EDVOTEK®

THE BIOTECHNOLOGY EDUCATION COMPANY®

AP BIOLOGY

BIG IDEA 1 INVESTIGATION 3

Edvo-Kit #AP03

Determining Evolutionary Relationships Using BLAST

Experiment Objective:

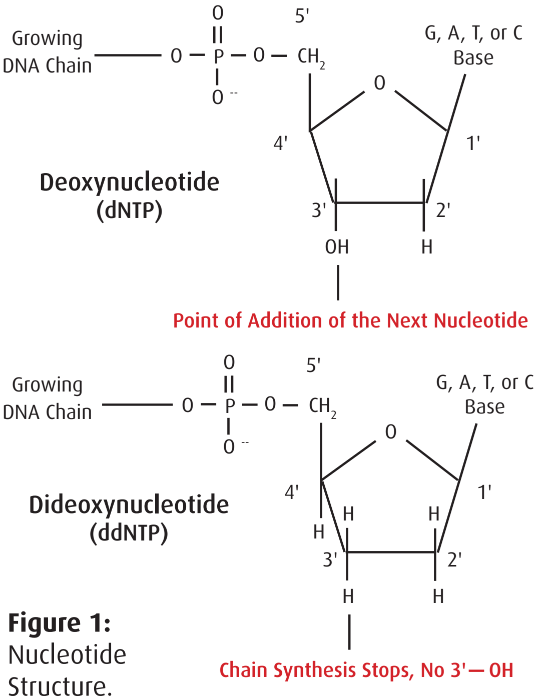
The objective of the experiment is for students to become familiar with databases that can be used to investigate gene sequences and to construct cladograms that provide evidence for evolutionary relatedness among species.

1.800.EDVOTEK • Fax 202.370.1501 • www.edvotek.com • info@edvotek.com

AP03.170426

Background Information

The basic unit of all living organisms, from bacteria to humans, is the cell. Contained within the nucleus of these cells is a molecule called deoxyribonucleic acid (or DNA). Today, we know that DNA is the blueprint used to build an organism – our genetic makeup, or genotype, controls our phenotype (observable characteristics). The directions coded for by our genes controls everything from growth and development to cell specification, neuronal function, and metabolism.



Point of Addition of the Next Nucleotide

Chain Synthesis Stops, No 3'— OH

Figure 1: Nucleotide Structure.

A strand of DNA is composed of building blocks known as nucleotides (Figure 1, top). Each deoxynucleotide (dNTP) comprises three basic parts: a phosphate group, a deoxyribose sugar, and a nitrogen-containing base (adenine, cytosine, guanine, or thymine — abbreviated as A, C, G, or T). The order of these nucleotides gives rise to genes, each with a unique sequence. The 3′ hydroxyl group on the sugar of one nucleotide forms a covalent bond with the 5′ phosphate group of its neighbor, making DNA a stable scaffold for genetic information. The nature of this bond results in DNA strands with a distinct polarity, ensuring that the strand of DNA is read in the correct direction (Figure 2).

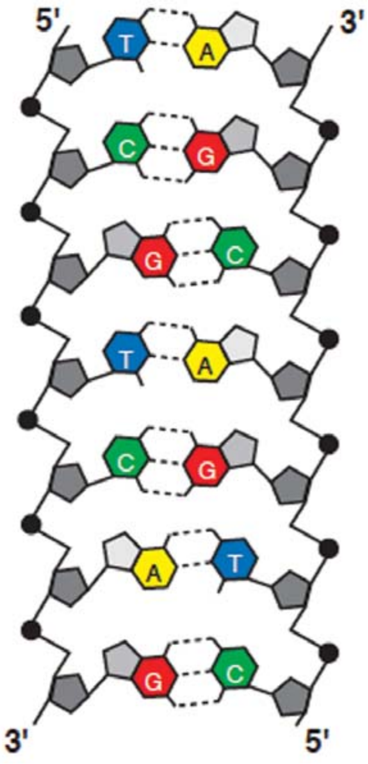


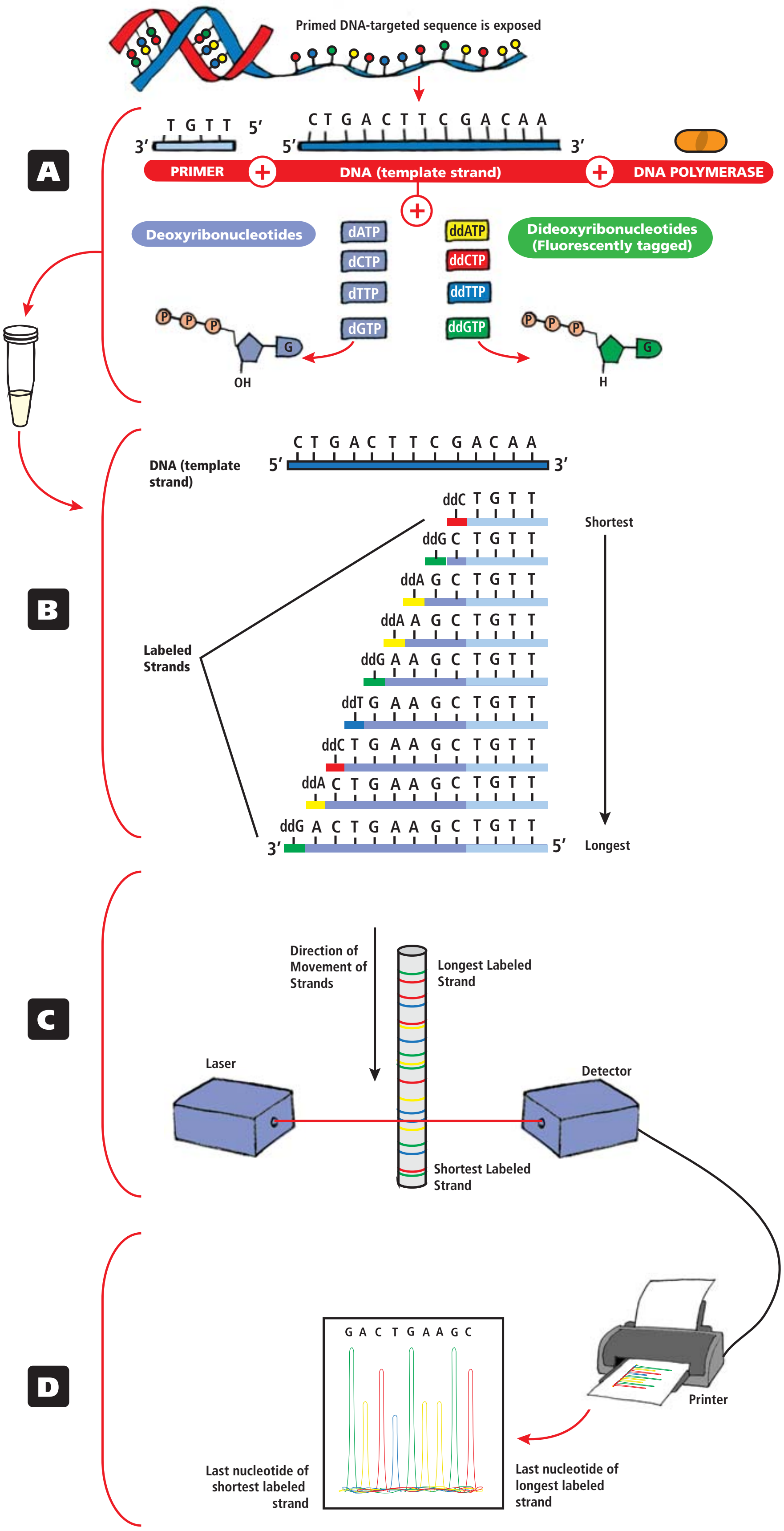
Figure 2: Structure of DNA.

In the early 1970′s, Frederick Sanger developed a method to determine the nucleotide sequence of DNA. This method creates a set of copies that are complementary to the original DNA sequence using a DNA primer to target the site to be sequenced, the enzyme DNA polymerase I (DNA Pol I), and free nucleotides. DNA Pol I uses the primer to start synthesis of the new strand of DNA in the 5′-3′ direction using the existing DNA as a template. To this mixture, Sanger added dideoxynucleotides (ddNTPs). These nucleotide analogs lack the 3′ hydroxyl group (Figure 1), making it impossible for the polymerase to add another nucleotide to the end of growing strand. This creates a series of DNA fragments of differing size that can be used to map the location of each nucleotide in a given piece of DNA (Figure 3B).

In the 1990′s, the entire process of collecting and analyzing sequencing data became automated (Figure 3). Despite this, the basic biochemistry of the chain-termination method of DNA sequencing remains unchanged. Fluorescently labeled ddNTPs are used in the sequencing reactions, both to terminate the growing nucleotide chain and to label the DNA for analysis. Each ddNTP is tagged with a different color fluorophore, allowing all four sequencing reactions to be performed in one tube.

Figure 3: Sanger DNA Sequencing.

(A) Setting up the sequencing reaction.
(B) Incorporation of the ddNTPs create different size DNA fragments.
(C) The labeled mixture is sequenced using capillary gel electrophoresis. A laser detects the fluorescent label on each of the ddNTPs.
(D) The information is analyzed using a computer.



To analyze the fluorescent sequencing reactions, automated machines utilize a polyacrylamide gel formed in a thin capillary tube. While the DNA fragments are separating through the gel matrix, a laser beam is focused on the capillary. The laser excites the fluorophores on the DNA, and the detector captures the fluorescent emission (Figure 3 C, D). This set-up allows the instrument to identify the bases in the sequencing reaction in real time as they are being separated through the capillary. These high-throughput machines can provide fast, high-quality sequence for a fraction of the cost of traditional sequencing strategies. As the demand for high-quality DNA sequencing information continues to grow, "next-generation" strategies are being developed to speed up sequencing efforts while simultaneously reducing costs.

THE BIRTH OF GENOMICS

With the rise of automated sequencing techniques in the early 1990's, scientists began large-scale genome sequencing efforts in order to understand the biological complexity of an organism at the DNA level. This new field of biology, called genomics, analyzes the sequence and structure of an organism's genome, as well as the interactions between genes themselves. In addition to the human genome, many scientifically and commercially relevant organisms have had their complete genome sequenced. Notable organisms with sequenced genomes include the bacteria *E. coli*, the baker's yeast *S. cerevisiae*, the nematode *C. elegans*, corn (*Z. mays*) and our closest living relative, the chimpanzee (*P. troglodytes*). As the DNA sequencing technology becomes faster and less expensive, scientists have been able to sequence the DNA of many other organisms, including many lesser-known species that are important for establishing evolutionary relationships. Furthermore, individuals from within a single species are also being sequenced and analyzed to explore genetic diversity. As a result of these projects, a vast amount of DNA sequence information has been made available for researchers.

However, an organism's DNA sequence was of limited use unless it could be converted to biologically useful information. In the early stages of genomic studies, researchers recognized the need for data management systems to store and analyze large quantities of sequence information. As computer scientists developed technology to address these needs, they established the interdisciplinary field of science known as Bioinformatics. This discipline blends computer science, biology, and information technology to develop extensive databases to analyze biological information. These databases are universally accessible online, making it easier for scientists around the world to share biological data and come up with greater discoveries.

Bioinformatics has allowed scientists to unlock mysteries coded for by DNA. For example, since the human genome was completed in 2003, the sequence information has been used to map specific genes to their chromosomal location and to identify novel genes. Since the genome varies about 0.2% between individuals (about one base in every 500 is different), specific variations in the DNA sequence can be used as markers for disease predisposition. Potential protein coding genes are easily identified by the presence of stop and start codons. Likewise, special programs have been developed to analyze non-coding sequences of DNA called promoters, which initiate gene expression by regulating the amount of RNA produced within a cell. The annotation of novel DNA sequences will continue as new sequence information is added and more powerful programs mine the information, allowing scientists to understand the function of every base within a given genome.

Furthermore, since the genome of many model organisms has also been sequenced, DNA sequence comparison software like the Basic Local Alignment Search tool (or BLAST) has allowed scientists to identify genes that are similar to those that are important for human health and development. Scientists can learn more about these genes by studying their function in a model organism. For instance, about 75% of the genes that cause disease in humans have homologs in the fruit fly, *D. melanogaster*. In fact, the fly model of Alzheimer's disease has provided new information, which has allowed scientists to identify novel targets for treatment.

The genetic revolution will continue to yield new discoveries. While scientists continue to identify genes that cause disease or phenotypic differences (tall versus short), there is a growing danger to see individuals as the sum of these DNA sequences. A simplistic view of this genetic information has the potential to cause inconvenience or harm. For example, researchers are working to identify genetic factors involved in psychiatric disorders like schizophrenia. While current research shows there is a strong genetic component, epidemiological studies demonstrate that environmental influences also play a crucial role. Additionally, prenatal screening for diseases in human embryos remains controversial. Therefore, the ethical, legal, and social implications of genetic information represent another major challenge for the field of genomics. Scientists, bioethicists, and lawmakers must work together to balance improvements to human health with the ethical implications of the genetic revolution.

DETERMINING EVOLUTIONARY RELATIONSHIPS USING DNA SEQUENCES

Traditionally, scientists sorted living organisms into different groups, or clades, using distinctive physical characteristics known as shared derived characters. This data is organized into a branched chart called a cladogram (or phylogenetic tree), which visually represents the evolutionary relationships between organisms (Figure 4). When analyzing a cladogram, the endpoint of a branch represents a specific grouping of organisms, whereas the intersection between two branches represents the shared common ancestor. Many times, the length of the branch is proportional to the amount of time since the two organisms diverged, meaning that they more recently shared a common ancestor. Each branch intersection corresponds to the evolution of a shared derived character that applies to each species above that point on the tree. For example, the cladogram in Figure 4 shows that dry skin evolved after jaws and lungs. This means that the iguana, panther and chimpanzee all have dry skin, whereas the hagfish, skate and frog do not.

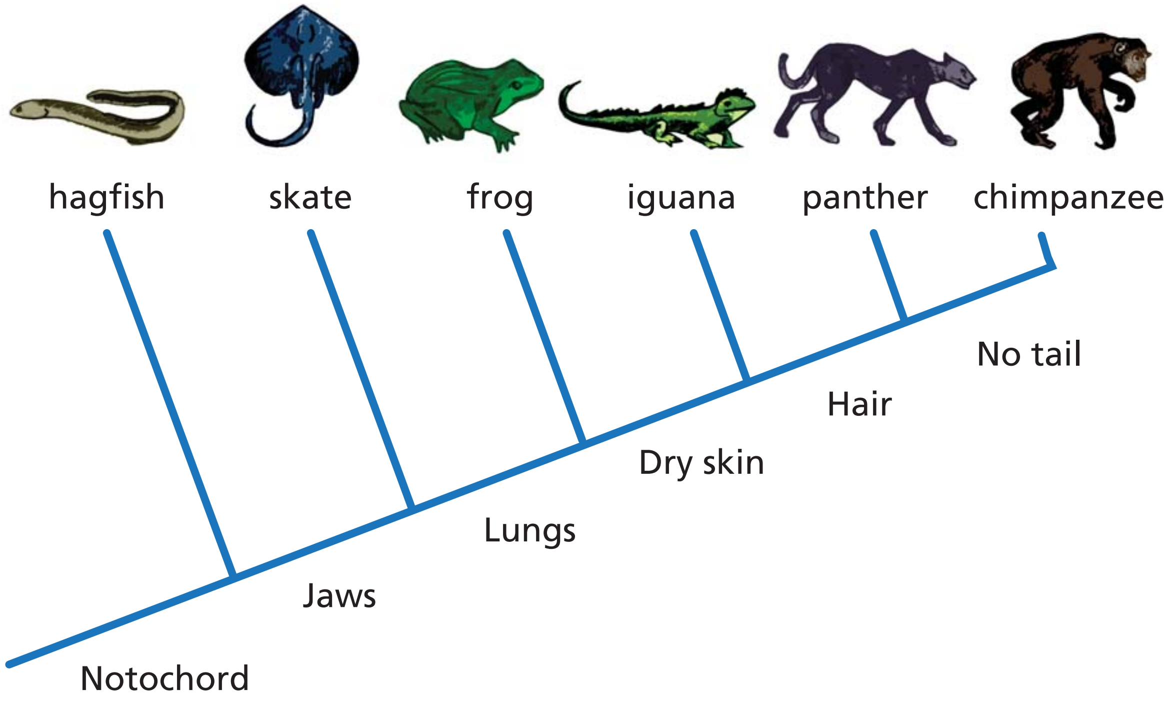


Figure 4: Using Shared Derived Characters to Create a Cladogram.

With the advent of fast, affordable DNA sequencing, scientists have been exploring the genomes of many animals. Molecular biologists can use the data from these genome-wide sequencing projects to establish evolutionary relationships. The new DNA sequence information is compared to that of several other species to determine which areas of the genome are different. Comparisons between the DNA sequences of two closely related species will uncover recent mutations that may contribute to phenotypic changes between the species; in contrast, comparisons between very different species help to identify genes that have been conserved for millions of years. Many times, at least one distantly related group is included in the analysis, as it allows the researcher to determine the polarity or evolutionary direction of sequence changes. When a cladogram has evolutionary direction it is known as a rooted tree.

Sequence differences between genomes can be as large as multi-gene duplication events or chromosomal rearrangements or as small as single base changes in the DNA sequence. Because single base changes occur at a relatively constant rate over evolutionary time they can be used like a molecular clock. For example, the human and chimpanzee genomes differ by about 4%, leading scientists to calculate that the most recent common ancestor lived about 6 million years ago. More ancient divergence times are harder to estimate but still provide important insights into gene function and evolution. Humans and fruit flies only share about 60% of the DNA sequence and estimates for their divergence time range from 678 to 917 million years ago. However, the regions that are similar between humans and fruit flies indicate that these genes are essential for normal growth and development in both organisms.

In this investigation, students will use explore the evolutionary relationships between different organisms using traditional and molecular methods. Students will create and interpret cladograms using shared derived characters. Modern techniques of bioinformatics and DNA sequencing technology are used to analyze several related genes. The molecular data is then used to create phylogenetic trees describing the relationships between different organisms. After completing this exploration, students will be able to connect and apply concepts pertaining to genetics and evolution.

Experiment Overview and General Instructions

EXPERIMENT OBJECTIVE

The objective of the experiment is for students to become familiar with databases that can be used to investigate gene sequences and to construct cladograms that provide evidence for evolutionary relatedness among species.

In this investigation, students will:

1. Create cladograms that depict evolutionary relationships among organisms.

2. Analyze biological data with online bioinformatics tools.

3. Connect and apply concepts pertaining to genetics and evolution.

LABORATORY NOTEBOOKS

Scientists document everything that happens during an experiment, including experimental conditions, thoughts and observations while conducting the experiment, and, of course, any data collected. Today, you'll be documenting your experiment in a laboratory notebook or on a separate worksheet.

BEFORE STARTING THE EXPERIMENT:

- Carefully read the introduction and the protocol. Use this information to form a hypothesis for this experiment.
- Predict the results of your experiment.

DURING THE EXPERIMENT:

- Record your observations.

AFTER THE EXPERIMENT:

- Interpret the results - does your data support or contradict your hypothesis?
- If you repeated this experiment, what would you change? Revise your hypothesis to reflect this change.

Investigation I: Constructing and Interpreting Cladograms Using Shared Derived Characters

EXERCISE 1: USING A CLADOGRAM

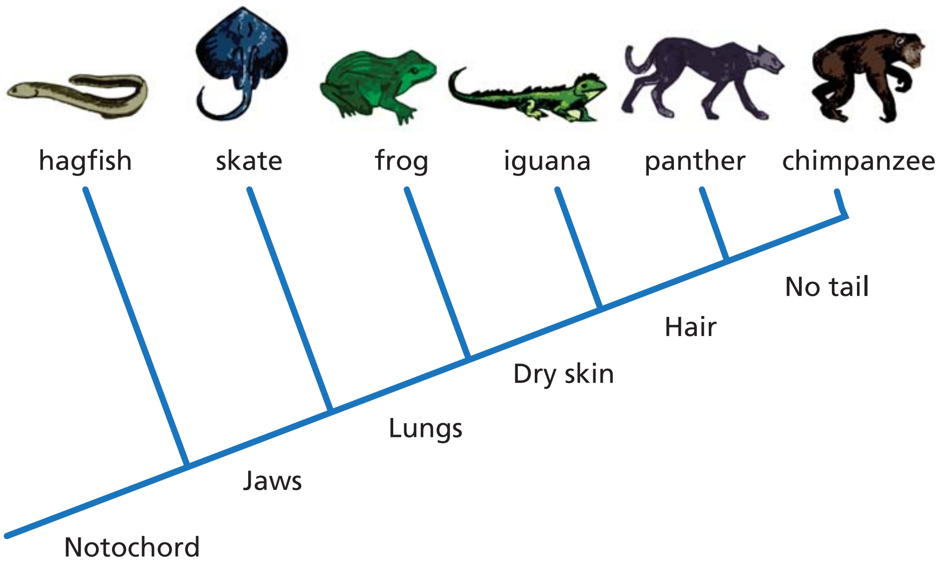
Answer the following questions using the cladogram, below.

1. Of the six organisms featured, which can grow hair?

2. According to the cladogram, which shared derived character is present in all six organisms?

3. Which five shared derived characters are present in panthers?

4. Which evolved first – notochord or dry skin? Jaws or hair?



Investigation I: Constructing and Interpreting Cladograms Using Shared Derived Characters, cont.

EXERCISE 2: CONSTRUCTING AND INTERPRETING A CLADOGRAM

In the space provided, construct a cladogram using the following data about animals and answer the related questions, below.

- **True tissues** – specialized groupings of cells that perform a specific function.
- **Bilateral Symmetry** – the animal's midline (sagittal plane) divides an organism into two roughly equivalent halves.
- **Notochord** – a flexible bar that provides skeletal support for an organism.

Organism (Phylum)	True tissues	Bilateral Symmetry	Notochord
Sea Sponge *(Porifera)*	-	-	-
Planaria *(Platyhelmenthes)*	+	+	-
Lancelet *(Chordata)*	+	+	+
Comb Jelly *(Cnidaria)*	+	-	-
Total	3	2	1

1. Of the four organisms, how many have true tissues?

2. From your cladogram, which evolved first? Bilateral symmetry or a notochord?

EXERCISE 3

Use the following data to construct a cladogram of the major plant groups in the space provided below.

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

Investigation II: Building Cladograms Using DNA Sequence Information

EXERCISE 1

Using DNA sequence alignment and other bioinformatics techniques, scientists have shown that humans and mice share about 92% of their genes. In comparison, humans and chickens share about 65% of their genes. Which animal would be closer to humans on a cladogram?

In the space provided, explore the evolutionary relationship between humans, mice, and chickens using a cladogram.

Investigation II: Building Cladograms Using DNA Sequence Info, cont.

EXERCISE 2

Cytochrome C (CytC) is an important protein for cellular respiration. During oxidative phosphorylation, CytC transports electrons between complexes III and IV. Because the function of CytC is essential for metabolism, this gene is conserved in the mitochondria of all eukaryotic organisms. The following table compares the DNA and protein sequences of CytC from humans to several different species. The comparisons can be used to create a cladogram.

Pairwise Comparisons of Human (*Homo sapiens*) CytC Nucleotide and Protein Sequence

Organism	Nucleotide Sequence Similarity to Human (percent)	Protein Sequence Similarity to Human (percent)
Chimpanzee *(Pan troglodytes)*	98.7	100
Mouse *(Mus musculus)*	90.0	91.3
Chicken *(Gallus gallus domesticus)*	81.4	82.7
Fruit Fly *(Drosophila melanogaster)*	72.1	76.9

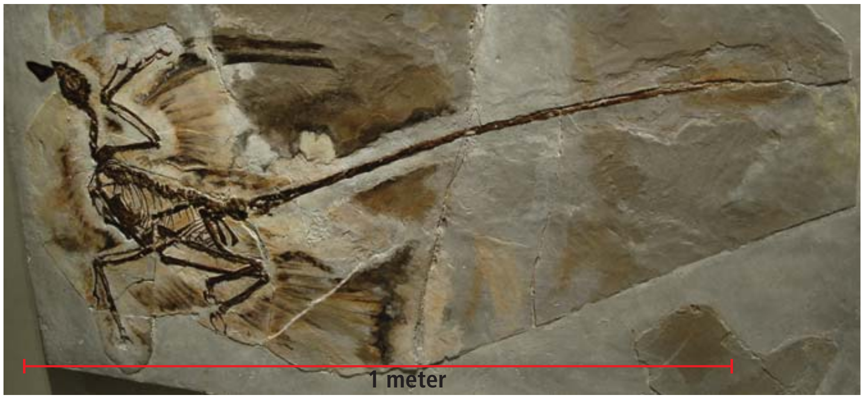
1. Which are more similar, the protein sequences or the DNA sequences? Why?

2. Draw a cladogram relating the evolutionary relationships between the species using the CytC gene sequence information.

Investigation III: Investigating Fossil Specimens Using BLAST

A. MORPHOLOGICAL OBSERVATION OF A FOSSIL SPECIMEN AND FORMATION OF A HYPOTHESIS

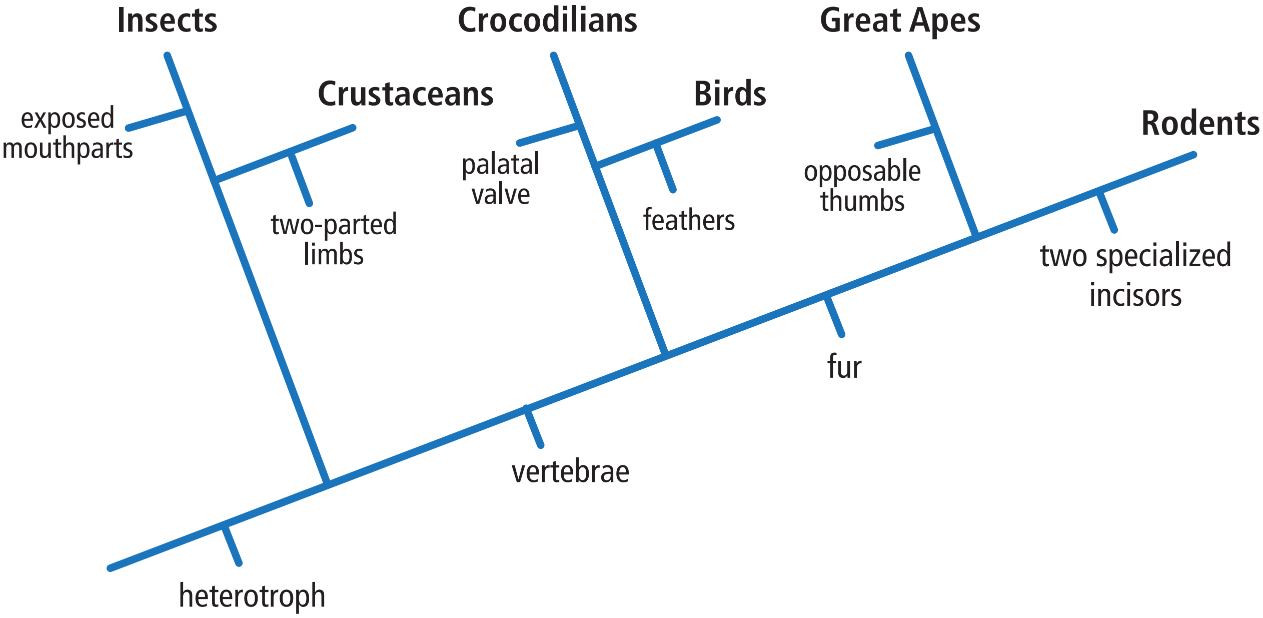
While hunting for fossils in southwestern Wyoming, you and your team of paleontologists uncovered the following specimen. After careful examination, it is determined to be a previously unidentified species, which warrants a careful study of the fossil.



© M. Martyniuk / Wikimedia Commons / CC-BY-SA 4.0

1. Carefully examine the fossil. Record some general observations about the morphology of the animal (e.g. bone structure, prominent physical features, number of appendages, evidence of body covering) in the space below. Does the fossil resemble any modern animals?

2. Based on your observations, form an initial hypothesis as to where the organism fits into the cladogram. Mark your hypothesis on the cladogram, below.



Investigation III: Investigating Fossil Specimens Using BLAST, cont.

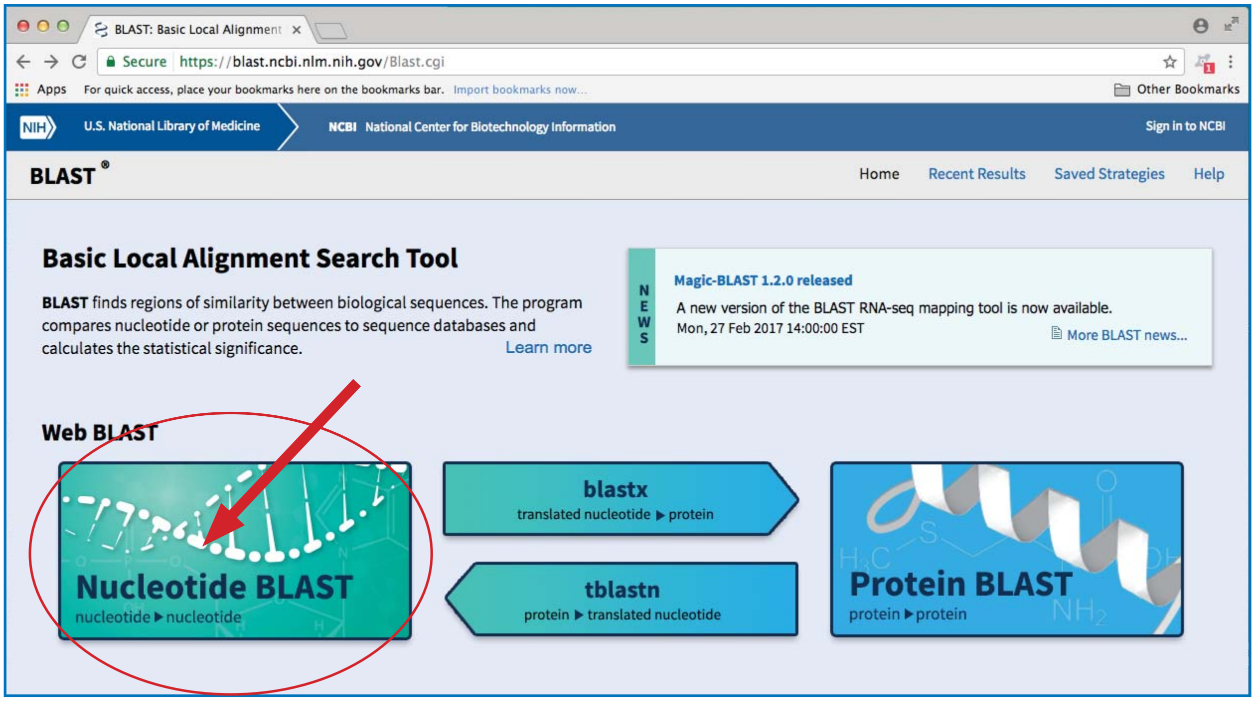
B. USING BLAST TO ANALYZE GENES AND DETERMINE THE MOST LIKELY PLACEMENT OF THE FOSSIL SPECIES

In your examination of the fossil, you discover small amounts of soft tissue. Using cutting-edge biotechnology techniques, you are able to extract small quantities of intact DNA from the tissue for sequencing. This information will be valuable to the classification of the newly discovered species because organisms with common ancestry share similar gene sequences. Animals with a higher percent of sequence identity most likely have a common ancestor. These two animals will be placed closer to one another on a cladogram. In this exercise, BLAST is used to compare the ancient DNA sequence to modern DNA sequences. This data will contribute to the final placement of the fossil species on the cladogram.

1. Four separate genes were sequenced from the ancient DNA. Download the results of the DNA sequencing study to a flash drive or a computer. Make sure you know where the files are located.

   These files are located at the following web addresses:
   Sequence 1 -- http://www.edvotek.com/site/img/AP_Biology_Lab3_Gene1.asn
   Sequence 2 -- http://www.edvotek.com/site/img/AP_Biology_Lab3_Gene2.asn
   Sequence 3 -- http://www.edvotek.com/site/img/AP_Biology_Lab3_Gene3.asn
   Sequence 4 -- http://www.edvotek.com/site/img/AP_Biology_Lab3_Gene4.asn

2. In your web browser, enter **https://blast.ncbi.nlm.nih.gov/Blast.cgi** to access the BLAST search engine. Click on the logo for "Nucleotide BLAST".


BLAST: Basic Local Alignment
Secure https://blast.ncbi.nlm.nih.gov/Blast.cgi
Apps
For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...
Other Bookmarks
NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
BLAST®
Home
Recent Results
Saved Strategies
Help
NEWS
Web BLAST
Nucleotide BLAST
nucleotide ▸ nucleotide
blastx
translated nucleotide ▸ protein
tblastn
protein ▸ translated nucleotide
Protein BLAST
protein ▸ protein

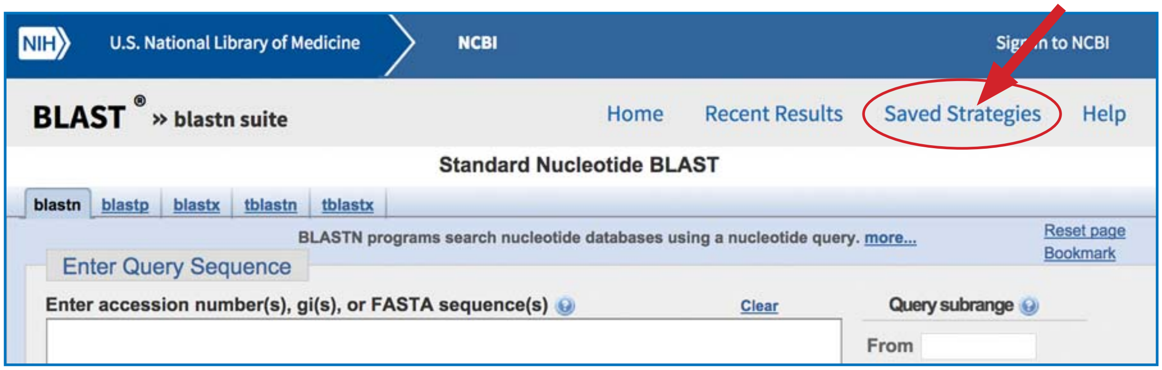
Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more

Magic-BLAST 1.2.0 released
A new version of the BLAST RNA-seq mapping tool is now available.
Mon, 27 Feb 2017 14:00:00 EST
More BLAST news...

Investigation III: Investigating Fossil Specimens Using BLAST, cont.

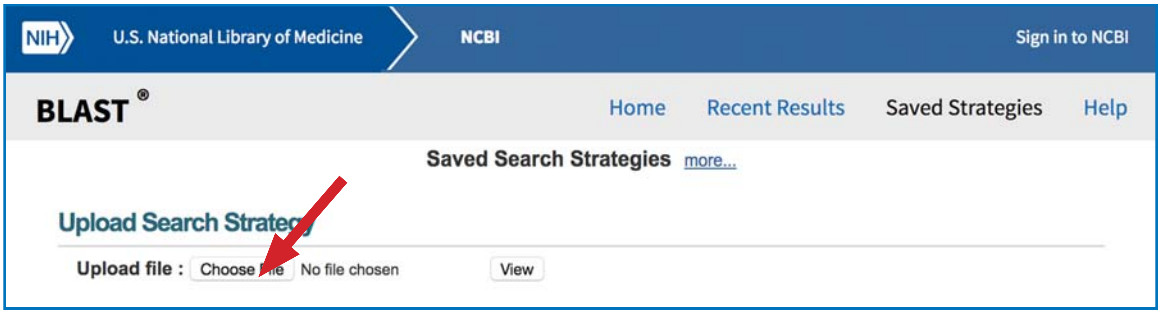
3. Use "Saved Strategies" to import your DNA sequence data into the BLAST search engine. Click on the link for "Saved Strategies" at the top right hand side of the screen.



BLASTN programs search nucleotide databases using a nucleotide query. more...

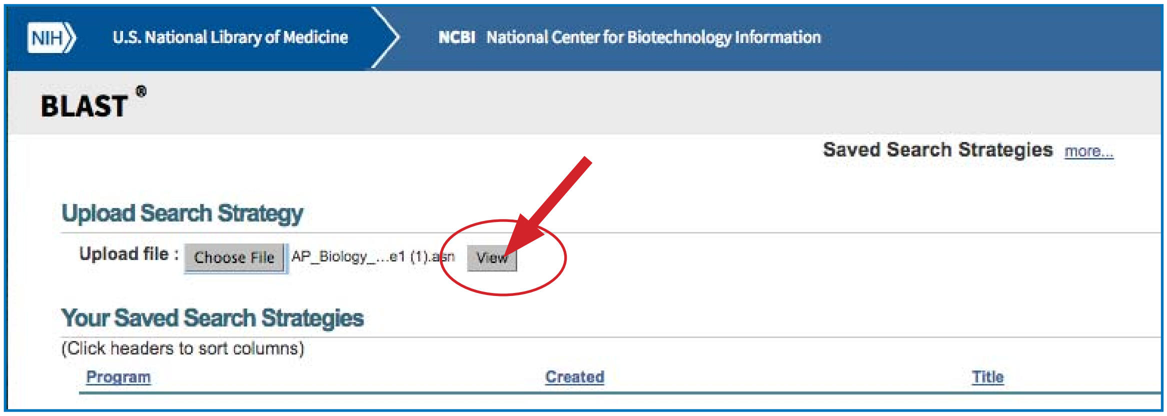
Enter accession number(s), gi(s), or FASTA sequence(s)

4. Click on the button for "Choose File" under the "Upload Search Strategy" header. Navigate to the appropriate file folder and click on one of the sequence titles to choose the file. Click the "Open" button to load the file.



Upload file : Choose File No file chosen

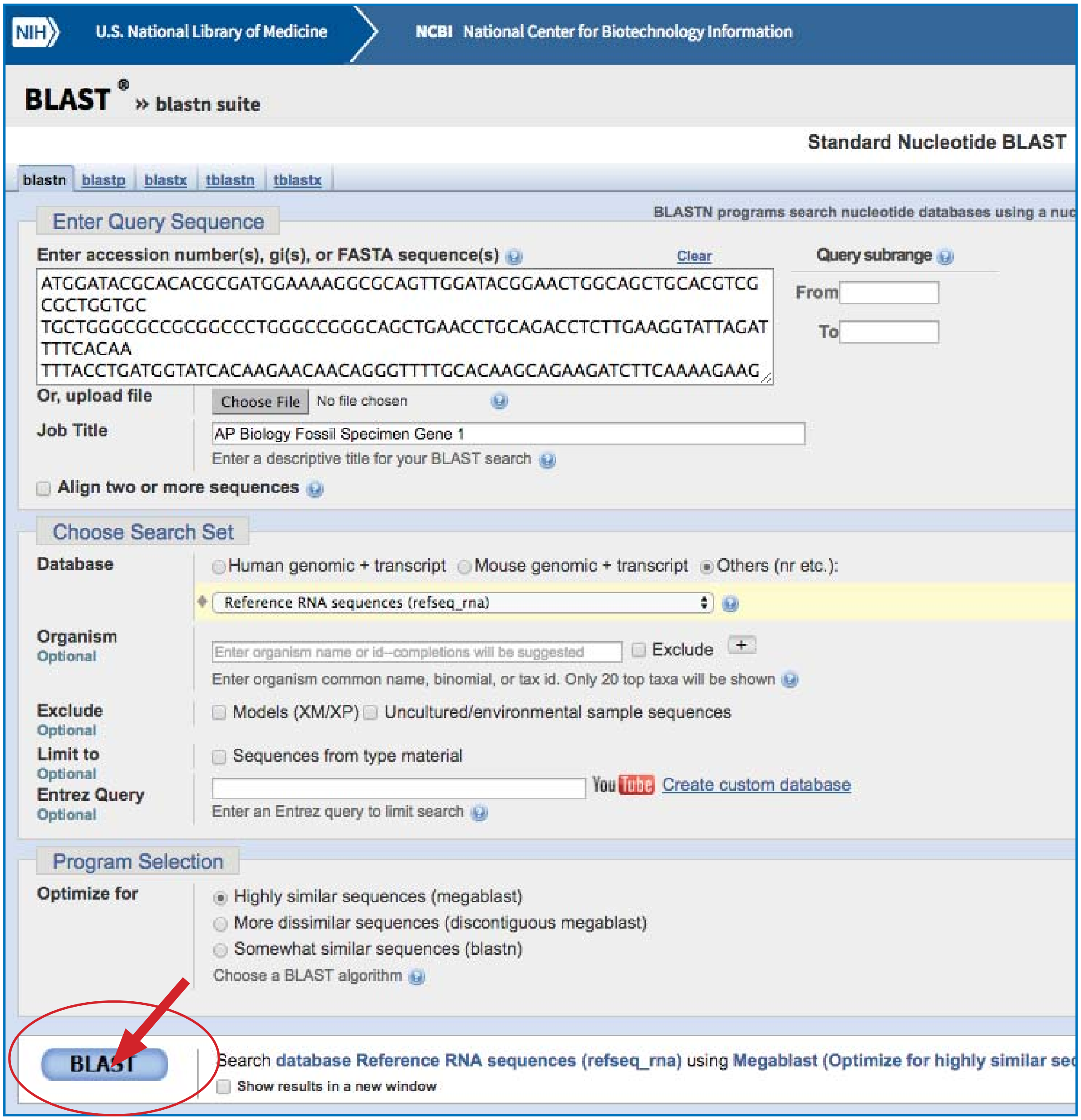
5. Once the sequence file is loaded, click "View" to open the sequence in the BLAST search engine.


NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
BLAST®
Saved Search Strategies more...
Upload Search Strategy
View
Your Saved Search Strategies
(Click headers to sort columns)
Program
Created
Title

Upload file : Choose File AP_Biology_...e1 (1).asn

Investigation III: Investigating Fossil Specimens Using BLAST, cont.

6. The sequence will load with the search parameters already set. To ensure experimental success, do not change any of these settings. Click the "BLAST" button to perform the database search.


NIH U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
BLAST ® » blastn suite
Standard Nucleotide BLAST
blastn blastp blastx tblastn tblastx
Enter Query Sequence
BLASTN programs search nucleotide databases using a nuc
Clear
Query subrange
From
To
Choose File No file chosen
Choose Search Set
Program Selection
Create custom database
BLAST

Enter accession number(s), gi(s), or FASTA sequence(s)

```
ATGGATACGCACACGCGATGGAAAAGGCGCAGTTGGATACGGAACTGGCAGCTGCACGTCG
CGCTGGTGC
TGCTGGGCGCCGCGGCCCTGGGCCGGGCAGCTGAACCTGCAGACCTCTTGAAGGTATTAGAT
TTTCACAA
TTTACCTGATGGTATCACAAGAACAACAGGGTTTTGCACAAGCAGAAGATCTTCAAAAGAAG
```

Or, upload file

Job Title: AP Biology Fossil Specimen Gene 1

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Database: ○ Human genomic + transcript ○ Mouse genomic + transcript ● Others (nr etc.):

Reference RNA sequences (refseq_rna)

Organism Optional: Enter organism name or id--completions will be suggested □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to Optional: □ Sequences from type material

Entrez Query Optional:

Enter an Entrez query to limit search

Optimize for:

● Highly similar sequences (megablast)

○ More dissimilar sequences (discontiguous megablast)

○ Somewhat similar sequences (blastn)

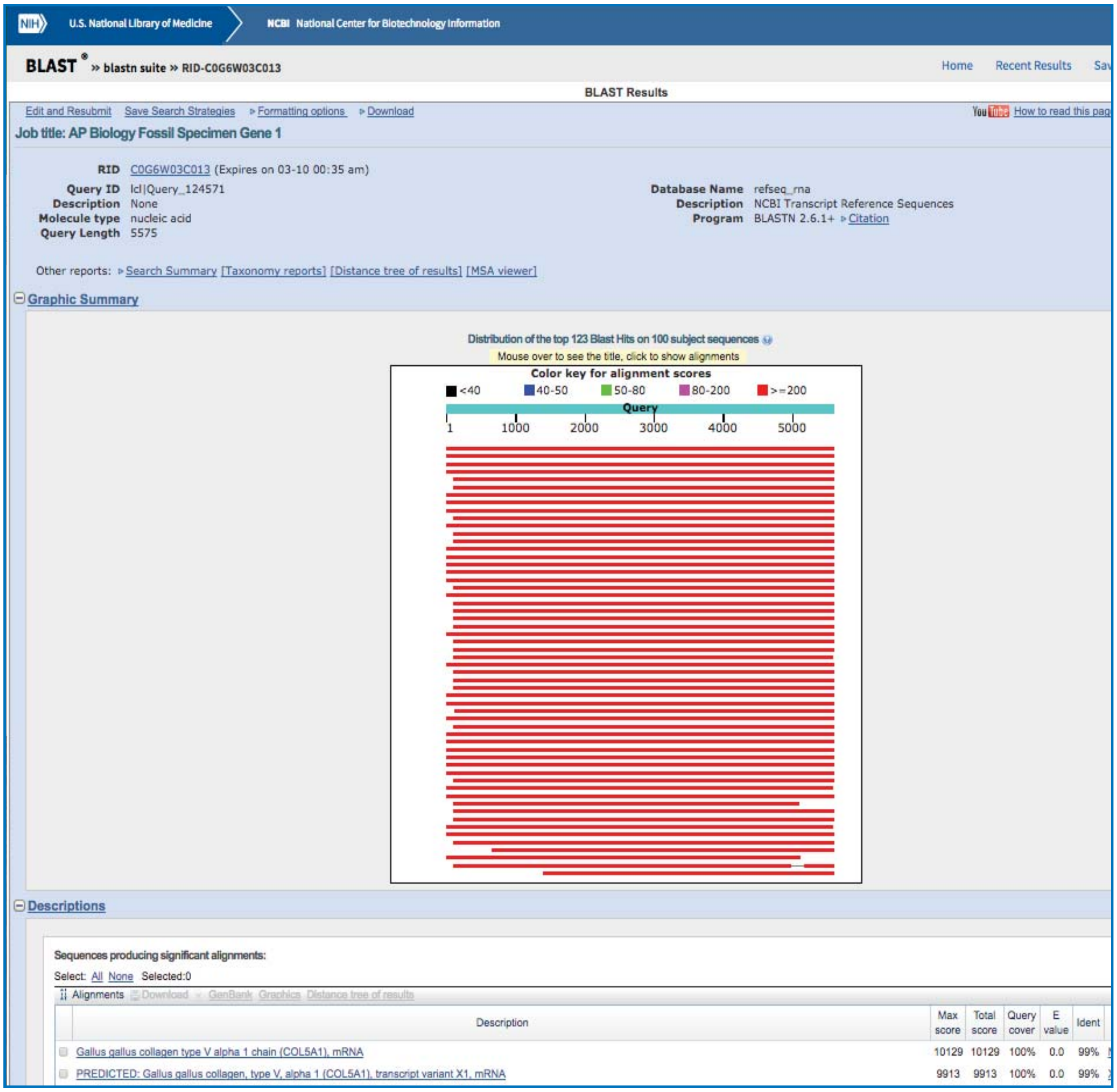
Choose a BLAST algorithm

Search database Reference RNA sequences (refseq_rna) using Megablast (Optimize for highly similar seq

□ Show results in a new window

Investigation III: Investigating Fossil Specimens Using BLAST, cont.

7. After BLAST searches through its database, it will output the data on a separate webpage. The data is arranged into four sections:
   a. Search summary report: displays the BLAST search parameters.
   b. Graphic Summary: Displays the alignment of the database matches to the query sequence. The color of the boxes corresponds to how closely the results match with the gene of interest, with red representing the highest degree of similarity. The most similar sequences are at the top of the image.
   c. Descriptions: Lists all the sequences in the database with significant sequence homology to the query sequence. The most similar sequences are at the top of the list.
   d. Alignments: Shows the alignment blocks for each BLAST hit. Data in this section includes the percent identity between the query sequence and the results. The most similar sequences are at the top of the section.


NIH U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
BLAST® » blastn suite » RID-C0G6W03C013
Home
Recent Results
Sav
Edit and Resubmit
Save Search Strategies
Formatting options
Download
YouTube How to read this pag

BLAST Results

Job title: AP Biology Fossil Specimen Gene 1

RID C0G6W03C013 (Expires on 03-10 00:35 am)
Query ID lcl|Query_124571
Description None
Molecule type nucleic acid
Query Length 5575

Database Name refseq_rna
Description NCBI Transcript Reference Sequences
Program BLASTN 2.6.1+ Citation

Other reports: Search Summary [Taxonomy reports] [Distance tree of results] [MSA viewer]

Graphic Summary

Distribution of the top 123 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

	Description	Max score	Total score	Query cover	E value	Ident
□	Gallus gallus collagen type V alpha 1 chain (COL5A1), mRNA	10129	10129	100%	0.0	99%
□	PREDICTED: Gallus gallus collagen, type V, alpha 1 (COL5A1), transcript variant X1, mRNA	9913	9913	100%	0.0	99%

AP03.170426

Investigation III: Investigating Fossil Specimens Using BLAST, cont.

8. To analyze the results, scroll down to the Description section. Be sure to record your results in your lab notebook.
   a. Examine the description for each match. Can you determine the type of protein that the ancient DNA codes for? Perform a quick search to explore the function of this protein in modern organisms.
   b. What organism's gene sequence is most similar to the fossil DNA sequence? What kinds of organisms have similar genes to the ancient DNA? To answer this you may have to select a few close matches and then use the Latin name to find out their kingdom, phylum, and class.
   c. Based on the DNA sequence, where would each animal fit into your cladogram? Is it close to your placement of the fossil species on the cladogram based on morphological features alone? How similar is the modern DNA to the fossil DNA?

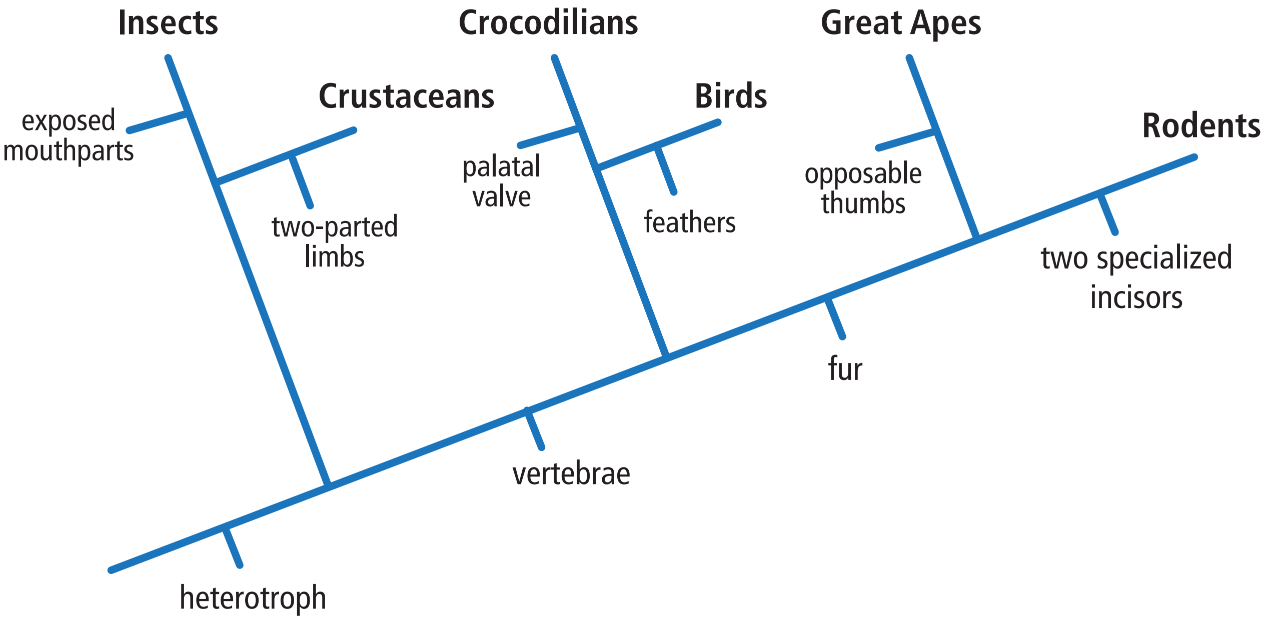
9. Create a cladogram comparing your DNA sequence to modern homologs. First, select 10-15 sequences from anywhere in the descriptions list. Next, click on the link for "Distance Tree of Results" in the Search summary report of the BLAST results page. This creates a cladogram using the results from your BLAST search. How does this help you place the fossil into the cladogram?

10. Repeat the BLAST search with the other three DNA sequences. Record your results in your lab notebook

11. Review the results from all four BLAST searches. Do they help you refine the placement of your fossil organism on the cladogram? Why or why not?

12. Using the information from your BLAST search, name your new species.

13. Using the information from your BLAST searches, mark the below cladogram with the most likely placement for the fossil species. Have you placed the new species in the same category as you did previously? Why or why not?



Investigation IV: Explore Interesting Genes with BLAST

Now that your are familiar with data analysis using BLAST, the next step is to choose and explore a gene that interests you. To locate a gene, go to the Entrez Gene website (www.ncbi.nlm.nih.gov/gene) and search for the gene. Once you have found the gene on the site, copy the gene sequence and input it into a BLAST query.

For practice, follow the instructions below to BLAST the human beta hemoglobin gene.

1. In your web browser, enter www.ncbi.nlm.nih.gov/gene to access the Entrez Gene search engine.
2. In the search bar at the top of the page, enter "human beta hemoglobin" and hit "Search."
3. Click on the first link in the results list. This brings up the full gene report.
4. In the side bar, under the "Table of Contents" header, click on "NCBI Reference Sequences (RefSeq)." This brings you to the genomic and mRNA nucleotide sequences.
5. The first accession number listed under "mRNA and Protein(s)" should read "NM_000518.4" or something similar. Click on the link to access detailed information about the mRNA sequence from GenBank.
6. To access the mRNA sequence in a simple text-based format, click on the link for "FASTA" just below the gene title. Note: The Nucleotide page also contains relevant citations and other details about the gene. Feel free to explore the content if time permits.
7. In the side bar on the FASTA sequence page, under the "Analyze this sequence" header, click on "Run BLAST". This copies the accession number of your mRNA sequence and enters it into the BLAST search box. (Alternatively, copy the gene sequence and go to the BLAST search page. Paste the sequence into the search box.)
8. If you plan on accessing the data at a later date, be sure to give your search a name in the "Job Title" box.
9. Select the database to search using the "Database" option under "Choose Search Set." This allows you to limit your search to the human genome or the mouse genome. For this exercise, consider using default setting of "Others (nr etc.)" to search through all available data.
10. Optimize your search conditions using the "Program Selection" dialog box. Choosing "Highly similar sequences (megablast)" will be very fast, but it will only compare your sequence of interest to closely related sequences. "More dissimilar sequences (discontiguous megablast)" allows for sequence mismatches in the matches, which may increase the number of matches. The final option, "Somewhat similar sequences (blastn)" is the slowest option because it uses the most general search terms, but it will display the most matches.
11. Click "BLAST" to perform your search. This may take a little while depending on the search parameters.
12. Analyze your results. The matching sequences represent the beta hemoglobin gene from different species.

Using BLAST, explore a gene that interests you. This could be something discussed in another lesson (i.e. the enzyme Catalase) or in the news (BRCA-1). Below are some questions to direct your research.

1. Describe the function of your gene in humans. Does it perform the same function in other organisms?
2. Is your gene found in many species or only a few?
3. You may see some differences in the DNA sequences, even between closely related species. Would any of these changes affect the function of the protein? Why or why not?