INVESTIGATION 3

COMPARING DNA SEQUENCES TO UNDERSTAND EVOLUTIONARY RELATIONSHIPS WITH BLAST

How can bioinformatics be used as a tool to determine evolutionary relationships and to better understand genetic diseases?

BACKGROUND

Between 1990–2003, scientists working on an international research project known as the Human Genome Project were able to identify and map the 20,000–25,000 genes that define a human being. The project also successfully mapped the genomes of other species, including the fruit fly, mouse, and *Escherichia coli*. The location and complete sequence of the genes in each of these species are available for anyone in the world to access via the Internet.

Why is this information important? Being able to identify the precise location and sequence of human genes will allow us to better understand genetic diseases. In addition, learning about the sequence of genes in other species helps us understand evolutionary relationships among organisms. Many of our genes are identical or similar to those found in other species.

Suppose you identify a single gene that is responsible for a particular disease in fruit flies. Is that same gene found in humans? Does it cause a similar disease? It would take you nearly 10 years to read through the entire human genome to try to locate the same sequence of bases as that in fruit flies. This definitely isn't practical, so a sophisticated technological method is needed.

Bioinformatics is a field that combines statistics, mathematical modeling, and computer science to analyze biological data. Using bioinformatics methods, entire genomes can be quickly compared in order to detect genetic similarities and differences. An extremely powerful bioinformatics tool is BLAST, which stands for Basic Local Alignment Search Tool. Using BLAST, you can input a gene sequence of interest and search entire genomic libraries for identical or similar sequences in a matter of seconds.

In this laboratory investigation, you will use BLAST to compare several genes, and then use the information to construct a *cladogram*. A cladogram (also called a phylogenetic tree) is a visualization of the evolutionary relatedness of species. Figure 1 is a simple cladogram.

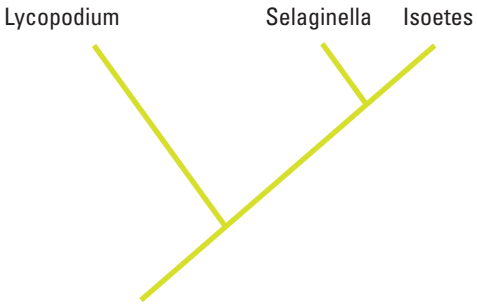

Figure 1. Simple Cladogram Representing Different Plant Species

Note that the cladogram is treelike, with the endpoints of each branch representing a specific species. The closer two species are located to each other, the more recently they share a common ancestor. For example, *Selaginella* (spikemoss) and *Isoetes* (quillwort) share a more recent common ancestor than the common ancestor that is shared by all three organisms.

Figure 2 includes additional details, such as the evolution of particular physical structures called shared derived characters. Note that the placement of the derived characters corresponds to when (in a general, not a specific, sense) that character evolved; every species above the character label possesses that structure. For example, tigers and gorillas have hair, but lampreys, sharks, salamanders, and lizards do not have hair.

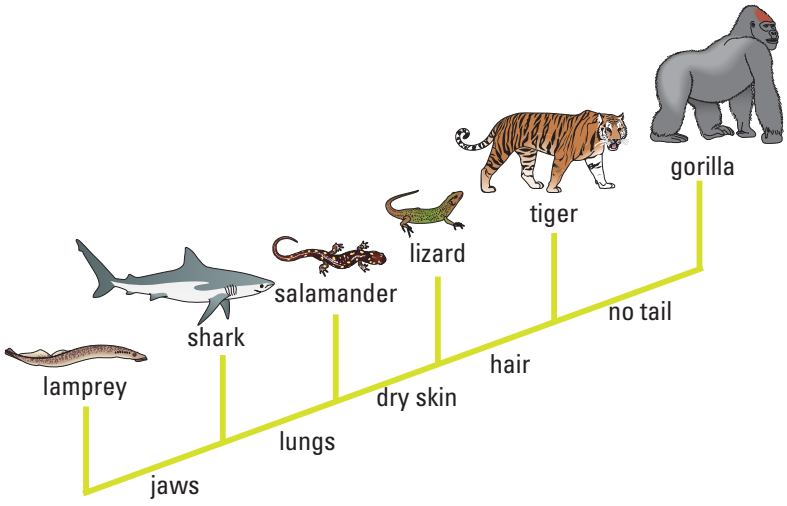

Figure 2. Cladogram of Several Animal Species

The cladogram above can be used to answer several questions. Which organisms have lungs? What three structures do all lizards possess? According to the cladogram, which structure — dry skin or hair — evolved first?

Historically, only physical structures were used to create cladograms; however, modern-day cladistics relies heavily on genetic evidence as well. Chimpanzees and humans share 95%+ of their DNA, which would place them closely together on a cladogram. Humans and fruit flies share approximately 60% of their DNA, which would place them farther apart on a cladogram.

Can you draw a cladogram that depicts the evolutionary relationship among humans, chimpanzees, fruit flies, and mosses?

Learning Objectives

- To create cladograms that depict evolutionary relationships
- To analyze biological data with a sophisticated bioinformatics online tool
- To use cladograms and bioinformatics tools to ask other questions of your own and to test your ability to apply concepts you know relating to genetics and evolution

General Safety Precautions

There are no safety precautions associated with this investigation.

THE INVESTIGATIONS

Getting Started

Your teacher may assign the following questions to see how much you understand concepts related to cladograms before you conduct your investigation:

1. Use the following data to construct a cladogram of the major plant groups:

Table 1. Characteristics of Major Plant Groups

Organisms	Vascular Tissue	Flowers	Seeds
Mosses	0	0	0
Pine trees	1	0	1
Flowering plants	1	1	1
Ferns	1	0	0
Total	3	1	2

2. GAPDH (glyceraldehyde 3-phosphate dehydrogenase) is an enzyme that catalyzes the sixth step in glycolysis, an important reaction that produces molecules used in cellular respiration. The following data table shows the percentage similarity of this gene and the protein it expresses in humans versus other species. For example, according to the table, the GAPDH gene in chimpanzees is 99.6% identical to the gene found in humans, while the protein is identical.

Table 2. Percentage Similarity Between the GAPDH Gene and Protein in Humans and Other Species

Species	Gene Percentage Similarity	Protein Percentage Similarity
Chimpanzee *(Pan troglodytes)*	99.6%	100%
Dog *(Canis lupus familiaris)*	91.3%	95.2%
Fruit fly *(Drosophila melanogaster)*	72.4%	76.7%
Roundworm *(Caenorhabditis elegans)*	68.2%	74.3%

a. Why is the percentage similarity in the gene always lower than the percentage similarity in the protein for each of the species? (Hint: Recall how a gene is expressed to produce a protein.)

b. Draw a cladogram depicting the evolutionary relationships among all five species (including humans) according to their percentage similarity in the GAPDH gene.

Online Activities

You can also prepare for the lab by working through the following online activities:

- "The Evolution of Flight in Birds"
  http://www.ucmp.berkeley.edu/education/explorations/reslab/flight/main.htm
  This activity provides a real-world example of how cladograms are used to understand evolutionary relationships.
- "What did T. rex taste like?"
  http://www.ucmp.berkeley.edu/education/explorations/tours/Trex/index.html
- "Journey into Phylogenetic Systematics"
  http://www.ucmp.berkeley.edu/clad/clad4.html

Figure 3. Fossil Specimen

Procedure

A team of scientists has uncovered the fossil specimen in Figure 3 near Liaoning Province, China. Make some general observations about the morphology (physical structure) of the fossil, and then record your observations in your notebook.

Little is known about the fossil. It appears to be a new species. Upon careful examination of the fossil, small amounts of soft tissue have been discovered. Normally, soft tissue does not survive fossilization; however, rare situations of such preservation do occur. Scientists were able to extract DNA nucleotides from the tissue and use the information to sequence several genes. Your task is to use BLAST to analyze these genes and determine the most likely placement of the fossil species on Figure 4.

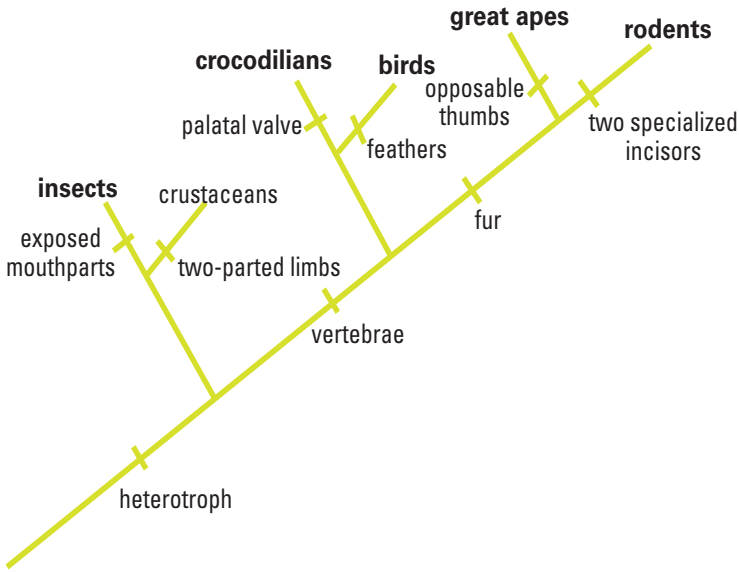

Figure 4. Fossil Cladogram

Step 1 Form an initial hypothesis as to where you believe the fossil specimen should be placed on the cladogram based on the morphological observations you made earlier. Draw your hypothesis on Figure 4.

Step 2 Locate and download gene files. Download three gene files from the AP Biology Investigative Labs page at AP Central: ***http://apcentral.collegeboard.com/apc/members/courses/teachers_corner/218954.html***.

Step 3 Upload the gene sequence into BLAST by doing the following:

a. Go to the BLAST homepage: ***http://blast.ncbi.nlm.nih.gov/Blast.cgi***

b. Click on "Saved Strategies" from the menu at the top of the page.

Figure 5

c. Under "Upload Search Strategy," click on "Browse" and locate one of the gene files you saved onto your computer.

d. Click "View."

Figure 6

e. A screen will appear with the parameters for your query already configured. NOTE: Do not alter any of the parameters. Scroll down the page and click on the "BLAST" button at the bottom.

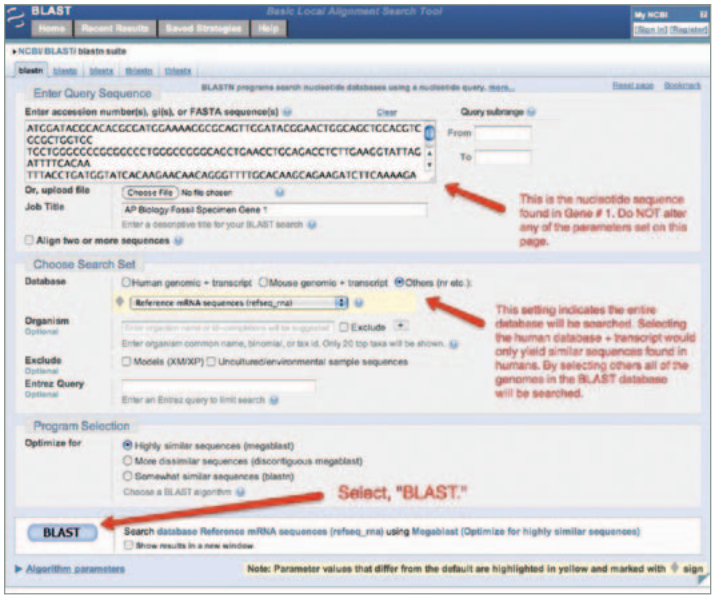

Figure 7

f. After collecting and analyzing all of the data for that particular gene (see instructions below), repeat this procedure for the other two gene sequences.

Step 4 The results page has two sections. The first section is a graphical display of the matching sequences.

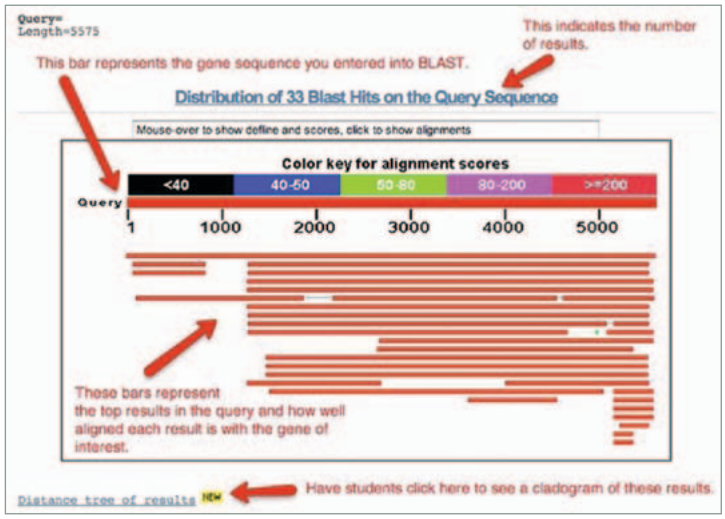

Figure 8

Scroll down to the section titled "Sequences producing significant alignments." The species in the list that appears below this section are those with sequences identical to or most similar to the gene of interest. The most similar sequences are listed first, and as you move down the list, the sequences become less similar to your gene of interest.

Figure 9

If you click on a particular species listed, you'll get a full report that includes the classification scheme of the species, the research journal in which the gene was first reported, and the sequence of bases that appear to align with your gene of interest.

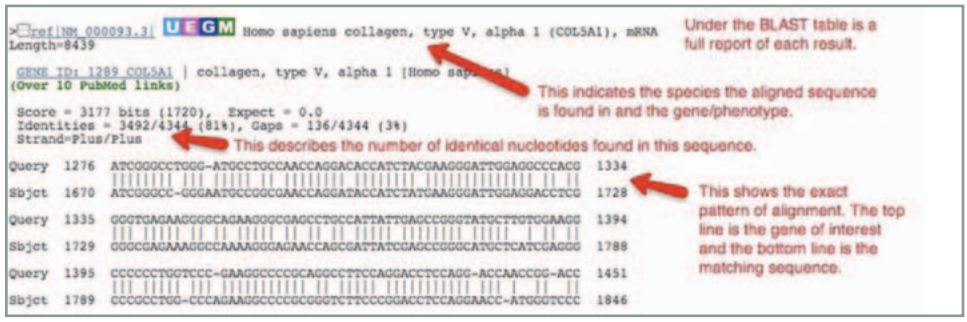

Figure 10

If you click on a particular species listed, you'll get a full report that includes the species' classification scheme, the research journal in which the gene was first reported, and the sequence of bases that appear to align with your gene of interest.

If you click on the link titled "Distance tree of results," you will see a cladogram with the species with similar sequences to your gene of interest placed on the cladogram according to how closely their matched gene aligns with your gene of interest.

Analyzing Results

Recall that species with common ancestry will share similar genes. The more similar genes two species have in common, the more recent their common ancestor and the closer the two species will be located on a cladogram.

As you collect information from BLAST for each of the gene files, you should be thinking about your original hypothesis and whether the data support or cause you to reject your original placement of the fossil species on the cladogram.

For each BLAST query, consider the following:

- The higher the score, the closer the alignment.
- The lower the e value, the closer the alignment.
- Sequences with e values less than 1e-04 (1 x 10-4) can be considered related with an error rate of less than 0.01%.

1. What species in the BLAST result has the most similar gene sequence to the gene of interest?

2. Where is that species located on your cladogram?

3. How similar is that gene sequence?

4. What species has the next most similar gene sequence to the gene of interest?

Based on what you have learned from the sequence analysis and what you know from the structure, decide where the new fossil species belongs on the cladogram with the other organisms. If necessary, redraw the cladogram you created before.

Evaluating Results

Compare and discuss your cladogram with your classmates. Does everyone agree with the placement of the fossil specimen? If not, what is the basis of the disagreement?

On the main page of BLAST, click on the link "List All Genomic Databases." How many genomes are currently available for making comparisons using BLAST? How does this limitation impact the proper analysis of the gene data used in this lab?

What other data could be collected from the fossil specimen to help properly identify its evolutionary history?

Designing and Conducting Your Investigation

Now that you've completed this investigation, you should feel more comfortable using BLAST. The next step is to learn how to find and BLAST your own genes of interest. To locate a gene, you will go to the Entrez Gene website (***http://www.ncbi.nlm.nih.gov/gene***). Once you have found the gene on the website, you can copy the gene sequence and input it into a BLAST query.

Example Procedure

One student's starting question: What is the function of actin in humans? Do other organisms have actin? If so, which ones?

1. Go to the Entrez Gene website (***http://www.ncbi.nlm.nih.gov/gene***) and search for "human actin."
2. Click on the first link that appears and scroll down to the section "NCBI Reference Sequences."
3. Under "mRNA and Proteins," click on the first file name. It will be named "NM 001100.3" or something similar. These standardized numbers make cataloging sequence files easier. Do not worry about the file number for now.
4. Just below the gene title click on "FASTA." This is the name for a particular format for displaying sequences.
5. The nucleotide sequence displayed is that of the actin gene in humans.
6. Copy the entire gene sequence, and then go to the BLAST homepage (***http://blast.ncbi.nlm.nih.gov/Blast.cgi***).
7. Click on "nucleotide blast" under the Basic BLAST menu.
8. Paste the sequence into the box where it says "Enter Query Sequence."
9. Give the query a title in the box provided if you plan on saving it for later.

10. Under "Choose Search Set," select whether you want to search the human genome only, mouse genome only, or all genomes available.

11. Under "Program Selection," choose whether or not you want highly similar sequences or somewhat similar sequences. Choosing somewhat similar sequences will provide you with more results.

12. Click BLAST.

Below is a list of some gene suggestions you could investigate using BLAST. As you look at a particular gene, try to answer the following questions:

- What is the function in humans of the protein produced from that gene?
- Would you expect to find the same protein in other organisms? If so, which ones?
- Is it possible to find the same gene in two different kinds of organisms but not find the protein that is produced from that gene?
- If you found the same gene in all organisms you test, what does this suggest about the evolution of this gene in the history of life on earth?
- Does the use of DNA sequences in the study of evolutionary relationships mean that other characteristics are unimportant in such studies? Explain your answer.

Suggested Genes to Explore	Families or Genes Studied Previously
ATP synthase	Enzymes
Catalase	Parts of ribosomes
GAPDH	Protein channels
Keratin	
Myosin	
Pax1	
Ubiquitin	