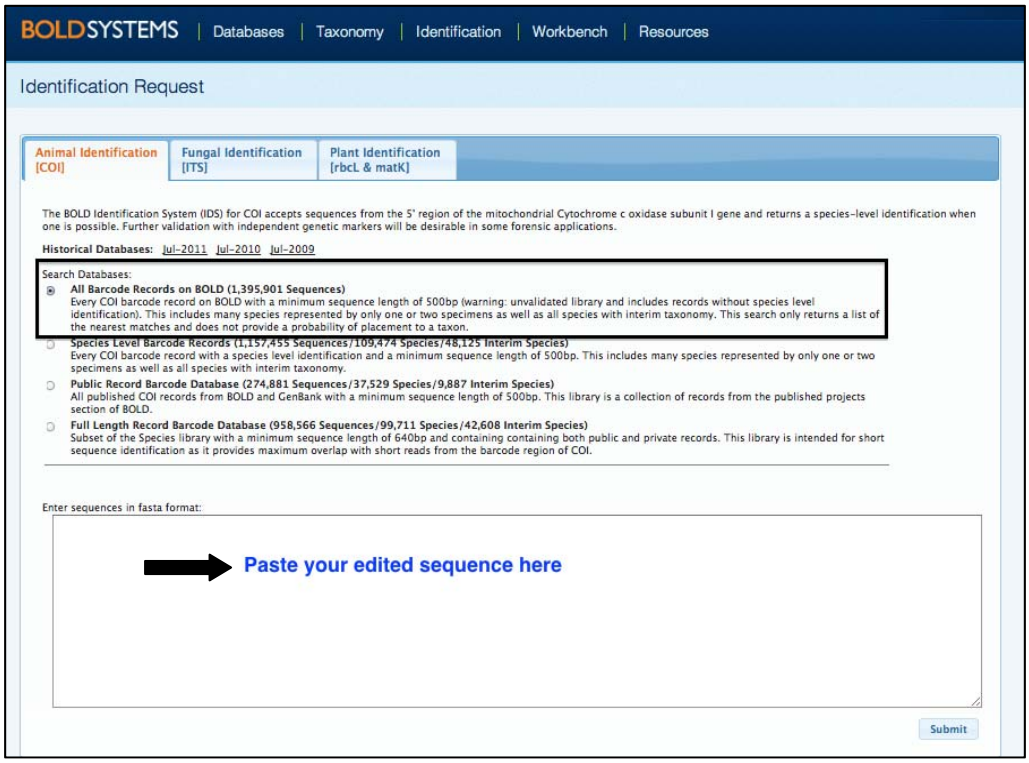

Figure 7: Entering your DNA Sequence to Identify it Using BOLD. Source: BOLD.

No sequences in the database matched ours closely enough to be considered a match by the BOLD identification algorithms. However, we can look at the data and find related sequences. The image in ***Figure 8*** shows that our sequence was 53.49% similar to a sequence from *Aphis craccivora*.

31. Look at your BOLD search results. ***What species matches your sequence most closely?*** What ***genus does that species belong to?*** Include the complete scientific name (Genus and species) in your lab notebook or on a separate sheet of paper.

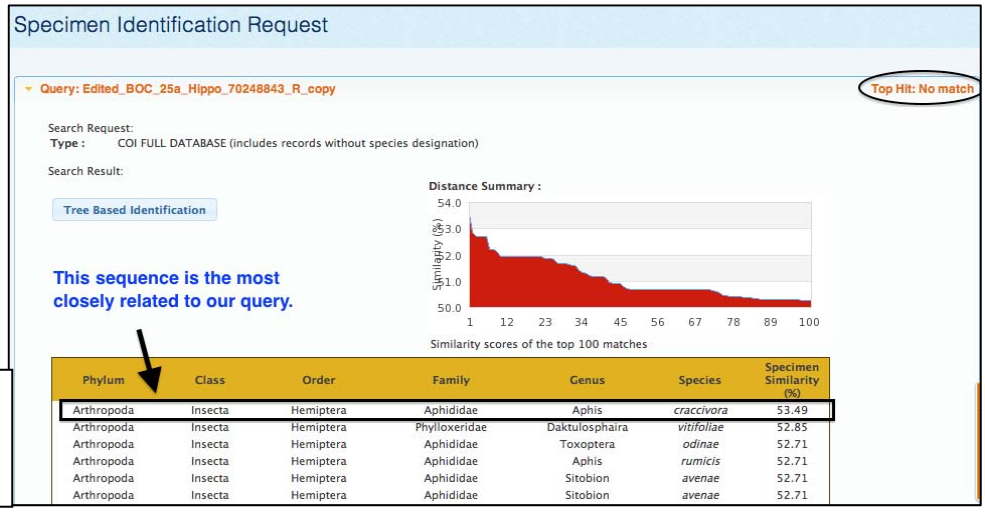

Figure 8: Results from a Search of the BOLD Database. Source: BOLD.

32. The search results also provide taxonomic information about the species from which the DNA sequence was isolated, as seen in the black box in ***Figure 10***. Fill in the following information for the species that matches yours most closely, in your lab notebook or on a separate sheet of paper.

Phylum: _____
Class: _____
Order: _____
Family: _____

33. Click the ***Tree Based Identification*** button to see where your sequence fits in a BOLD-generated phylogenetic tree (black box, ***Figure 9***).

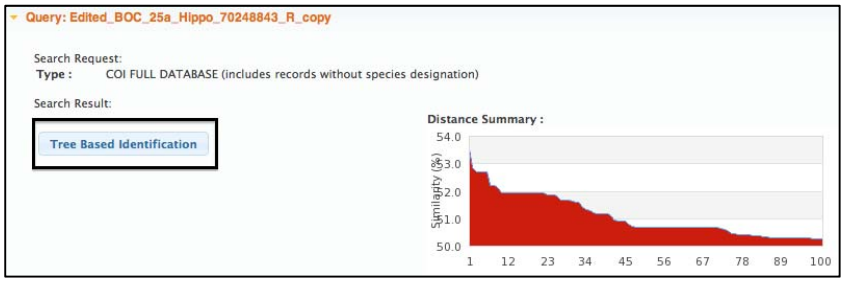

Figure 9: Select "Tree Based Identification" Button to See Where Your Sequences Fits in with the BOLD Phylogenetic Tree. Source: BOLD.

34. Select the ***View Tree*** button to download a multi-page PDF file containing your tree (black box, ***Figure 10***). You may also wish to click the ***View Image List*** button to view images of related species.

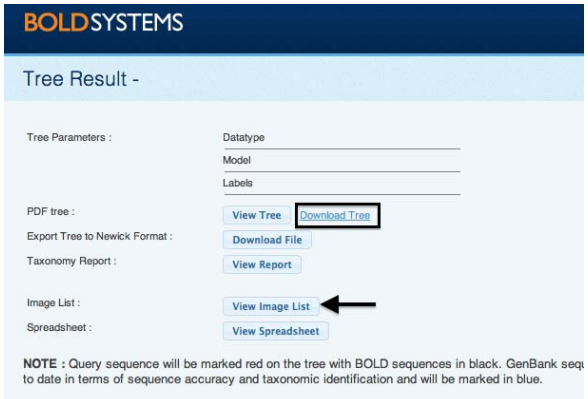

Figure 10: Select "View Tree" to Download PDF of Your Tree. Source: BOLD.

35. Find your sample in the tree file. Your sample will be identified in red, as seen in the example shown in ***Figure 11***. You will probably have to scroll to the second or third page in your PDF.

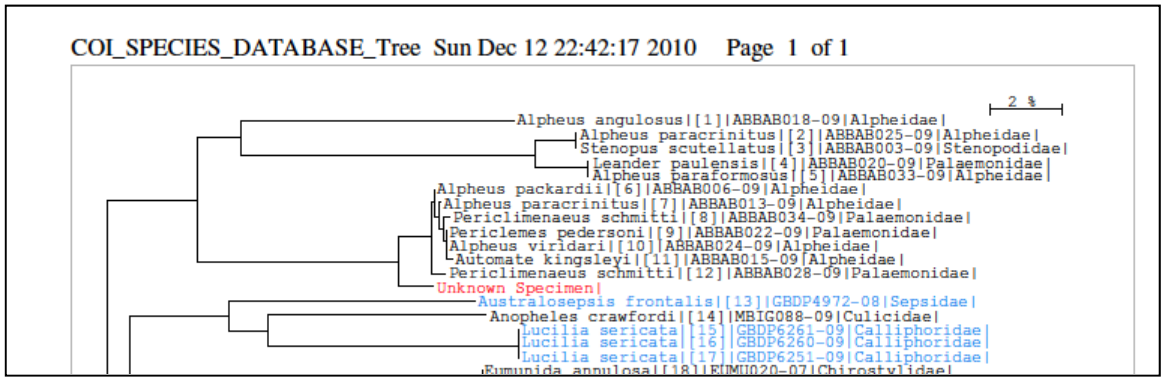

Figure 11: A Phylogenetic Tree from BOLD, with the Unknown DNA Sequence Highlighted in Red. Source: BOLD.

36. Copy and paste a screen capture image of your tree into a word processing document.

37. Based on what you see in your phylogenetic tree, record in your lab notebook or on a separate sheet of paper ***at least two organisms that are closely related to the species from which your sequence was obtained. Be sure to include the complete scientific name for each.***

38. Using online tools such as Google, Wikipedia, and/or the Encyclopedia of Life (http://www.eol.org), search for these closely-related organisms and ***list their common names*** in your lab notebook or on a separate sheet of paper, next to your answers to the previous question.

39. Using these same tools (Google, Wikipedia, and/or the Encyclopedia of Life), ***determine the common name of the species from which your DNA was obtained*** and record in your lab notebook or on a separate sheet of paper.

40. Read the Wikipedia and Encyclopedia of Life entries about all three of these species. If these species are not found in Wikipedia or the Encyclopedia of Life, you may need to find other information from a Google search.

41. ***What do all of these organisms have in common? Habitat? Diet?*** In your lab notebook or on a separate sheet of paper, ***list any similarities that these organisms share. Also note any important differences, and other facts that you find interesting and/or surprising.***

42. Finally, thinking about what you have learned in all nine lessons about DNA barcoding, ***how have these lessons contributed to your understanding of the process of how genetic research is performed?*** Write your answer(s) in your notebook or on a separate sheet of paper.