Using BLAST and MAB Phylogeny Analysis to study how dinosaurs fit into the evolutionary tree of life

Background

During Sessions 1 and 2 you were given amino acid sequences that had been obtained from a fossilized bone specimen from a *Tyrannosaurus rex* (as well as sequences from a Hadrosaur and a Mastodon). You also learned how to input the *T. rex* amino acid sequence into BLAST to identify what present-day animals are most closely related to *T. rex*. In this activity you will learn how to use a computer to analyze related amino acid sequences from a variety of animals to gain insight on their evolutionary relationships.

Mutations in DNA are the driving force for evolution. When mutations arise in the DNA sequence of an organism, they can result in changes to the translated amino acid sequence of a protein. For example, the original DNA sequence in a small portion of a gene might have read ATAAGT, but after the mutation it reads ATAACT (i.e., a G was replaced with a C). This changes the amino acid in the sequence from Leucine to a stop codon (signaling the end of the protein), which results in the cell making a shortened protein whose function may substantially differ from the original full-length protein. When a mutation is present in an organism's cell, it can be passed on from the organism to its offspring, which is how animals evolve on a molecular scale. The genetic differences between two species, such as a bird and a lizard, represent the accumulation of billions of mutations over many millions of years. The differences in the DNA (or, as we will study today, protein) sequences among a set of representative species can be used to determine how the species are related, which we will visualize as a **phylogenetic tree.** As we will discover, the more closely related organisms will have more similar protein sequences, and the more distantly related organisms will have more dissimilar protein sequences.

The first step of this activity is to obtain protein sequence data from a set of animal species that we want to compare. We will be searching for the "alpha-2 type 1 collagen" protein sequence since that is the what scientists were able to extract from the fossilized femur bone of the *T. rex*. Collagen is evolutionarily rather well-conserved across species, which is why it is a good choice for using amino acid sequences to build a phylogenetic tree. When a protein is “well-conserved” it means that the protein is found in multiple species that are distantly related, collagen is a well-conserved protein found in all animals with true bone. In order to find the collagen sequence, you will conduct a search in an online database called GenBank. The alpha-2 type 1 collagen protein sequences have already been collected for you for most of the animals, however you still need to collect the appropriate amino acid sequence for the *T. rex*.

The animal species for which you will be building a phylogenetic tree are chicken, rainbow trout, human, dog, cattle, toxodon (*Toxodon platensis*), mastodon (*Mammut americanus*), salamander, frog, and *T. rex*. (side note: these species were selected for this activity because they have alpha-2 type 1 collagen protein sequences available in the database). These animals will allow you to analyze where the *T. rex* fits in the tree of relative to birds, mammals, amphibians, and fish (no reptile data were available), which are four out

of the five major vertebrate taxonomic groups. It also allows you to see how other extinct animals like mastodon and toxodon relate to present-day animals.

Learning Objectives

- Understand how amino acid sequences can be compared using a computer program in order to reconstruct a phylogenetic tree
- Learn how to perform a search in the GenBank database
- Learn how to obtain protein or peptide sequence data in the correct formatting
- Learn how to use the web-based software tool MAB (Methods and Algorithms for Bioinformatics) Phylogeny Analysis to generate phylogenetic trees from a set of species' related amino acid sequences
- Understand how to interpret a phylogenetic tree

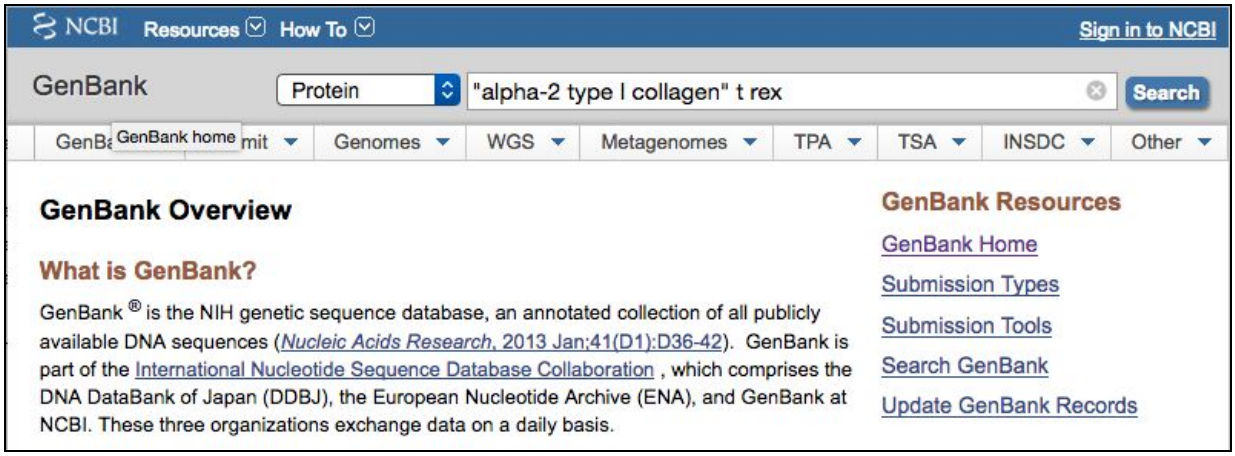
Procedure

Obtain *T. rex* amino acid sequence data from the "alpha-2 type 1 collagen" protein:

1. **Go to the folder/club website** and open up the document entitled alpha 2 collagen sequences.
2. Then navigate to the following link: http://www.ncbi.nlm.nih.gov

(NCBI stands for the National Center for Biotechnology Information, which is a branch of the National Library of Medicine that hosts the GenBank database). The NCBI website is free for the public to access, and it contains libraries of genomic, genetic, and biomedical data. We will be using it to access protein sequences in GenBank. GenBank contains the sequences of many genes and their protein products, for hundreds of thousands of different species.

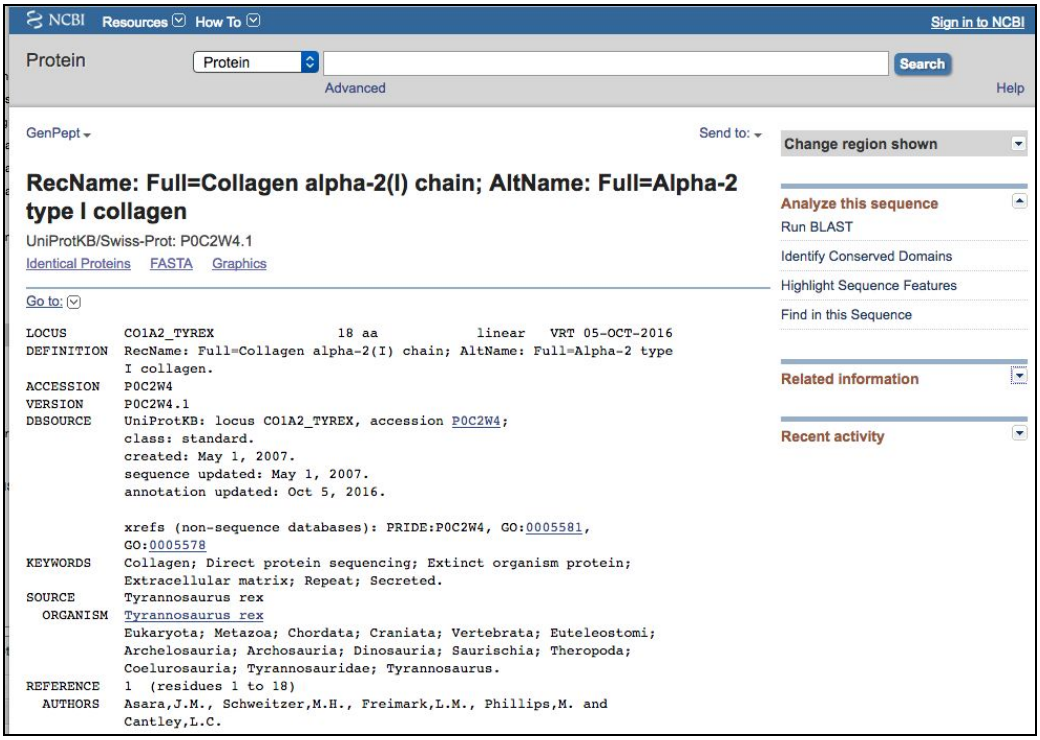
3. In the search bar at the top of the web page type in “collagen type I alpha 2 T rex“ or “α2t1 collagen T rex” and select the protein database from the drop-down menu.



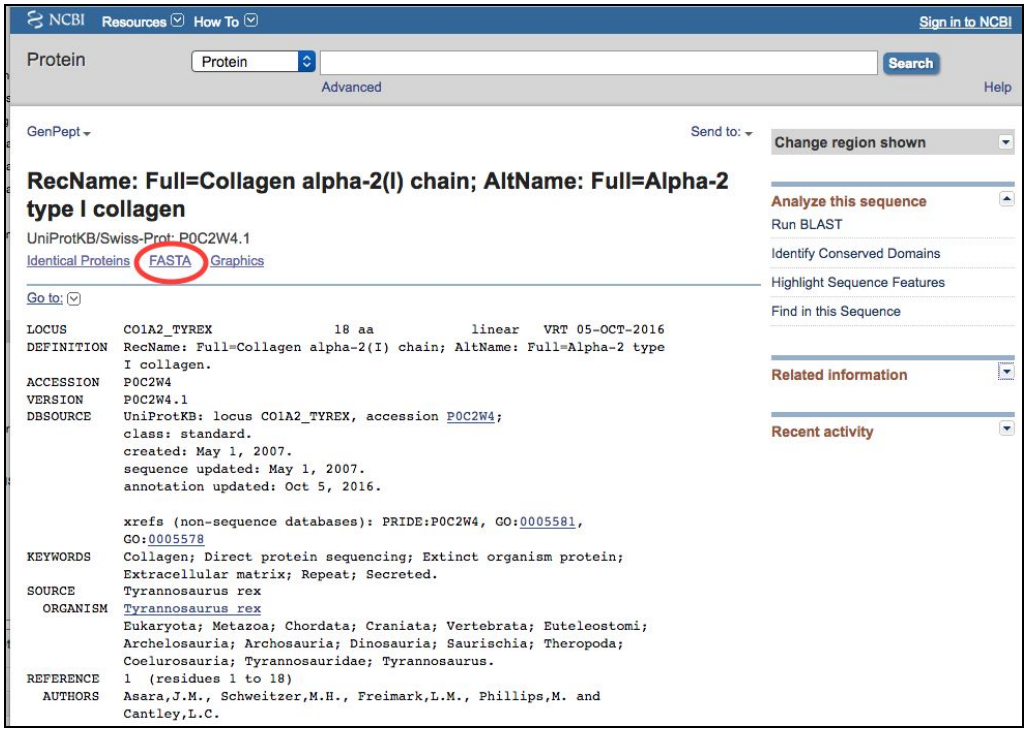
4. Click on the blue Search button.
5. It will provide a list with the top results relevant to your search. It should load 3 to 4 items, be sure to click on the result that says alpha-**2**(I) chain, **not** alpha-1(I).

The results display every known protein sequence that matches with the key words alpha 2, type 1, and *T. rex*. It will show results of things that do not precisely match your search, so be sure to fully read the names of the results. If you were to broaden your search to “alpha 2 collagen” it will result in hundreds of matches, rather than only three or four.

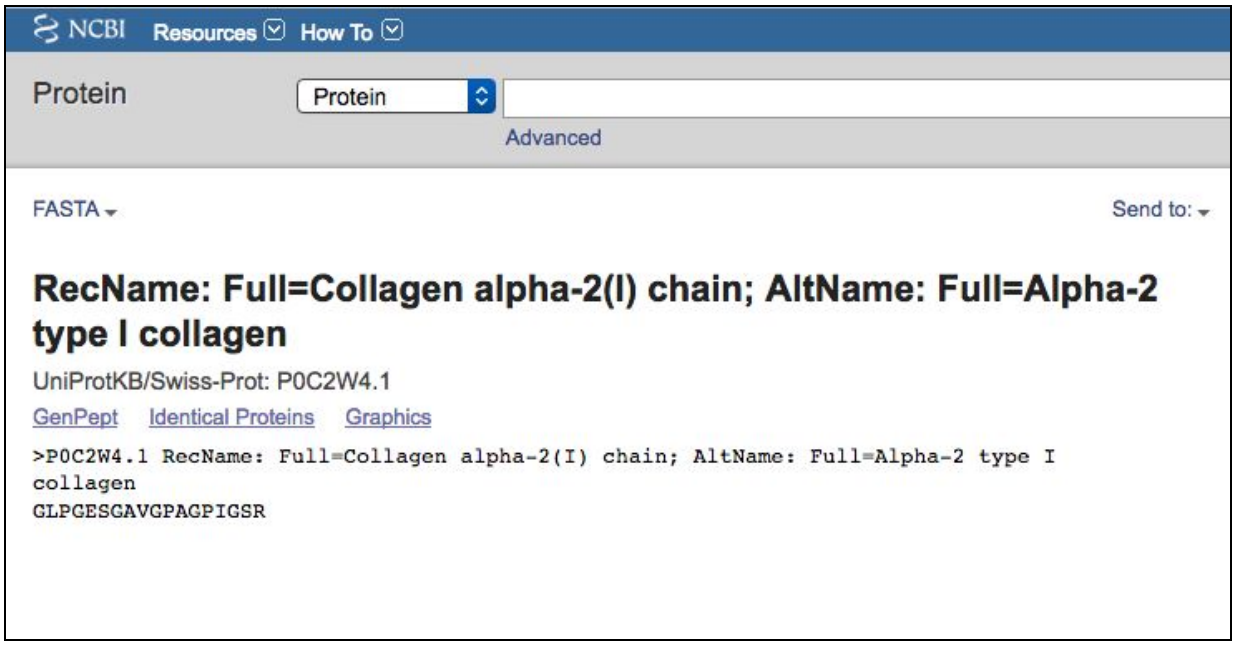
6. Once you select the correct result, it will open up this page (pictured below). In order to make sure that you have selected the correct result, look at the column on the left hand side of the page. The fourth item down should say “source organism”, and the organism should be *Tyrannosaurus rex*. If it does not say it, hit the back button and retype the search query exactly as shown in Step 3.



7. Once you reach the correct Protein record page, click on the FASTA button underneath the protein’s name in black bold writing.



GenBank should then display a FASTA record page, like this:



"FASTA" (an abbreviation for "Fast-All") is the simple text-based file format that is often used to transmit DNA or amino acid sequences from one computer program to another. In a FASTA file, the DNA nucleotides or protein amino acids are represented by individual letter codes. The FASTA file format begins with a ">" (greater than) character followed by a description, which is then followed by lines of sequence data.

8. On FASTA record page, select and copy all of the text from the “>” all the way to the end of the amino sequence.

9. Paste the whole FASTA formatted sequence at the top of the Word document that you opened in Step 1. Make sure that there is a line break before the "GLPGESGAVGPAGPIGSR":

```
>P0C2W4.1 RecName: Full=Collagen alpha-2(I) chain; AltName: Full=Alpha-2 type
I collagen
GLPGESGAVGPAGPIGSR

>Human
MLSFVDTRTLLLLAVTLCLATCQSLQEETVRKGPAGDRGPRGERGPPGPPGRDGEDGPTGPPGPPGPPG
PPGLGGNFAAQYDGKGVGLGPGPMGLMGPRGPPGAAGAPGPQGFQGPAGEPGEPGQTGPAGARGPA
GPPGKAGEDGHPGKPGRPGERGVVGPQGARGFPGTPGLPGFKGIRGHNGLDGLKGQPGAPGVKGEPG
APGENGTPGQTGARGLPGERGRVGAPGPAGARGSDGSVGPVGPAGPIGSAGPPGFPGAPGPKGEIGAV
GNAGPAGPAGPRGEVGLPGLSGPVGPPGNPGANGLTGAKGAAGLPGVAGAPGLPGPRGIPGPVGAAGA
TGARGLVGEPGPAGSKGESGNKGEPGSAGPQGPPGPSGEEGKRGPNGEAGSAGPPGPPGLRGSPGSR
GLPGADGRAGVMGPPGSRGASGPAGVRGPNGDAGRPGEPGLMGPRGLPGSPGNIGPAGKEGPVGLPG
```

10. Now you need to change the description to say "T-rex". In the Word document, delete "P0C2W4.1 RecName: Full=Collagen alpha-2(I) chain; AltName: Full=Alpha-2 type I collagen" and replace it with "T-rex". (Be sure to leave the > otherwise it will not recognize the format. This step is important in order to read and make sense of your phylogenetic tree. Now instead of the tree reading the full protein name, it wil read the name of the animal.) The Word document should now look like this:

```
>t-rex
GLPGESGAVGPAGPIGSR

>Human
MLSFVDTRTLLLLAVTLCLATCQSLQEETVRKGPAGDRGPRGERGPPGPPGRDGEDGPTGPPGPPGPPG
PPGLGGNFAAQYDGKGVGLGPGPMGLMGPRGPPGAAGAPGPQGFQGPAGEPGEPGQTGPAGARGPA
GPPGKAGEDGHPGKPGRPGERGVVGPQGARGFPGTPGLPGFKGIRGHNGLDGLKGQPGAPGVKGEPG
APGENGTPGQTGARGLPGERGRVGAPGPAGARGSDGSVGPVGPAGPIGSAGPPGFPGAPGPKGEIGAV
GNAGPAGPAGPRGEVGLPGLSGPVGPPGNPGANGLTGAKGAAGLPGVAGAPGLPGPRGIPGPVGAAGA
TGARGLVGEPGPAGSKGESGNKGEPGSAGPQGPPGPSGEEGKRGPNGEAGSAGPPGPPGLRGSPGSR
GLPGADGRAGVMGPPGSRGASGPAGVRGPNGDAGRPGEPGLMGPRGLPGSPGNIGPAGKEGPVGLPG
```

11. Save the changes made to the document. Under the "File" menu, select "Save as", name the file "yourlastname_sequence" and be sure to save the file as a .txt (plain text file).

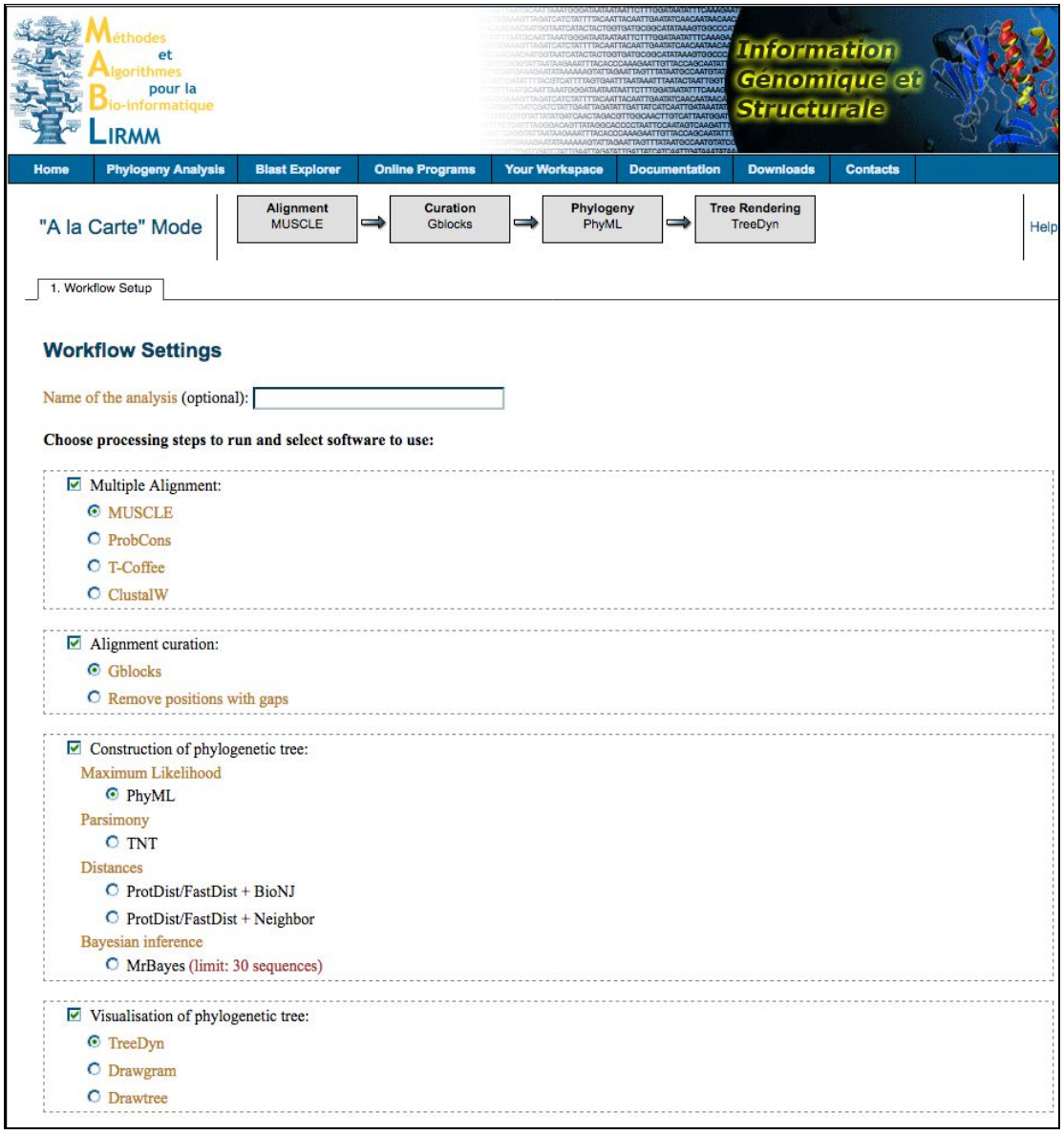
This file is now ready to be used to build your phylogenetic tree. It contains the alpha 2 type 1 collagen sequences for the *T. rex*, chicken, rainbow trout, human, dog, cattle, toxodon, mastodon, salamander, and frog.

Constructing a phylogenetic tree using MAB:

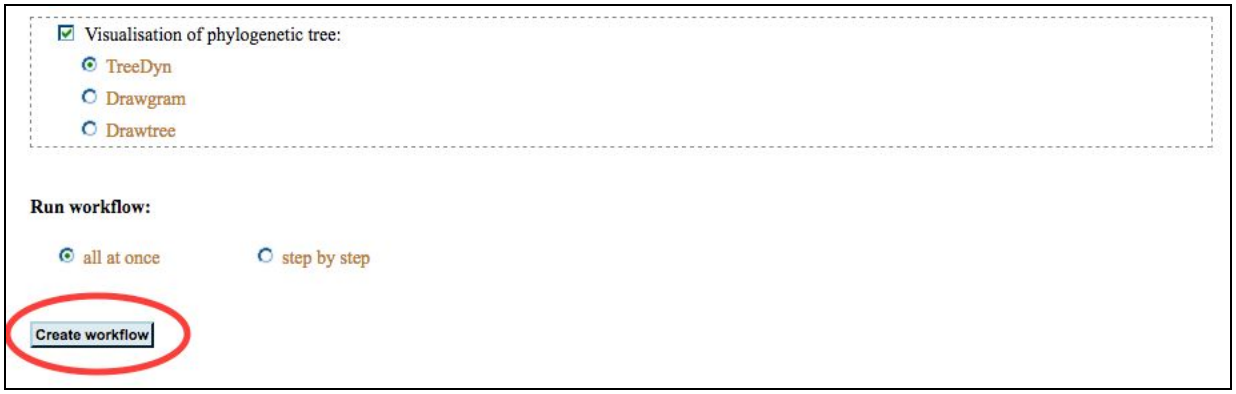
1. With your web browser, navigate to the following web page: http://www.phylogeny.fr/alacarte.cgi

This page is the MAB (Methods and Algorithms for Bioinformatics) Phylogeny Analysis tool, which you will use to generate a phylogenetic tree. (Note: most of the MAB website is in French, but the form that you will use to run the Phylogeny Analysis tool is in English.)

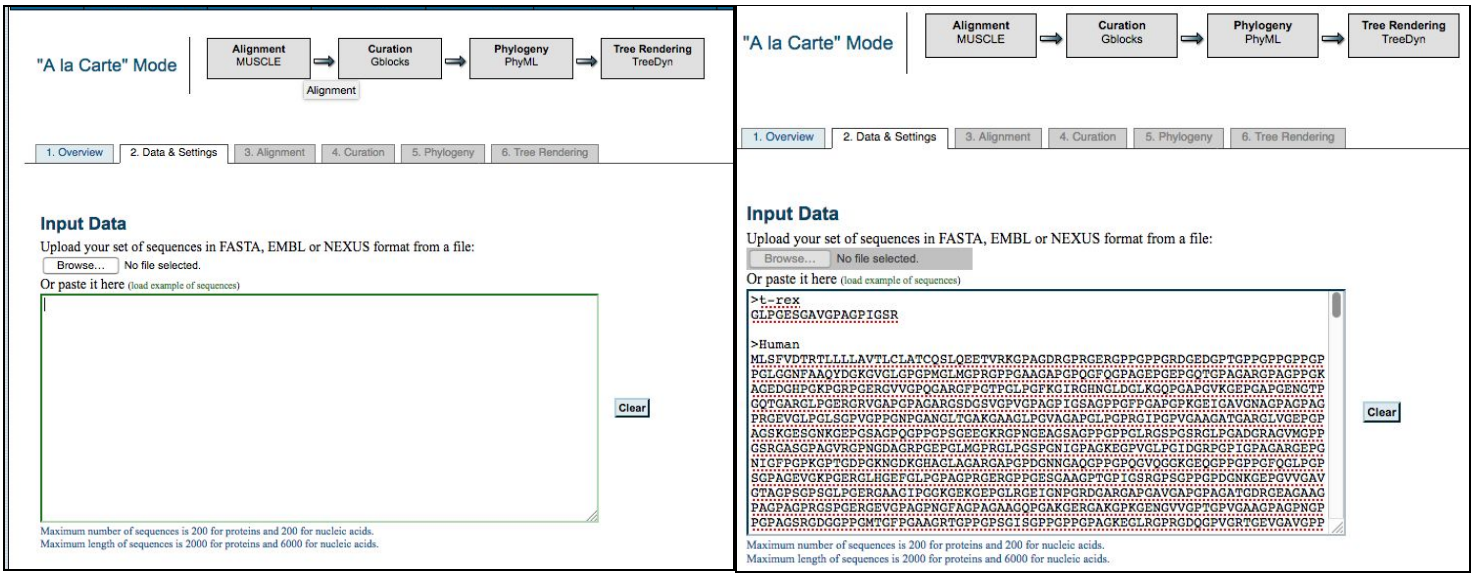
2. This link will open up directly to “A la carte” mode. Under “Workflow Settings” insert a name for your analysis.



3. Scroll to the bottom of the page and select “create workflow”. Do <u>not</u> change any of the settings, they are already set to the correct options for creating your phylogenetic tree.



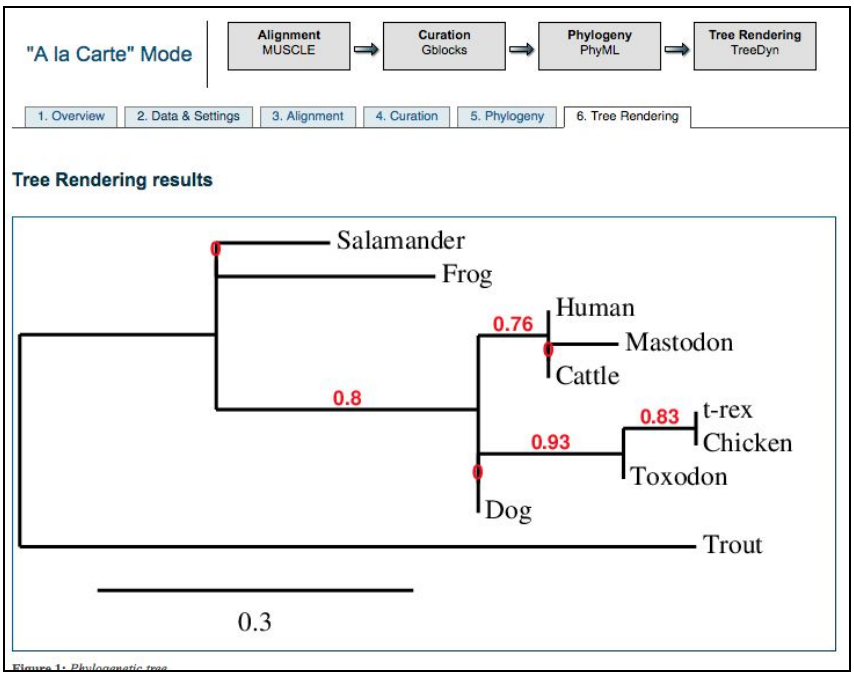
4. MAB should open up to the second browser tab, “data and settings”. This tab gives you the option to upload your file or paste the sequence. Copy and paste all of the sequences (the whole Word document) into the big text box below "Input Data" in the MAB browser window.



5. Scroll down to the bottom and enter your email address if you wish to be emailed your tree, if not select “submit”. Do not change any settings before hitting submit.

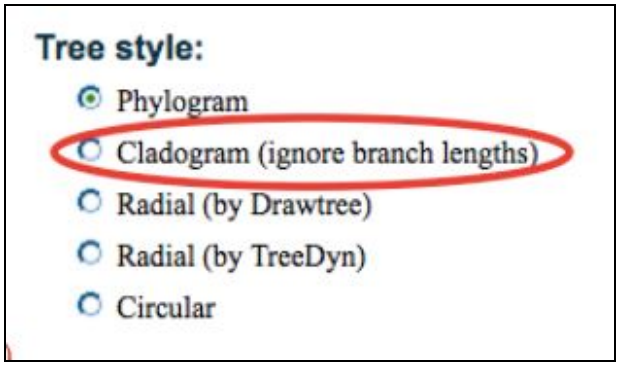
After clicking the Submit button, MAB will display a brief animation of a phylogenetic tree. During this time, MAB is aligning the sequences and then comparing them.[1]

6. MAB Phylogeny Analysis can take anywhere between 1 to 5 minutes to construct the phylogenetic tree. Once it loads, it should look like this:

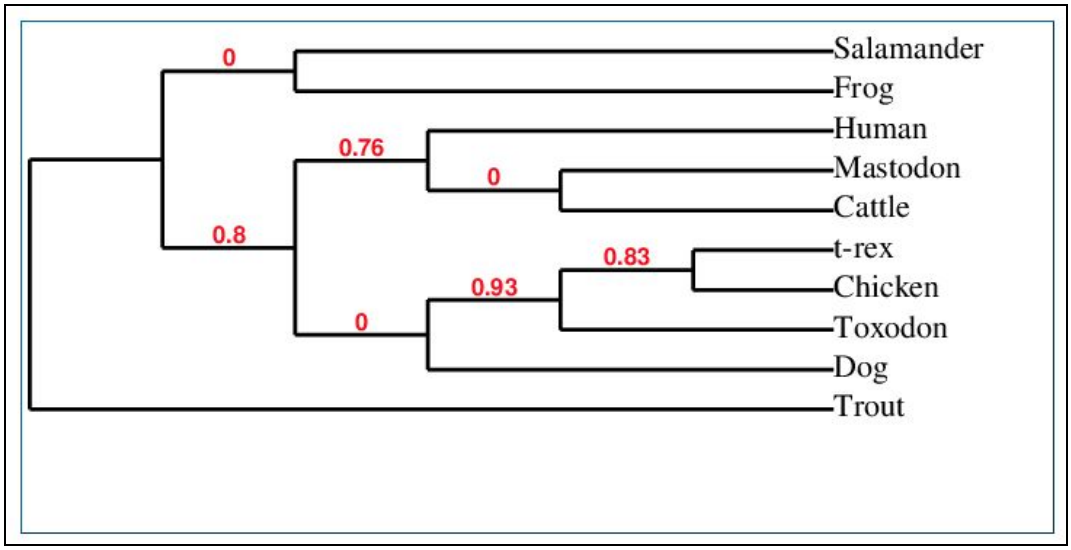


[1] This program uses a common method for aligning the sequences, called MUSCLE (Multiple Sequence Comparison by Log-Expectation). This step is important because it uses an algorithm to align each peptide sequence in order to accurately predict where mutations occured that signal how the animals evolved. If the sequences are not aligned they can not be used to generate a phylogenetic tree.

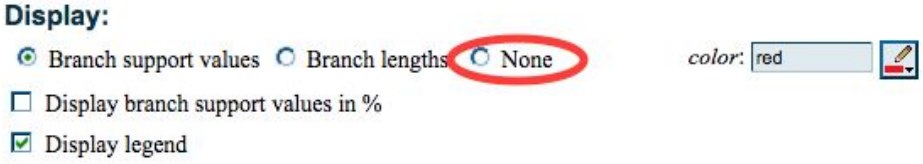
7. Scroll down to the “Tree Styles” section toward the bottom of the web page: Click on the radio button for "Cladogram” (in this context, "cladogram" is telling MAB to show a phylogenetic tree without scaling the length of tree branches based on degree of dissimilarity).



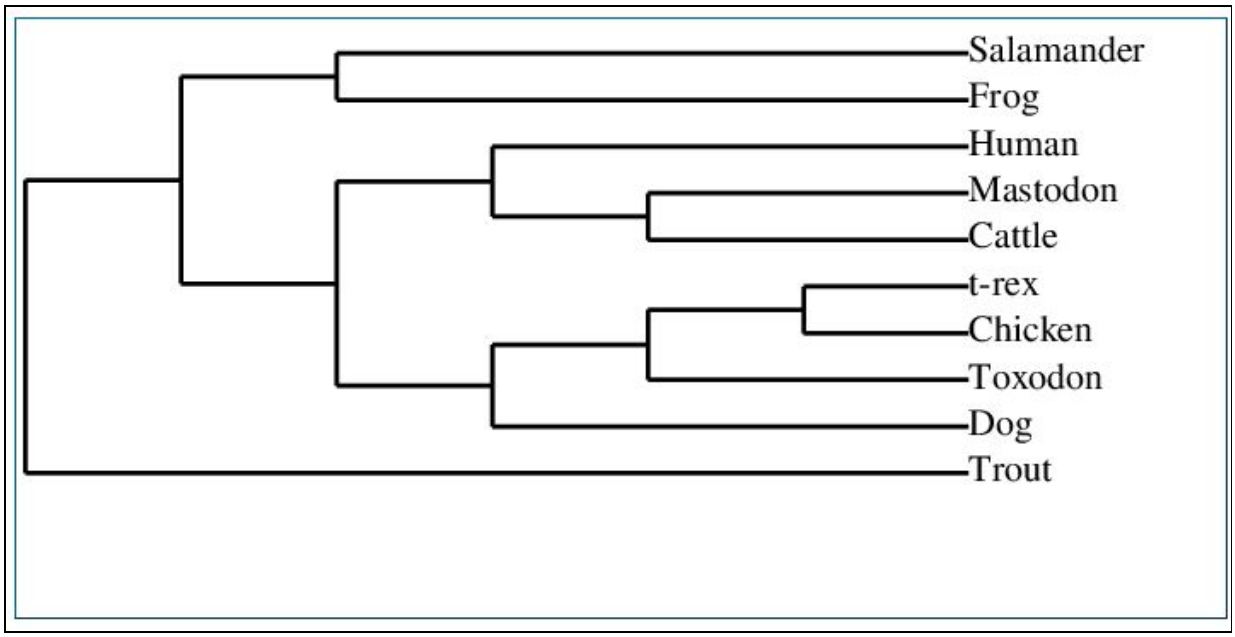
This will make it easier to read and understand the evolutionary relationships. The tree should now look like this:



8. The final setting that needs to be adjusted is under “Display:”, change the setting from “Branch support values” to “none”.



9. You now have your finished phylogenetic tree. It should look like this:



10. If you want to save your tree, you can click on the "PNG" or "PDF" links underneath the tree:

Analyzing results

1. Which of the species that you analyzed, is the *T. rex* most closely related to? Does this match with the BLAST results from Session 2?
2. Which pairs of animal species are "sister species"? (i.e., which animals are most closely related?)
3. What species is the "out-group" (i.e., the least related to the rest of the species) in this phylogenetic tree?
4. Can you find anything puzzling with the relationships depicted in this phylogenetic tree? (hint, look at dog). Do you suppose this might reflect the fact that only a very short amino acid sequence from a single gene was analyzed?

Evaluating results

1. Why is it important to understand evolutionary relationships among animals?
2. Why is it important to learn more about extinct animals?
3. Why is it important for scientists to publish their findings, such as genetic sequences, in public databases?
4. What other questions could these same techniques be used to answer?

Tree style:

- (•) Phylogram
- ( ) Cladogram (ignore branch lengths)
- ( ) Radial (by Drawtree)
- ( ) Radial (by TreeDyn)
- ( ) Circular