

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 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 required.

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, students 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.

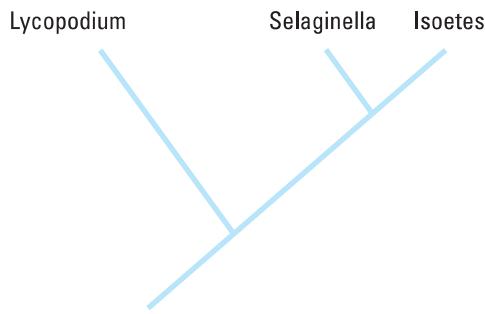


Figure 1. Simple Cladogram Representing Different Plant Species

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 species of moss.

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 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.

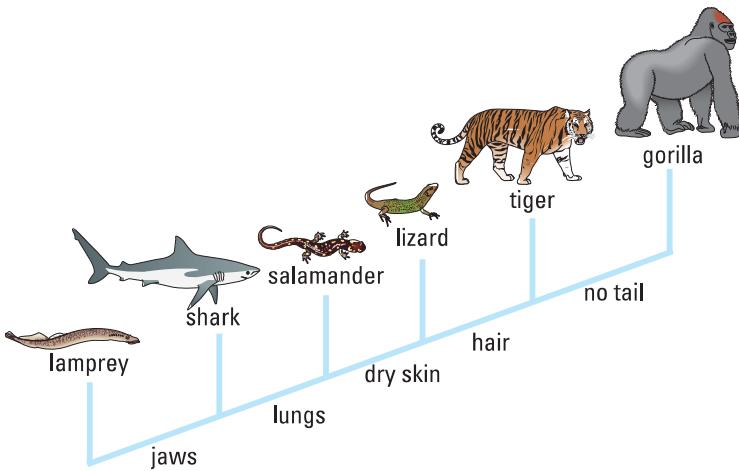


Figure 2. Cladogram of Several Animal Species

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, physical structures were used to create cladograms; however, modern-day cladistics relies more heavily on genetic evidence. 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?

■ PREPARATION

Materials and Equipment

One computer with Internet access per student or per group is needed to complete this investigation.

■ Timing and Length of Lab

It is recommended that teachers use a minimum of one hour of preparation time before the lab to download the gene files, review the screenshots, and practice uploading the gene files and analyzing the data. The prelab assessment can be completed in one 45-minute class period or assigned as homework the day before the lab. The summative assessment can be completed in one 45-minute class period.

■ Safety and Housekeeping

There are no safety precautions associated with this investigation.

■ ALIGNMENT TO THE AP BIOLOGY CURRICULUM FRAMEWORK

This investigation can be conducted while covering concepts pertaining to evolution (big idea 1) and/or genetics and information transfer (big idea 3). As always, it is important to make connections between big ideas and enduring understandings, regardless of where in the curriculum the lab is taught. The concepts align with the enduring understandings and learning objectives from the AP Biology Curriculum Framework, as indicated below.

■ Enduring Understanding

- SYI-3: Naturally occurring diversity among and between components within biological systems affects interactions with the environment.
- EVO-3: Life continues to evolve within a changing environment.

Learning Objectives

- SYI-3.A: Explain the connection between variation in the number and types of molecules within cells to the ability of the organism to survive and/or reproduce in different environments.
- EVO-3.B: Describe the types of evidence that can be used to infer an evolutionary relationship.
- EVO-3.C: Explain how a phylogenetic tree and/or cladogram can be used to infer evolutionary relatedness.

Science Skill

- 2D: Represent relationships within a biological model

ARE STUDENTS READY TO COMPLETE A SUCCESSFUL INQUIRY-BASED, STUDENT-DIRECTED INVESTIGATION?

This investigation can be conducted while covering concepts pertaining to evolution. It is recommended that the students already have a solid of understanding of the structure and function of DNA and gene expression, specifically how the order of nucleotides in DNA codes for the production of proteins.

Skills Development

Students will develop the following skills:

- Formulating, testing, and revising a hypothesis based on logic and evidence
- Using a sophisticated online bioinformatics program to analyze biological data
- Analyzing evolutionary patterns using morphological data and DNA analysis
- Analyzing preconstructed cladograms to demonstrate an understanding of evolutionary patterns
- Designing cladograms to depict evolutionary patterns
- Discussing and debating alternative interpretations of data based on evidence

Potential Challenges

This lab is designed to be flexible and can be modified as desired. The amount of information on the BLAST website is a bit overwhelming — even for the scientists who use it on a frequent basis! Reassure students that a big part of this investigation is inquiry and exploration of the data provided and that they are not expected to know every detail of the BLAST program.

It is recommended that you use a computer projector to demonstrate the steps of the procedure and work through the first gene sequence with the entire class after you work through the steps yourself. After modeling the analysis of the first gene, the students should then continue the lab in groups.

Screenshots of each step in the procedure are provided in the Student Manual version of this lab. In addition to the screenshots, the following video tutorials may be helpful. However, please note that these tutorials do not match the exact procedures of this lab.

- <http://www.youtube.com/watch?v=HXEpBnUbAMo>
- <http://www.howcast.com/videos/359904-How-To-Use-NCBI-Blast>

Additional videos can be found by searching “NCBI BLAST” on YouTube.

To help you and your students use BLAST, you might review the tutorials developed by NCBI at

http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs.

1. Navigate to NCBI and the BLAST page as usual.
2. At the top, click “Help” to go to the BLAST Documentation Page.
3. Under the heading “About BLAST,” there is a handbook that has both novice-friendly sections as well as information for experienced users.

BLAST documentation can be viewed online as well as downloaded and distributed for educational purposes. There are practice data sets that teachers can use to demonstrate how to read the results, and all users are free to email or call the BLAST Help Line.

Those who want to dig deeper can visit the NCBI Educational Resources Web page (<http://www.ncbi.nlm.nih.gov/education/>) for videos, tutorials, project descriptions, and other tools designed for teaching.

This inquiry-based investigation has no definite right answer. This will frustrate some students. Reassure them that their performance in this investigation is based on the use of the data they collect to construct and test a reasonable hypothesis.

Students are unlikely to understand what BLAST is doing when it searches for sequence similarities. A simple analogy is the sticky note and the library. Tell students that they have a three-word phrase written on a sticky note. Their job is to go to the school library, look for every book that has that three-word phrase, and write down the exact page number and name of every book they find. Next, they must search for every book that has their three-word phrase, even if the spelling is not perfect. They must keep doing this until they find every last book that has a part of their three-word phrase. Their last chore is to put all the names and page numbers of the books they found in order, from most to least similar to their original phrase. If students are not impressed with the library analogy, tell them to use Google to search for a three-word phrase (with near matches) and categorize the hits for the entire Web. That is essentially what BLAST is doing in a few seconds.

To clarify this idea, ask students to align the first five bases or amino acids in three to five sequences (such as the sequences they download from <http://blogging4biology.edublogs.org/2010/08/28/college-board-lab-files/>).

Which ones are more similar/less similar to one another? Once students understand the principle behind matching alignments, they can even calculate the percentage similarity by dividing the number of matching sequence bases by the total number compared. The following is a simplified example of the concept:

Organism A Sequence: ATGATCCAGT

Organism B Sequence: ACGACTCAGT

Organism C Sequence: TTGATCCAGT

In addition, you can have students align gene sequences on paper to simulate what the BLAST program is doing for them. When uploaded into the BLAST website, each gene sequence will appear in the query sequence. Students can copy the gene sequence on paper and compare it to the results once the gene is submitted on the BLAST website.

THE INVESTIGATIONS

Getting Started: Prelab Assessment

You may assign the following questions for homework; as a think, pair/group, share activity, in which pairs or small groups of students brainstorm ideas and then share them with other groups; or as a whole-class discussion to assess students' understanding of key concepts pertaining to cladograms:

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 in the process of 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.

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

Species	Gene Percentage Similarity	Protein Percentage Similarity
Chimpanzee (<i>Pan troglodytes</i>)	99.6%	100%
Dog (<i>Canis lupus familiaris</i>)	91.3%	95.2%
Fruit fly (<i>Drosophila melanogaster</i>)	72.4%	76.7%
Roundworm (<i>Caenorhabditis elegans</i>)	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 may also assign 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>

Designing and Conducting Independent Investigations

Now that students have completed this investigation, they should feel more comfortable using BLAST. The next step is to have students find and BLAST their own genes of interest. They might investigate something they have heard the name of, or you could ask them to think about and explore an enzyme or protein they studied before (e.g., DNA polymerase). They could look online for additional information to inform their questions (e.g., Are there diseases where DNA polymerase does not function normally? Do viruses make DNA polymerase?) Another option is to ask students to identify a disease that they know is related to proteins, such as spinocerebellar ataxia or various storage diseases. Search for the normal versus mutant versions of the protein or DNA. What is different about their sequences?

To locate a gene, go to the Entrez Gene* section of the NCBI website (<http://www.ncbi.nlm.nih.gov/gene>) and search for the gene. Once you have found the gene on the website, copy the gene sequence and input it into a BLAST query. Ask students to determine the function of proteins in humans and then to predict if they will find the same protein (and related gene) in other organisms. Do students understand that BLAST analyses provide only one piece of evidence about speciation and the phylogenetic relationships of organisms? Is DNA evidence more or less important to evolutionary studies as compared to morphological evidence?

Example Procedure

1. On the Entrez Gene website, search “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 “NM 001100.3.”
4. Just below the gene title, click on “FASTA.”
5. The nucleotide sequence displayed is that of the actin gene in humans.
6. Copy the gene sequence and 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 you want highly similar sequences or somewhat similar sequences. Choosing somewhat similar sequences will provide you with more results.
12. Click BLAST.

*Entrez Gene is a global database of genetic information. When you use it, you search a number of databases for specific gene information. Entrez Gene is separate from BLAST in that it searches for a specific gene's sequence. BLAST then compares the sequence of the single, specific gene with other sequences in the database. An example procedure of how Entrez Gene and BLAST work together is described in the following example.

Suggested Genes to Explore		
Actin	GAPDH	Pax1
ATP synthase	Keratin	Ubiquitin
Catalase	Myosin	Zinc finger

Examining Gene Sequences Without BLAST

One of the benefits of learning to use BLAST is that students get to experience a scientific investigation in the same manner as the scientists who use this tool. However, it is not necessary to BLAST common genes of interest. Many researchers have saved common BLAST searches into a database. The following video demonstrates how to access these saved BLAST queries:

<http://www.wonderhowto.com/how-to-use-blast-link-244610/view/>.

Summative Assessment

Have students consider the following when analyzing the gene sequences:

- 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 (1×10^{-4}) can be considered related with an error rate of less than 0.01%.

Students should analyze and discuss the data and try to form logical hypotheses based on evidence. While the evidence is leading toward a close relatedness with birds and/or reptiles, you should assess students on their understanding of cladogram construction, in general, and the evidence they use to defend their hypothesis.

The following questions are suggested as guidelines to assess students' understanding of the concepts presented in the investigation, but you are encouraged to develop your own methods of postlab assessment:

- Are students able to make predictions about where the fossil species could be placed on the cladogram based on information they collected from the BLAST queries?
- How did the students handle any disagreements about the cladogram? Was their reasoning evidence based?
- Did students have an adequate background in genetics to understand the data they had to analyze in this investigation?
- Are students able to construct their own cladograms using provided data?

Determine if students truly understand the evolutionary patterns seen in cladograms by asking them to include concepts such as speciation, extinction, and natural selection when describing a particular cladogram.

SUPPLEMENTAL RESOURCES

Other Labs

Another inquiry-based cladogram investigation that uses simple household items can be found at the following website:

<http://blogging4biology.edublogs.org/2010/08/26/cladogram-lab-activity/>

This cladogram investigation also uses simple household items:

http://www.pbs.org/wgbh/nova/teachers/activities/2905_link.html

Online Activities

The following online activities are included in the Student Manual:

“The Evolution of Flight in Birds”: This activity provides a real-world example of how cladograms are used to understand evolutionary relationships:

<http://www.ucmp.berkeley.edu/education/explorations/reslab/flight/main.htm>

“What did T. rex taste like?”:

<http://www.ucmp.berkeley.edu/education/explorations/tours/Trex/index.html>

References

The plant group cladogram table (and answer key) is available at

<http://petrifiedwoodmuseum.org/Taxonomy.htm>

The following resources illustrate common misconceptions in reading and interpreting phylogenetic trees:

Baum, David A., Stacey DeWitt Smith, and Samuel S. S. Donovan. “The Tree-Thinking Challenge.” *Science* 310, no. 5750 (November 11, 2005): 979–980.

Baum, David A. and Susan Offner. “Phylogenetics & Tree-Thinking.” *70(4)*, (2008): 222–229.

Gregory, T. Ryan. “Understanding Evolutionary Trees.” *Evolution: Education and Outreach* 1 (2008): 121–137.

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.

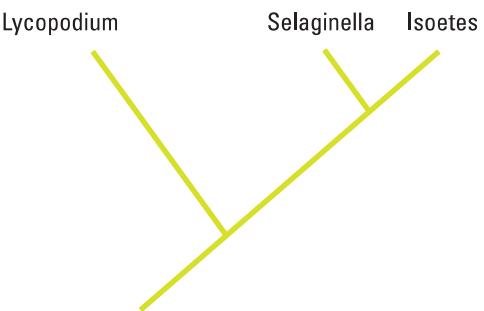


Figure 1. Simple Cladogram Representing Different Plant Species

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.

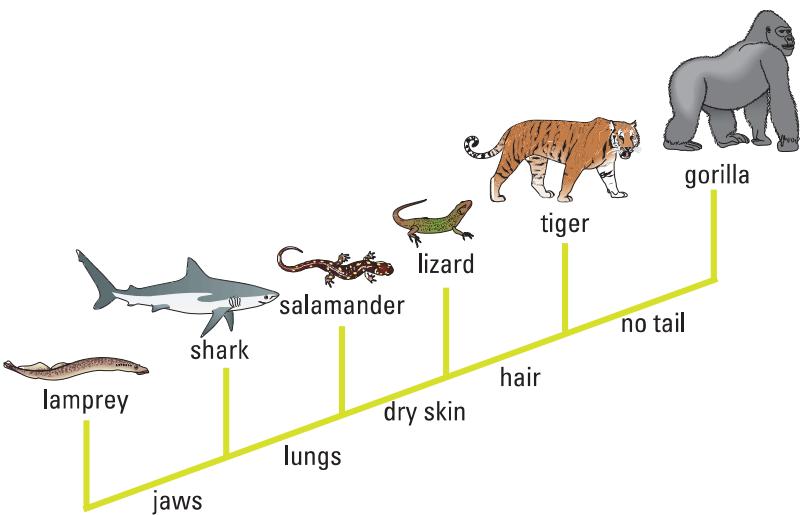


Figure 2. Cladogram of Several Animal Species

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

- SYI-3.A: Explain the connection between variation in the number and types of molecules within cells to the ability of the organism to survive and/or reproduce in different environments.
- EVO-3.B: Describe the types of evidence that can be used to infer an evolutionary relationship.
- EVO-3.C: Explain how a phylogenetic tree and/or cladogram can be used to infer evolutionary relatedness.

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:

- 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

- 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 (<i>Pan troglodytes</i>)	99.6%	100%
Dog (<i>Canis lupus familiaris</i>)	91.3%	95.2%
Fruit fly (<i>Drosophila melanogaster</i>)	72.4%	76.7%
Roundworm (<i>Caenorhabditis elegans</i>)	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>

Procedure

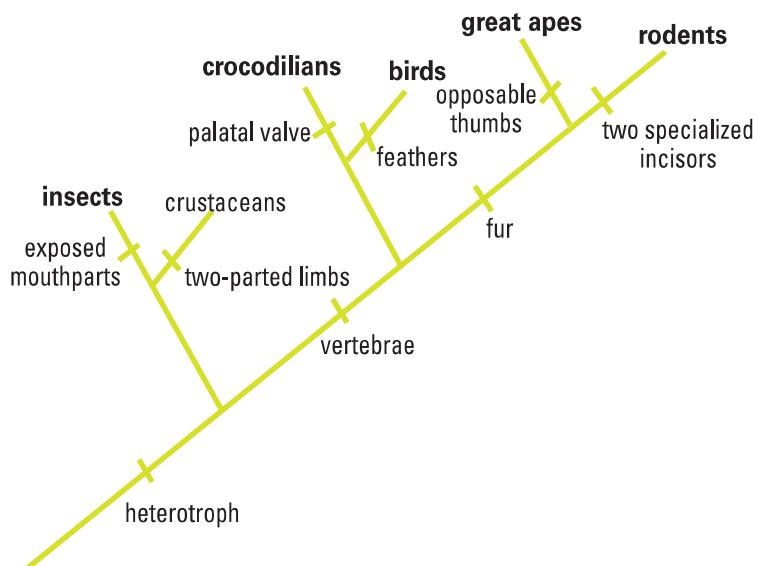
A team of scientists has uncovered the fossil specimen in Figure 3 near Liaoning Province, China. Make some general observations about the morphology (physical structure) of the fossil, and then record your observations in your notebook.

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.

© AMNH, Mick Ellison



Figure 3. Fossil Specimen

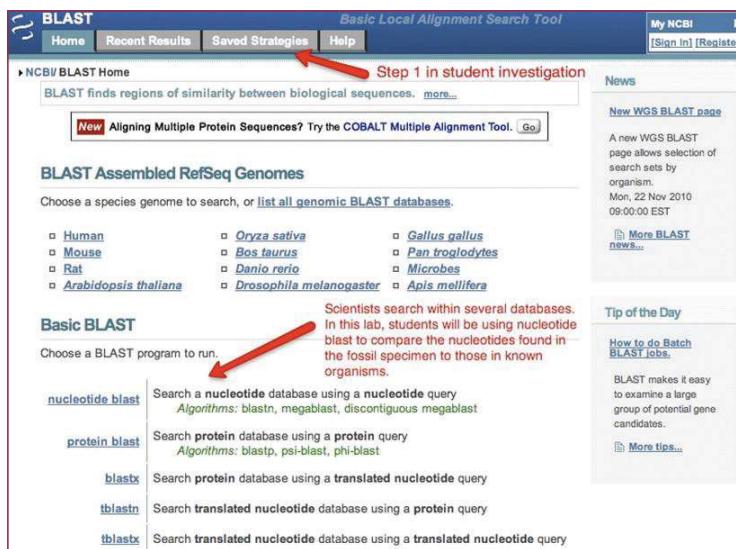
**Figure 4. Fossil Cladogram**

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 <http://blogging4biology.edublogs.org/2010/08/28/college-board-lab-files/>.

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

- Go to the BLAST homepage: <http://blast.ncbi.nlm.nih.gov/Blast.cgi>
- 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.

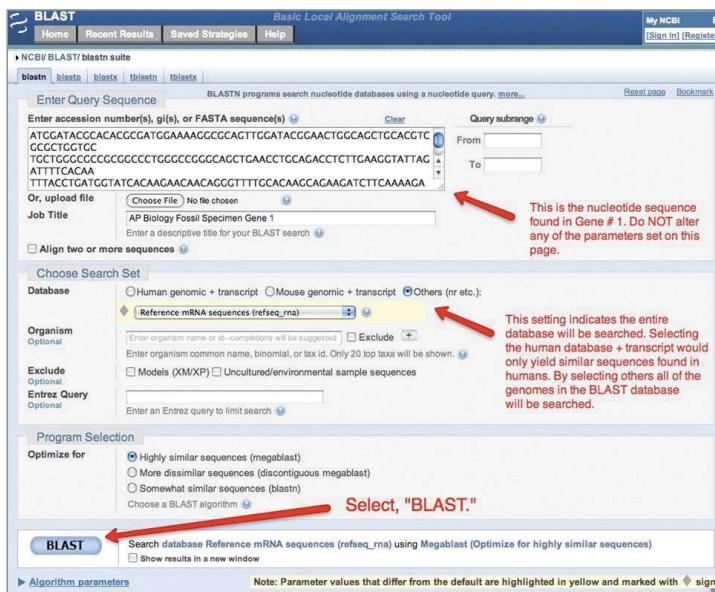


Figure 7

- 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.

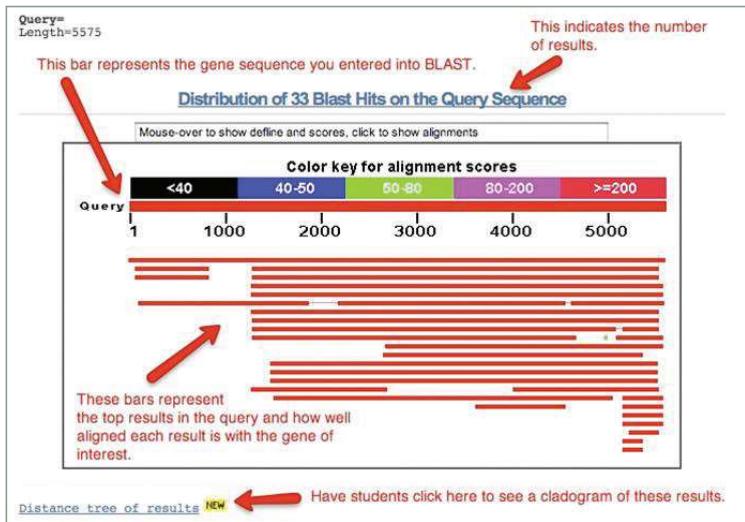


Figure 8

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.

Sequences producing significant alignments:	Score (bits)	E value	
ref NM_204790_1 Gallus gallus collagen, type V, alpha 1 (COL5A1)	1.029e+04	0.0	UEGM
ref XM_015062461_1 PREDICTED: Ornithorhynchus anatinus simili...	2476	0.0	GM
ref XM_01372383_1 PREDICTED: Monodelphis domestica similar to ...	1465	0.0	GM
ref XM_01317369_1 PREDICTED: Equus caballus similar to Coll...	3227	0.0	UGM
ref NM_000993_1 Homo sapiens collagen, type V, alpha 1 (COL5A1)	3177	0.0	UEGM
ref XM_054662_1 PREDICTED: Taenioptygia guttata misc_RNA (LOC...	3001	0.0	GM
ref NM_0104971_1 Sus scrofa collagen, type V, alpha 1 (COL...	2992	0.0	LGM
ref XM_02320740_1 PREDICTED: Alirocopra melanoleuca collagen	2976	0.0	GM
ref NM_537892_1 PREDICTED: Canis familiaris similar to proco...	2545	0.0	LGM
ref XM_01118214_2 PREDICTED: Pan troglodytes similar to coll...	2314	0.0	GM
ref XM_028475_1 PREDICTED: Macaca mulatta hypothetical protein...	2138	0.0	GM
ref NM_02935320_1 PREDICTED: Xenopus (Silurana) tropicalis	2132	0.0	GM
ref NM_01372244_1 PREDICTED: Monodelphis domestica similar to ...	1663	0.0	GM
ref NM_000994_1 Homo sapiens collagen, type V, alpha 1 (COL5A1)	1590	0.0	GM
ref XM_013182_1 PREDICTED: Monodelphis domestica similar to ...	1563	0.0	GM
ref NM_0029170_1 PREDICTED: Bos taurus collagen alpha-1(VI)	1218	0.0	GM
ref XM_01322721_1 PREDICTED: Sus scrofa collagen alpha-1(VI)	953	0.0	GM
ref NM_0107939_1 Danio rerio collagen type XI alpha-2 (col...	944	0.0	LGM
ref XM_0283319_1 PREDICTED: Pongo abelii hypothetical protein...	717	0.0	GM
ref XM_02724221_1 PREDICTED: Cryctocula cunicularia collagen	431	3e-117	GM
ref NM_00118209_2 PREDICTED: Macaca mulatta collagen alpha-1	403	6e-105	UGM
ref XM_02833217_1 PREDICTED: Pongo abelii collagen alpha-1(VI)	398	3e-107	UEGM
ref NM_134521_1 Ratmus norvegicus collagen, type V, alpha 1	398	3e-107	UEGM
ref XM_114314_1 PREDICTED: Homo sapiens hypothetical LOC1005...	390	5e-105	GM
ref XM_0108955_1 PREDICTED: Homo sapiens hypothetical LOC1005...	390	5e-105	GM
ref NM_015734_2 Mus musculus collagen, type V, alpha 1 (COL5A1)	350	8e-93	UEG
ref XM_03122271_1 PREDICTED: Sus scrofa hypothetical protein...	320	7e-84	GM
ref NM_02833148_1 PREDICTED: Pongo abelii collagen alpha-1(VI)	250	9e-63	GM
ref XM_01168200_1 PREDICTED: Pan troglodytes similar to proco...	233	9e-58	GM
ref XM_02733481_1 PREDICTED: Saccoglossus kowalewskii fibrin	62.1	4e-06	GM

Click the reference number for a specific sequence to learn more about that sequence.

This is the species and gene name that matches the gene of interest. Phenotype is sometimes identified as well.

The score (bits) refers to how many gaps or substitutions are associated with the sequence. The higher the score the more similar the alignment.

The e value is the likelihood that a match occurred purely by chance. The lower the e value, the better the match.

These links refer to related entries in other BLAST databases. They are not used in this lab.

Alignments

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.

>ref NM_000093.3	UEGM	Homo sapiens collagen, type V, alpha 1 (COL5A1), mRNA	Under the BLAST table is a full report of each result.
Length=8439			
GENE ID: 1289 COL5A1 collagen, type V, alpha 1 [Homo sapiens]	(Over 10 PubMed links)		This indicates the species the aligned sequence is found in and the gene/phenotype.
Score = 3177 bits (1720), Expect = 0.0			
Identities = 3492/4344 (81%), Gaps = 136/4344 (3%)			
Strand=Plus/Plus			This describes the number of identical nucleotides found in this sequence.
Query 1276 ATCGGGCCTGGG-ATGCCCTGCCAACCCAGGACACCATCTACGAAGGGATTGGAGGCCACG	1334		
Sbjct 1670 ATCGGGCC-GGGATTCGGCGAACCCAGGATACCATCTATGAAGGGATTGGAGCACCTCG	1728		
Query 1335 GGGTGAGAAGGGCAGAAGGGCAGCCATTATTGAGCCGGTATGCTGTGGAAGG	1394		
Sbjct 1729 GGGCGAGAAAAGGGAAACCCAGGGAGATTATCGAGCCGGATGCTCATGGAGG	1788		
Query 1395 CCCCCCTGGTCCC-GAAGGCCCGCAGGCCCTCCAGGACCTCCAGG-ACCAACCG-ACC	1451		
Sbjct 1789 CCCCGCTGG-CCCGAGGCCCCCGGGCTTCCCGAACCTCCNGAAC-ATGGGTCCC	1846		

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 $1e-04$ (1×10^{-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	