FROM FOSSILS TO PHYLOGENIES PART 2: USING BLAST TO IDENTIFY PROTEINS THAT ARE EVOLUTIONARILY RELATED ACROSS SPECIES

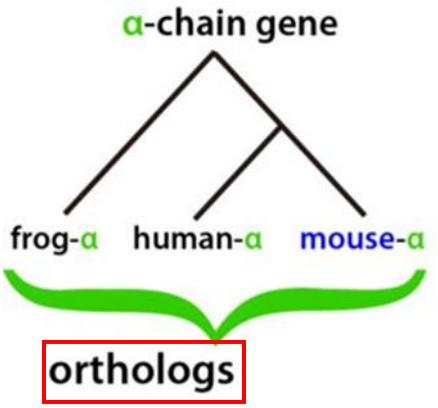
HOW CAN BIOINFORMATICS BE USED AS A TOOL TO DETERMINE EVOLUTIONARY RELATIONSHPS AND TO BETTER UNDERSTAND PROTEIN HERITAGE?

[Adapted by Dane Besser, Baylee Goodwin, and Stephen Ramsey from a CollegeBoard Investigation]

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 genes that define a human being. As you learned in Activity 2, a gene's DNA sequence is the template that dictates – according to a three-letter code – the sequence of amino acids out of which a specific protein is made. Amino acids have their own code as well, seeing as there are 20 amino acids and 64 codes.[1] Protein-coding genes are an important class of molecular "building blocks" for the human body. In addition to human genes, scientists have also sequenced the genes of hundreds of other species across the tree of life. These gene sequences are freely available for anyone in the world–*including you*–to access via a web browser and examine.

How are gene sequences useful for science? First, mapping DNA sequences to locate specific genes allows scientists to align the genes across species (for example, a pair of human and mouse genes). These genes might be "similar" in that they evolved from the same common ancestral gene. We call the two genes in such a pair orthologs. Figure 1 is an example of orthologs.

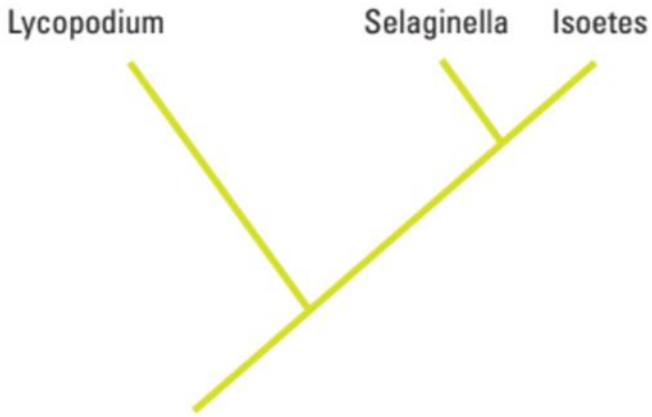


[1] See "The 20 Amino Acids and Their Role in Protein Structure" for the full list of amino acid codes.

Often, genes that are orthologs will have similar functions in their respective species, so scientists can learn about the function of a human gene by studying that gene's ortholog in another species, such as in a fruit fly or a mouse. Second, comparing related genes among two or more species can provide insight into the species' evolutionary relationships, more than comparing the species' physical appearance or characteristics. Finally, knowing the sequences and locations of human genes helps enable scientists to investigate how variation in a gene's sequence across humans leads to variation in human traits: eye color, hair color, height, or risk of various health conditions.

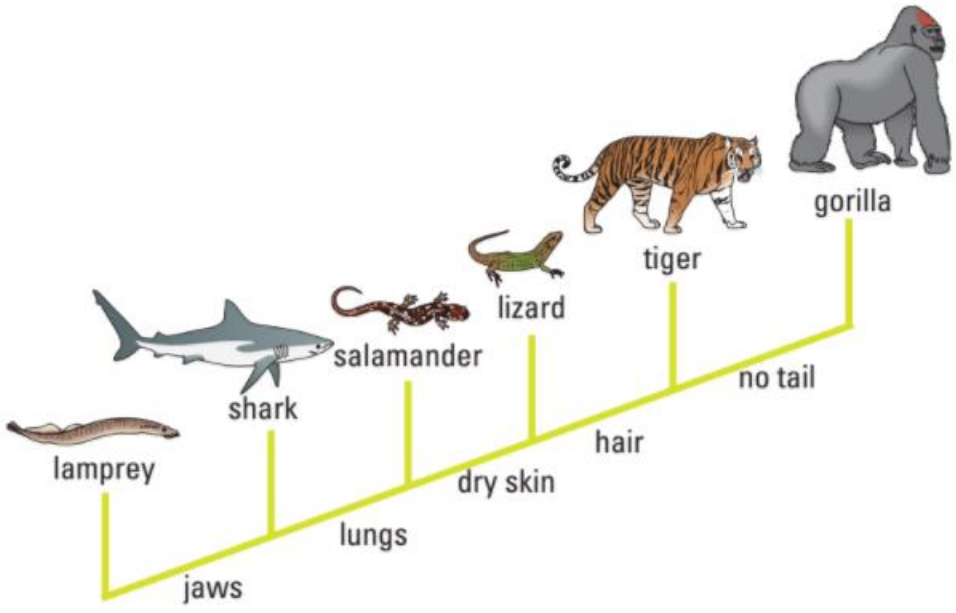
Suppose you are a scientist who has identified a fruit fly gene that, when the gene is disrupted in fruit fly embryos, results in a developmental abnormality. You would likely want to know: does that gene have an ortholog in humans? If so, do mutations in that human gene cause disease, and how do the fruit fly gene and human gene differ in terms of DNA or amino acid sequence? In theory, you could answer these questions by comparing paper printouts of the sequences of each of the ~20,000 human protein-coding genes to a printout of the sequence of your fruit fly gene, in order to find a potential ortholog that would have a close sequence match. But this process, if carried out by hand, would take many years. Fortunately, computers can carry out the same search in seconds or minutes. The software program that would be used for such a search is but one example of a broad class of bioinformatics software tools and computational methods. More precisely, bioinformatics is a field of science that blends biology, computer science, statistics, and mathematics in the systematic analysis of biological data and information. Using bioinformatics tools, entire genomes can be quickly compared to detect genetic similarities and differences. An extremely powerful and versatile bioinformatics tool is the Basic Local Alignment Search Tool (BLAST). Using BLAST, you can input a DNA or amino acid sequence and search entire genomic libraries for identical or similar known sequences.

In this activity, you will use BLAST to analyze amino acid sequences from several extinct species, determine what proteins they come from, and find the proteins' orthologs in modern-day animal species. You will then use the information from your BLAST analysis to create a phylogenetic tree. A phylogenetic tree is a diagram that depicts the evolutionary relatedness of species or groups of closely related species. Figure 2 is a simple phylogenetic tree.



Note that the phylogenetic tree is shaped like a tree, with the endpoints of each branch representing a specific group of organisms. The closer the two groups are located to each other, the more recently they shared 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 3 includes additional details, such as the evolution of particular physical structures called derived characteristics. Note that the placement of the derived characteristics 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 are not hairy.



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

Historically, physical characteristics were used for deciphering the evolutionary relationships among species; however, today scientists rely heavily on gene sequence information as well. Chimpanzees and humans share 95%+ of their DNA, which would place them closer together on a phylogenetic tree. Humans and fruit flies share approximately 60% of their genes, which would place them farther apart on a phylogenetic tree.

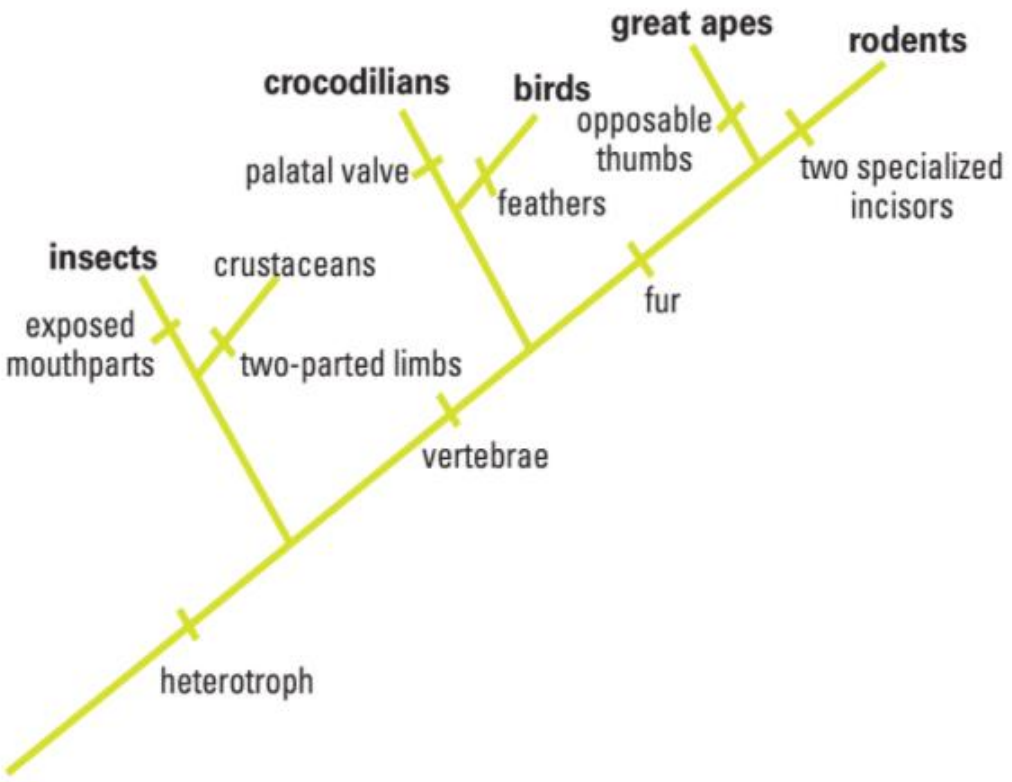
Can you draw a phylogenetic tree that depicts the evolutionary relationship among humans, chimpanzees, fruit flies, and mosses?

Learning Objectives

- Understand how phylogenetic trees depict evolutionary relationships
- Understand how the bioinformatics tool BLAST enables the identification of evolutionarily related proteins in different species (orthologs)
- Be able to critically analyze the results form a BLAST analysis to assess consistency with the current phylogenetic tree for various animal species groups.

Procedure

You are a member of a scientific team that has discovered three unusually well-preserved fossilized bone specimens from an extinct mastodon species (*Mammut americanum*) and two dinosaur species: *Tyrannosaurus rex* and the hadrosaur *Brachylophosaurus canadensis.* Upon careful examination of the fossil, small amounts of *soft tissue* have been discovered, which is unusual because normally soft tissue does not survive over this time-scale. From the soft tissue in the bone specimen, your team was able to extract amino acid sequences of several protein fragments–*the first time an actual dinosaur protein fragment has ever been sequenced!* Your task is to use BLAST to compare these amino acid sequences to protein sequences from other species. Then, use the results from the BLAST analysis to determine where these extinct species branch off from the evolutionary tree (Figure 4) in relation to modern animals like birds, crocodiles, and mammals.



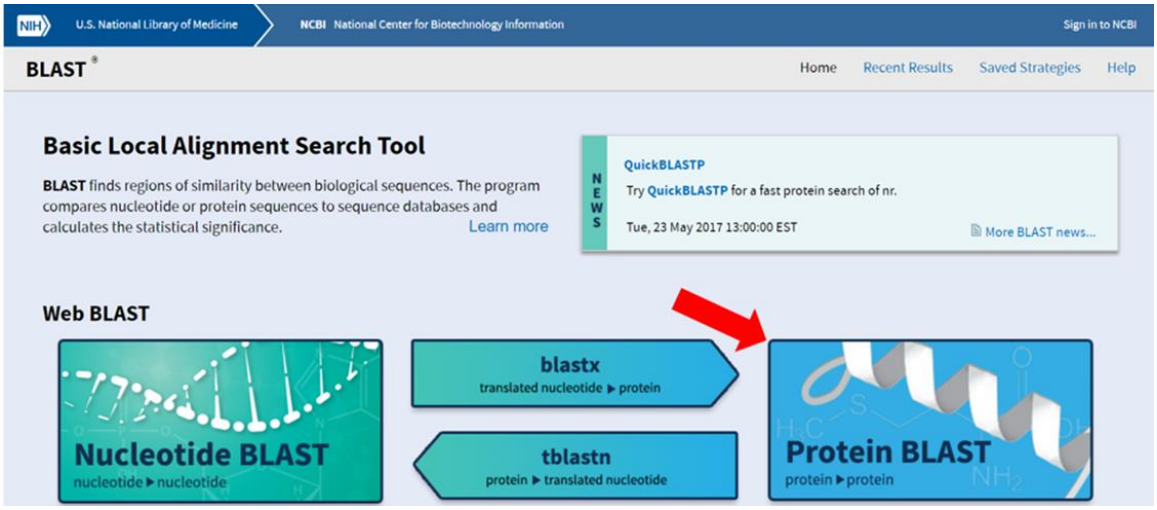
I) Step 1: Form an initial hypothesis about where the mastodon and the two dinosaur species belong on the phylogenetic tree (Figure 4) based on what you know about the physical characteristics of mastodons and dinosaurs. Mark the locations as "branches" from the tree on Figure 4.

II) Step 2: Locate the protein fragment files for the *T. rex* bone specimen at the end of this document.

III) Step 3: Copy the gene sequence into BLAST by doing the following:

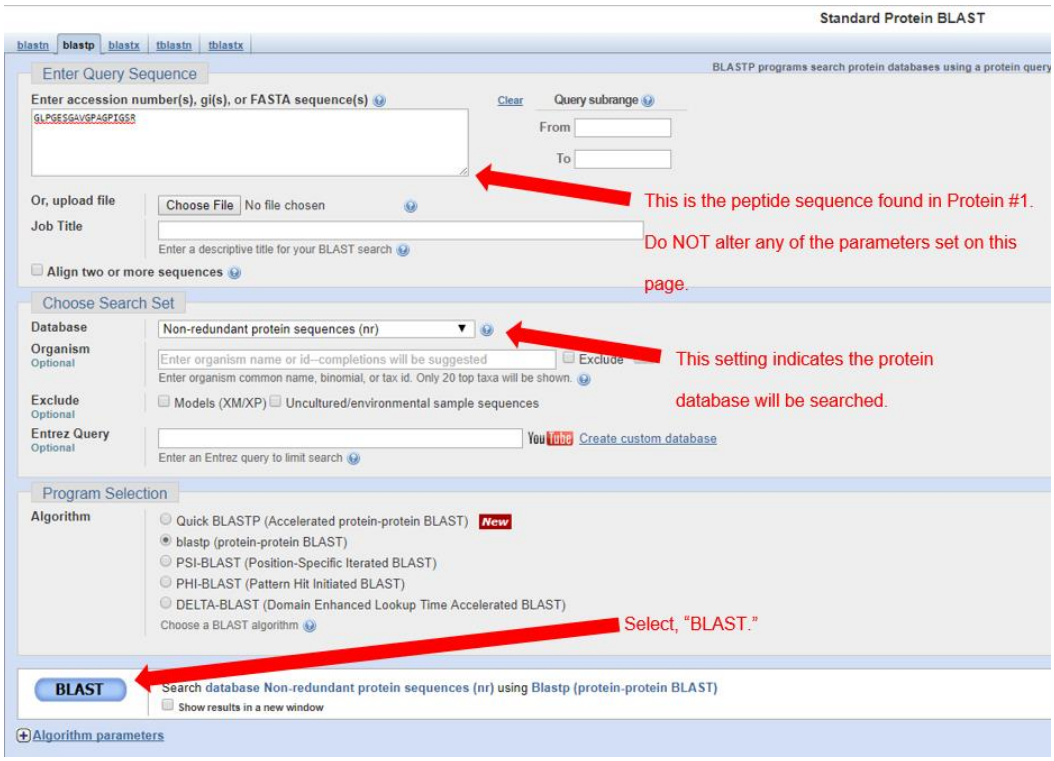
a) Use your web browser to access the BLAST homepage: https://blast.ncbi.nlm.nih.gov/Blast.cgi

b) Click on “Protein BLAST” from the menu at the bottom of the page



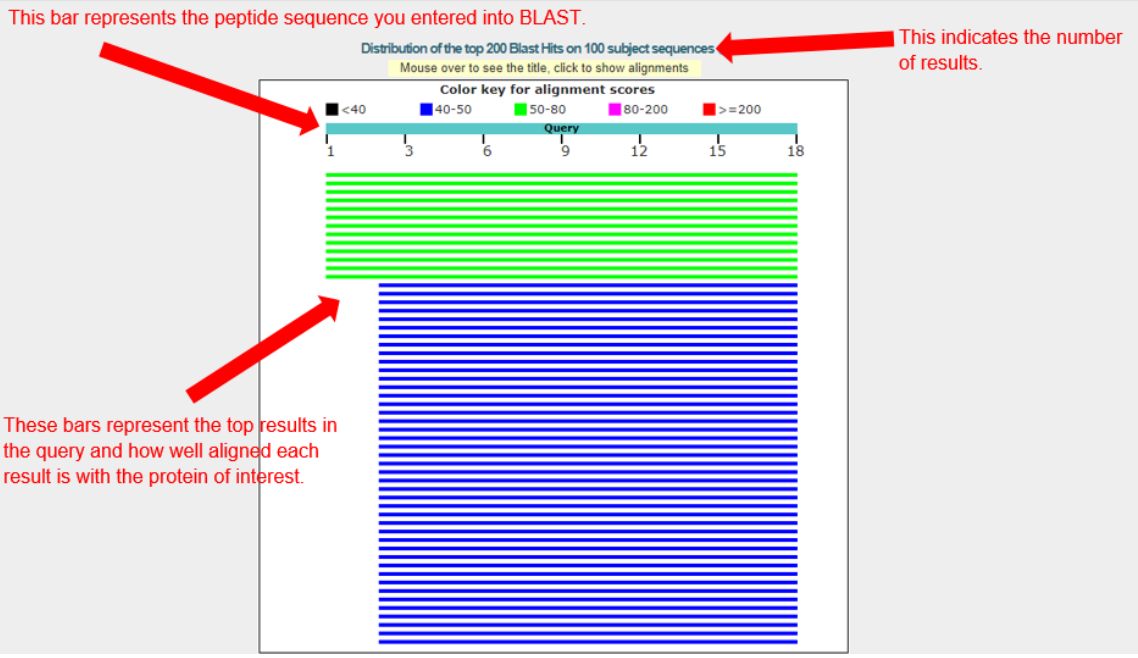
c) Under “Enter Query Sequence,” paste the first amino acid sequence from “t-rex.”

d) 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 left.

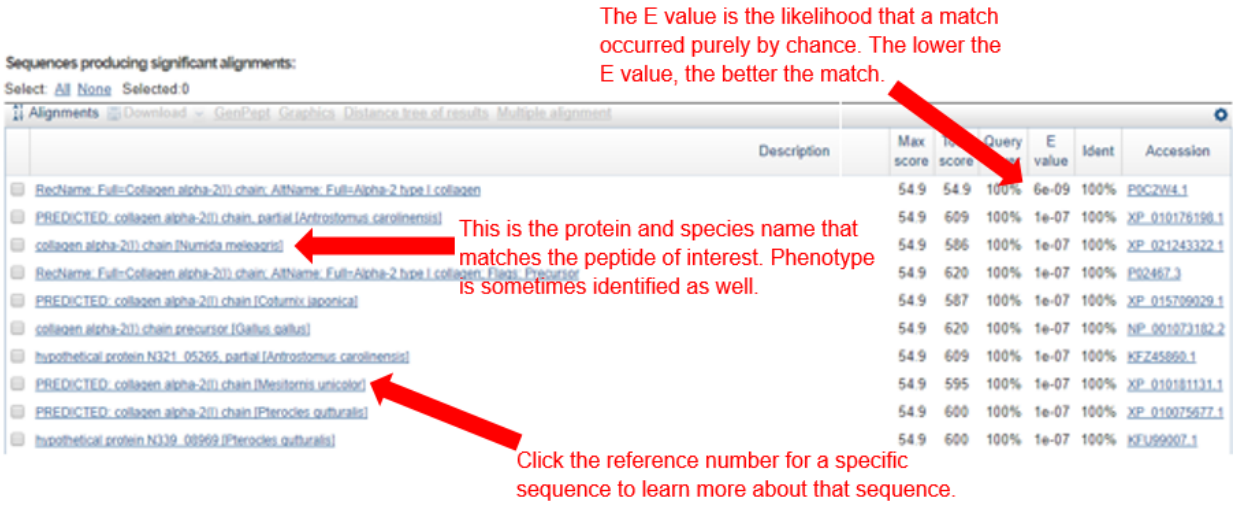


e) After collecting and analyzing all the data for that particular amino acid sequence (see instructions below), repeat this procedure for the other two amino acid sequences (mastodon and hadrosaur).

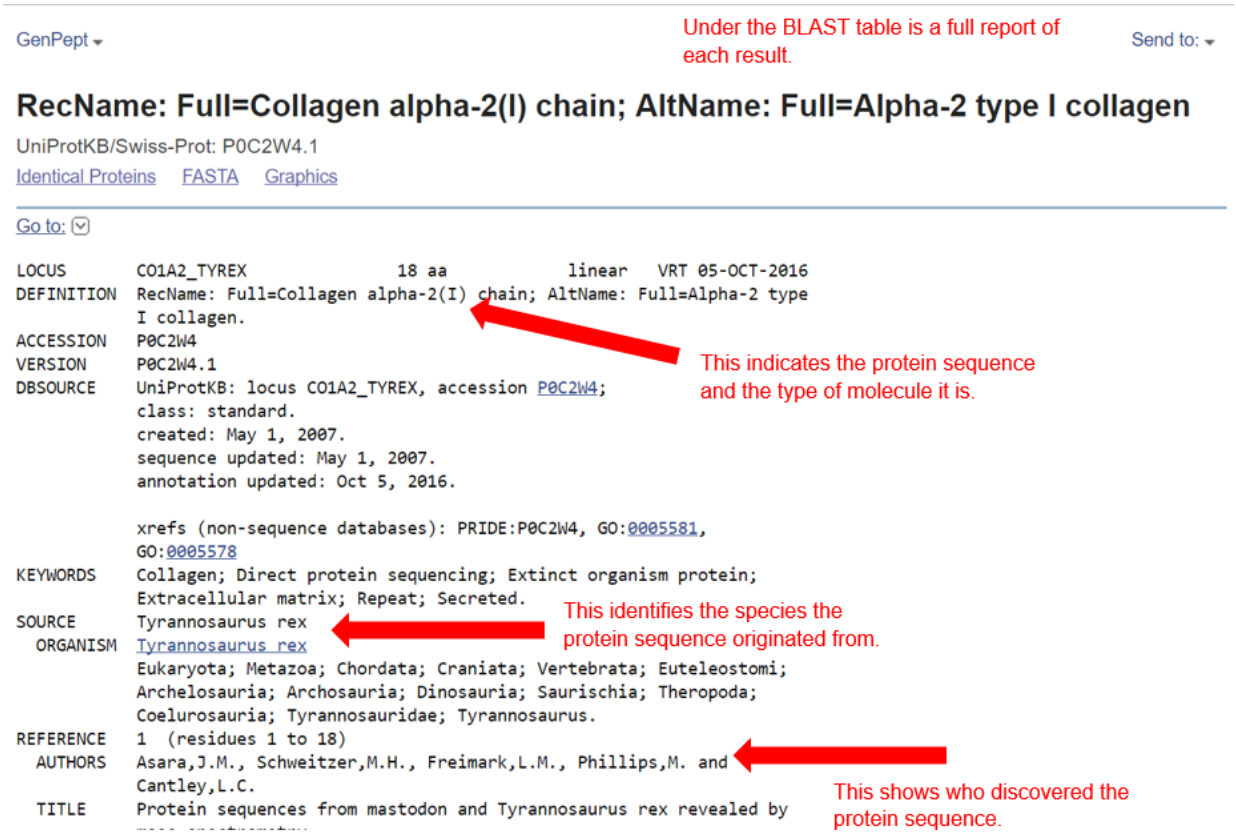
IV) Step 4: The results page has two sections. The first section is a graphical display of the matching sequences.



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 protein of interest. The most similar sequences are listed first, and as you move down the list, the sequences become less similar to your protein of interest. Each matching protein sequence is annotated with a description on the left. Based on scanning the descriptions in the table, what type of protein did your amino acid sequence come from? Do a Wikipedia search for this protein name. Does it make sense that this type of protein would be found in a bone sample?



If you click on a particular species listed, you’ll get a full report that includes the classification of the species, the research journal in which the protein was first reported, and the sequences of bases that appear to align with your protein 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 phylogenetic tree.

As you collect information from BLAST for each of the protein 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 phylogenetic tree.

For each BLAST query, consider the following:

- The higher the alignment score, the closer the alignment (the more similar the fossil protein and its matching protein from the database).
- The lower the *E* value, the less likely the alignment score this high occurred "by chance".
- Sequences with *E* values less than 10^{-4} (depicted as 1e-04 in the BLAST results table) can be considered highly likely to be evolutionarily related, i.e., orthologs.

1. What is the likely protein that your fossil-derived amino acid sequence came from?
2. What species in the BLAST result has the most similar amino acid sequence to your fossil-derived amino acid sequence?
3. Where is that species located on the Figure 4 phylogenetic tree?

4. How similar is that amino acid sequence to your fossil-derived amino acid sequence?
5. What species has the next most similar amino acid sequence to your fossil-derived amino acid sequence?

Based on what you have learned from the sequence analysis and what you know from the structure, decide where the fossil specimens (*M. americanum*, *T. rex*, or *B. canadensis*) belong on the phylogenetic tree for modern-day animals. If necessary, redraw the phylogenetic tree you created before.

Evaluating Results

Compare and discuss your phylogenetic tree with your classmates. Does everyone agree with the placement of the fossil specimens? If not, for which species is there disagreement?

On the main page of BLAST, under “Specialized searches,” click on the link “SmartBLAST.” What phylogenetic trees do you see when you put in different collagen sequences for the BLAST search? How does the lack of other sequenced species impact the proper analysis of the protein data used in this lab?

What other data could be collected from the fossil specimens to more convincingly determine their species' locations in the evolutionary tree?

>t-rex

GLPGESGAVGPAGPIGSR

>hadrosaur

GSNGEPGSAGPPGPAGLRGLPGESGAVGPAGPPGSR

>mastodon

QYDAKGVGLGPGPMGLMGPRGPPGATGPPGSPGFQGPPGEPGEPGQTGPAGSRGPAGPPGKAG
EDGHGKPGRPGERGVVGPQGARGFPGTPGLPGFKGIRGHNGLDGLKGQPGAPGVKGEPGAPGEN
GTPGQIGARGLPGERGRVGGPGPAGARGSDGSVGPVGPAGPIGSAGPPGFPGAPGPKGEIGPVGN
PGPSGPAGPRGEAGLPGVSGPVGPPGNPGANGLAGAKGAAGLPGVAGAPGLPGPRGIPGPVGAA
GATGARGIVGEPGPAGSKGESGSKGEPGSAGPQGPPGPSGEEGKRGPNGEAGSAGPAGPPGLRG
GPGSRGLPGADGRAGVMGPPGSRGASGPAGVRGPSGDSGRPGEPGVMGPRGLPGSPGNVGPAG
KEGPAGLPGIDGRPGPIGPAGARGEPGNIGFPGPKGPAGDPGKNGDKGHAGLAGPRGAPGPDGNN
GAQGPPGLQGVQGGKGEQGPAGPPGFQGLPGPSGTAGEAGKPGERGIPGEFGLPGPAGPRGERG
PPGQSGAAGPTGPIGSRGPSGPPGPDGNKGEPGVVGAPGTAGPSGPVGLPGERGAAGIPGGKGEK
GETGLRGDTGNTGRDGARGAPGAVGAPGPAGATGDRGEAGPAGSAGPAGPRGSPGERGEVGPA
GPNGFAGPAGAAGQAGAKGERGTKGPKGENGPVGPTGPVGAAGPAGPNGPPGPAGSRGDGGPP
GATGFPGAAGRTGPPGPAGITGPPGPPGAAGKEGLRGPRGDQGPVGRTGETGASGPPGFAGEKG
SSGEPGTAGPPGAPGPQGILGPPGILGLPGSRGERGLPGVAGAVGEPGPLGIAGPPGARGPPGAVG
SPGVNGAPGEAGRDGNPGSDGPPGRDGLPGHKGERGYPGNAGPVGTAGAPGPQGPLGPAGKHG
NRGEPGPAGSVGPVGAVGPRGPSGPQGARGDKGEAGDKGPRGLPGFKGHNGLQGLPGLAGQHG
DQGSPGSVGPAGPRGPAGPSGPVGKDGRPGHAGAVGPAGVRGSQGSQGPSGPPGPPGPPGPPG
PSGGGYDFGYDGDFYRA