2 Using BLAST and BOLD for Genetic Research Instructions

Student Researcher Background:

DNA Barcoding, BLAST, and BOLD

DNA barcoding is the use of a standardized DNA sequence as a means to identify new species, identify unknown samples, and compare relatedness and evolution among different species. Certain genes can be used in this manner because some regions of these genes are **conserved**—that is, they show a very slow rate of evolution and very little change in their DNA sequence—unlike other regions that evolve more rapidly and show more changes in their DNA. The gene that scientists in the Barcode of Life community have decided to use for animal species is the cytochrome c oxidase subunit 1 gene, or *COI*.

Conserved: DNA or protein sequences are said to be "conserved" if the sequences are the same or very similar.

DNA barcodes, if they are known, can be accessed from several databases, including the National Center for Biotechnology Information (NCBI) and the Barcode of Life Database (BOLD).

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

1. Use the bioinformatics tool ***BLAST*** to identify the source of an unknown DNA barcode sequence.
2. Find important ***taxonomic information*** for your species from ***BOLD***.
3. Find your ***scientific collaborators*** to discuss your findings and then ***generate a hypothesis*** about the relatedness of the species within your group.

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

How can we use BLAST in genetic research?

- **To identify a sample.** For example, many fish, birds, and marine mammals were killed in the 2010 Gulf of Mexico oil spill. Identifying all of these animals when they are coated in oil or are in their juvenile stage can be difficult. Or perhaps you believe that the packaged fish sold by a particular market is not the wild caught Alaskan Coho salmon they are advertising, but actually farmed Atlantic salmon.
- **To find related species.** For example, perhaps you work for the Centers for Disease Control and Prevention (CDC) and your boss just identified a patient with what looks like avian influenza (bird flu). It is your job to determine whether the infection really is influenza, and where the patient may have gotten infected.

PART I: Performing a BLAST Search

Your instructor will provide information on how to obtain an "unknown DNA sequence." Copy this sequence using the ***Copy*** command from the ***Edit*** menu of your web browser. Alternatively, you can save it as a text file (.txt) to your computer or memory stick.

Nucleotide: The basic building blocks of DNA and RNA: guanine (G), cytosine (C), adenine (A), thymine (T), and uracil (U). Each nucleotide contains a nitrogenous base, a five-carbon sugar, and a phosphate group.

1. Once you have your unknown DNA sequence, go to the BLAST homepage: http://blast.ncbi.nlm.nih.gov/Blast.cgi.
2. Choose a BLAST program to run, as shown in ***Figure 1***.

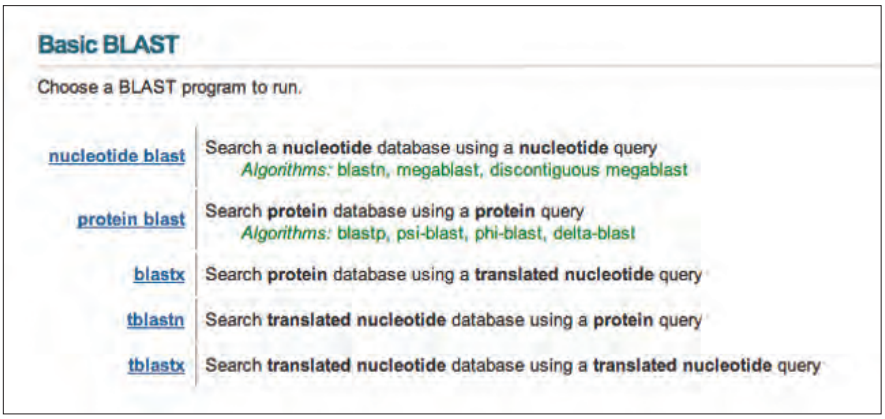


Figure 1: Choosing a BLAST Program. Source: NCBI BLAST.

You will be using your DNA sequence to search for other DNA sequences. Which BLAST program will you run?

3. Paste your DNA sequence into the search box, ***Enter Query Sequence*** (green box, ***Figure 2***), or upload a file containing the sequence (.txt files).
4. Select the database to search, as shown in ***Figure 2***. To identify an unknown sequence, or to find related sequences, select the ***Nucleotide Collection (nr/nt)***.
5. Make your ***Program Selection***, based on the goal of your search. In this exercise, you are identifying a sample.
   - ***Highly similar sequences (megablast)***. Example use: identifying a sample.
   - ***More dissimilar sequences (discontinuous megablast)***. Example use: find closely related species.
   - ***Somewhat similar sequences (blastn)***; best for shorter searches (fewer than 20 bases).
6. Click the ***BLAST*** button to start your search, as shown in ***Figure 3***.

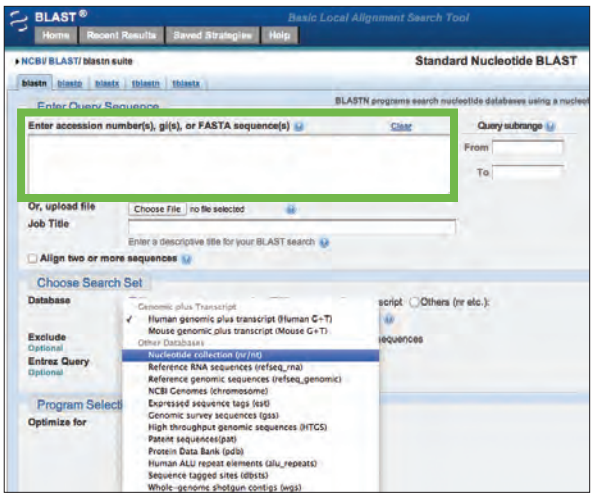


Figure 2: Selecting the Nucleotide Collection. Source: NCBI BLAST.

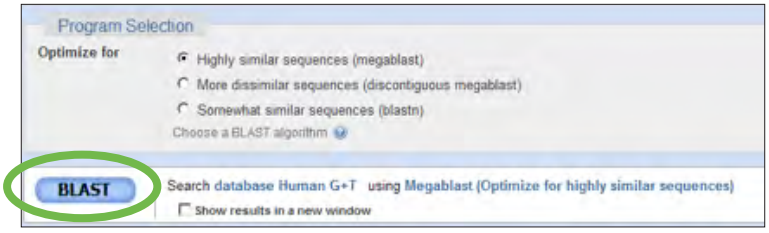


Figure 3: Starting the BLAST Search. Source: NCBI BLAST.

PART II: Understanding BLAST Search Results and Pairwise Comparisons

7. Each BLAST result is a pairwise comparison between your DNA sequence and the DNA sequences in an NCBI database. ***Figure 4*** provides information on the different parts of the BLAST search result.

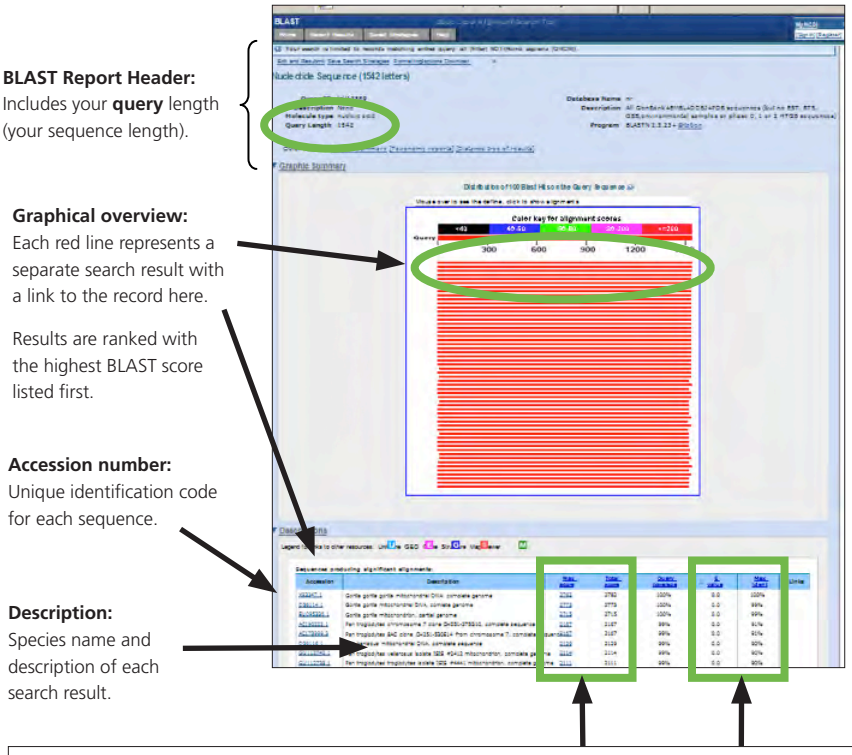


Figure 4: Parts of a BLAST Result. Source: NCBI BLAST.

BLAST Scores and Statistics: How to Rank Your Results

Max Score/Total Score: Score of the pairwise comparison between your DNA sequence and the DNA sequence in the NCBI database (EX: +2 for match; -1 for mismatch, -2 for a gap). Higher scores mean better alignments.

A	T	G	G	T	T	A	C	T	
A	-	G	G	A	A	T	C	T	
+2	-2	+2	+2	-1	-1	-1	+2	+2	Total Score = 5

E-value: Would this result be "expected" by chance alone? Lower E values show results that are less likely to be obtained by chance. Lower E- value numbers are better.

- **Example 1:** A hit with an E-value of 0.0 is less likely to be due to chance than a hit with an E-value of 5.4
- **Example 2:** A hit with an E-value of 2e-8 (2.0 x 10^{-8} or 0.00000002) is less likely to be due to chance than a hit with an E- value of 7.5

Query: When searching databases like those at the NCBI, your "query" is the sequence you are searching with and trying to match. In this case, your query is your unknown sequence.

Subject: When searching the databases at the NCBI, the subject sequences are sequences from the database that match the query. In this case, if a subject sequence is identical to your query, your query sequence probably came from the same type of organism that contributed the subject sequence.

Accession number: A unique identifier or code assigned to every entry in the National Center for Biotechnology Information (NCBI) databases. This unique code can be used to search the databases to find your gene or protein of interest.

8. Select your first search result, either by clicking on the first/top red line, or scrolling down and clicking on the ***Max Score*** to the right of the ***Description***.

Figure 5: Deciphering BLAST Search Results. Source: NCBI BLAST.

9. What is the description of your first (best) BLAST result? (See ***Figure 5***). Be sure to include the genus, species, and subspecies name (if applicable).

For example, the scientific name for the domestic dog is *Canis lupus familiaris*.

Genus | Species | Subspecies (if applicable)

What is scientific name for ***your*** species? Be sure to include the Genus, species and subspecies (if applicable).

10. What is the accession number of your search result?

11. Do you feel confident that you have correctly identified the species from which your sequence was taken? Justify your answer in terms of the BLAST scores (percent identity and E-value).

12. Your BLAST search result will include a pairwise comparison for each result. Each alignment is a comparison between two sequences: your **query** (unknown) sequence and a **subject** (result) sequence. Look through your pairwise comparison data. Do you see any differences between the query sequence and the best matching subject sequence in the pairwise alignment?

13. Your query will likely begin at nucleotide #1, but the best matching subject sequence might not (see "DNA Pairwise Alignment" in ***Figure 5***). The *COI* gene is encoded by the mitochondrial genome, which is over 13,000 nucleotides long. Look for the nucleotide number of your subject sequence (abbreviated "Sbjct") just below where your query starts (at #1), as shown in ***Figure 6***. This is the nucleotide position where your subject sequence starts.

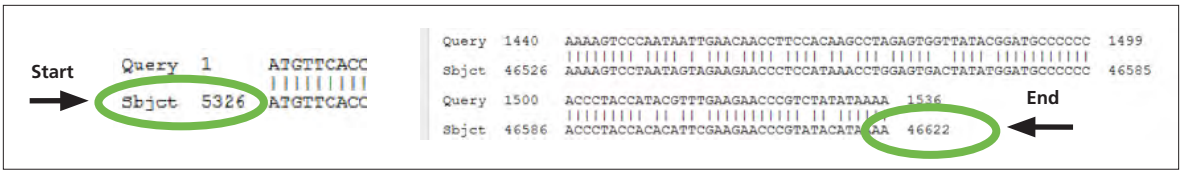


Figure 6: Locating the Subject Sequence. Source: NCBI BLAST.

a. At what nucleotide position does your subject sequence start?

b. At what nucleotide position does your subject sequence end?

PART III: Finding Taxonomic Information in the Barcode of Life Database (BOLD)

The Barcode of Life Database (BOLD) contains a great deal of information. Like the NCBI, it is a collection of databases that allows scientists around the world to collaborate by having a single location where researchers can submit or retrieve scientific data—specifically, DNA barcode information.

At BOLD, you will:

1. ***Determine*** whether your species has been barcoded.
2. Learn important **taxonomic** information about your species, and use that information to inform your hypothesis and to choose which other researchers in your class you will ***collaborate with***.

Taxonomic: Taxonomy is the science of classifying organisms into groups based on similarities in their physical characteristics. The goal of taxonomy is to organize species in groups based on their evolution and common ancestors. The word comes from the Greek "taxis" which means "order" or "arrangement."

Understanding Taxonomy: What's the Point?

Scientists categorize all life on planet Earth into a hierarchy called a **taxonomy**—Kingdom: phylum: class: order: family: genus: species. The taxonomic hierarchy is derived from evolutionary relationships, and ***makes it possible for scientists around the world to be sure they are all talking about the same species*** by providing a standard format for identification. The last two levels of the taxonomic hierarchy, the *Genus* and *species*, constitute the scientific name for a species, which you found in your BLAST results. All the species you will be studying are animals, which are part of Kingdom Animalia.

Determine Whether Your Species Has Been Barcoded

Barcoding species involves matching the DNA barcode sequence from an identified, known physical specimen (such as in a zoo, aquarium, or museum) with your DNA barcode sequences. You can use BOLD to identify unknown sequences, similar to what you did with your BLAST search. You can also search BOLD using the scientific name of your species, which you learned from your BLAST search. Now that you know the scientific name of your species, it is time to find the DNA barcode sequence for your species at the BOLD.

14. Go to the BOLD database: http://www.barcodinglife.org.

15. Type the scientific name of your species in the search box in the lower left portion of the screen, as shown in ***Figure 7***. If your search does not return any results, try searching with just the Genus name.

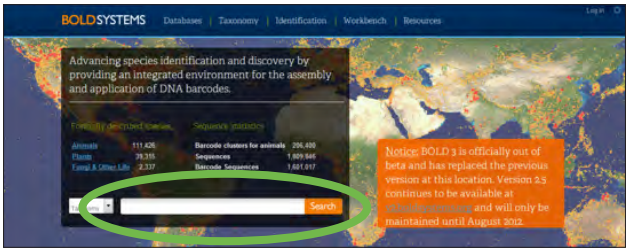


Figure 7: Searching for a Species in BOLD.
Source: Barcode of Life Data Systems.

16. Has your species been barcoded?

17. If your species has been barcoded, you will find a page for your species in BOLD. Take a moment to look at the information available, as shown in ***Figure 8***.

    a. How many specimens are available with Barcodes? [**Note:** "Barcodes" are complete DNA sequences, while "Sequences" may include only fragments.]

    b. How many of these records are Public Records?

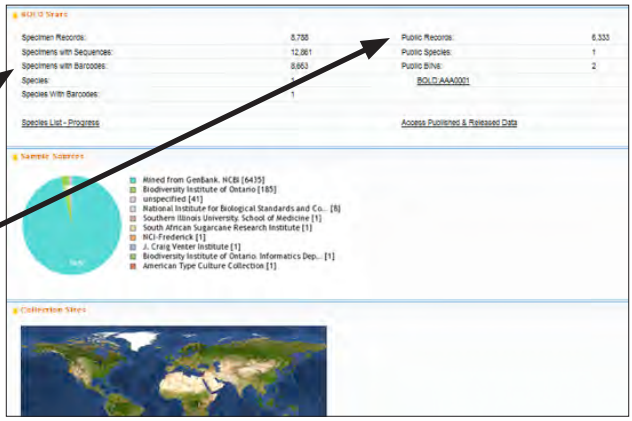
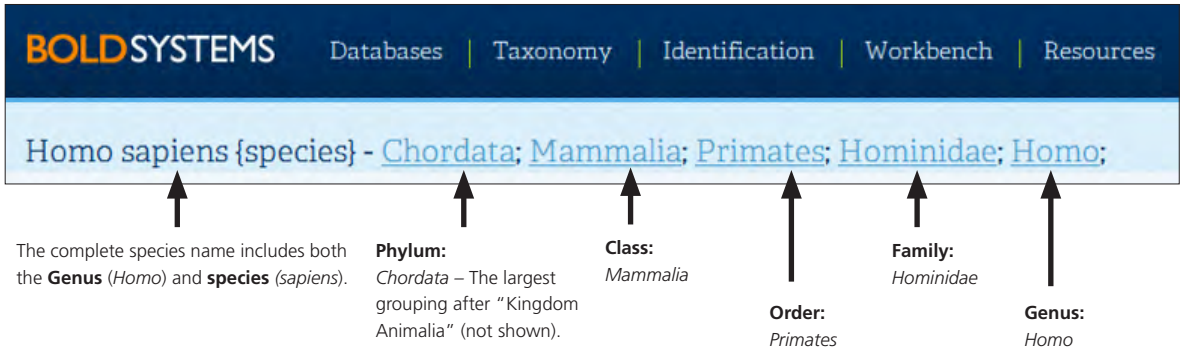


Figure 8: Finding information on a BOLD Species Page.
Source: Barcode of Life Data Systems.

18. Find the ***taxonomic information*** about the species across the top left side of the web page. Other taxonomic categories are also listed on the BOLD species page. Each is a subcategory of the one before it, from more general (larger) categories to more specific (smaller) categories. In the example shown in ***Figure 9***, taxonomic information for the species *Homo sapiens* is displayed.

Figure 9: Decoding the Taxonomic Information. Source: Barcode of Life Data Systems.



19. What is the Family for your species?

20. What is the Order for your species?

21. What is the Class for your species?

PART IV: Finding Your Collaborators and Generating a Hypothesis about Species Relatedness

Scientists often specialize in studying one particular group or type of species, such as canines, fish, or marine mammals. They form collaborations with other scientists who study the same or related species. ***Your teacher will help you find the other student researchers in your class who are studying species in the same taxonomic class as your species. Let your teacher know when you are ready to find your collaborators.***

Your teacher will also provide you with a handout that contains pictures of your group's species. Have a discussion with your collaborators about the characteristics of the species within your group. You can use these pictures and your discussion to help you come up with your research hypothesis about how your species are related to one another. Each group member should write down her own hypothesis. It is all right if your hypothesis is the same as or similar to the hypotheses of other members in your group; your hypothesis may also be completely different from or even contradict others in your group.

22. What is your Group Name (includes the class of your species)?

23. List the names of all the collaborators in your group and the species they are studying in the table on your *Worksheet*, or draw a similar table in your lab notebook or on your answer sheet. Be sure to list yourself as well!

Collaborators	Scientist's Name	Species Studied
Collaborator #1		
Collaborator #2		
Collaborator #3		
Collaborator #4		
Collaborator #5		
Collaborator #6		

24. What is your hypothesis about the relatedness of the species within your group?