

Project 3.2.6 Genomic Data

Introduction

We now have the power to perform billions of calculations on billion-byte datasets. Why and how does our new superpower change other disciplines? Biology is only one of many examples, but it’s an important one! How can billions of data points from millions of organisms be compared? Not by hand! We are uncovering the intricacies of life at the molecular level: the comings, doings, and goings of some 20,000 proteins – their precursors, mechanisms, and byproducts. You could be the first to use these proteins to uncover new information about life. So much data are becoming available, so quickly, that high school students can make discoveries that scientists have not yet had the chance to extract from the data.

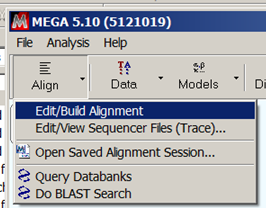
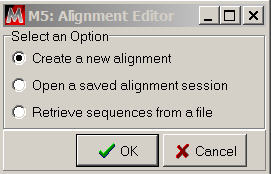
Equipment

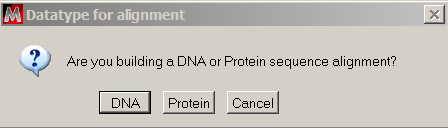
* MEGA algorithm suite (Molecular Evolution Genetics Analysis) <http://www.megasoftware.net>
* Internet connection

Procedure

In this project you will use computational power to:

1. Search a protein database for keywords and retrieve the encoding human (*Homo sapiens*)DNA sequence
2. Find DNA sequences in other species that are similar to *Homo sapiens* DNA sequence for that protein
3. Align the various species’ DNA sequences
4. Build a tree that shows relationships between species with similar DNA sequences
5. Form pairs as directed by your teacher.
6. Launch MEGA.
7. Search a protein database for keyword to retrieve a human DNA sequence, as follows.
   1. In MEGA, Choose **Align** > **Edit/Build Alignment**. Choose to **Create a new alignment** in the dialog box, and select **DNA** for the datatype. This will open a second MEGA window, the MEGA Alignment Explorer.

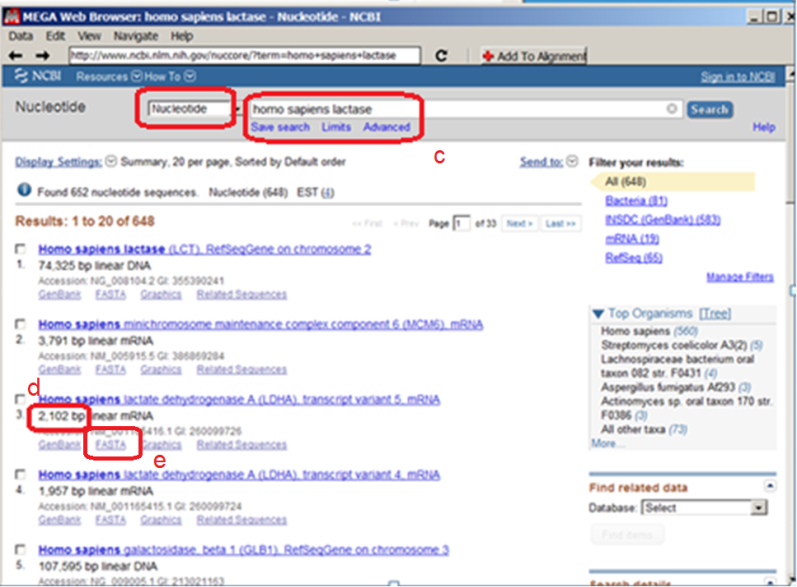


* 1. In the MEGA Alignment Explorer, choose **Web** > **Query Genbank**. This will open a third MEGA window, the MEGA Web Browser. The browser will already be directed to a search page at the National Center for Biotechnology Information (NCBI), an agency of the U.S. government within the National Institutes for Health (NIH). NCBI is in charge of GenBank, a repository of DNA sequences published by scientists. As of 2013, it contains over 160 million DNA sequences totaling over 150 billion base pairs.



* 1. On the NCBI webpage in the MEGA Web Browser, enter ‘homo sapiens’ and the name of a human protein in the search box, as shown in ‘c’ in the figure below. Note that you are searching for nucleotide sequences tagged with these keywords. You can use any human protein; some are suggested on the Project 3.2.6 Supplement A: Human Proteins. Use the all capital-letter symbol for the protein from the protein’s Wikipedia page as shown in the figure below.

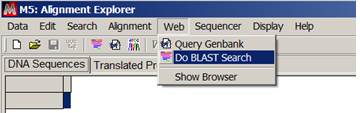




* 1. If one of the sequences seems like the protein you were looking for, check the length of the DNA sequence in basepairs, as shown in ‘d’ in the figure above. You want a result that is around 500-2000 basepairs; shorter sequences will match too many sequences from other species, and longer sequences will take too much computation to align with other sequences in the time you have. You also want a result that is mRNA.
  2. Select the FASTA file for your sequence. See ‘e’ in figure above.
  3. The MEGA Web browser will navigate to the webpage for this DNA sequence that has been uploaded by the scientist(s) who collected the DNA sequence data. Copy the NCBI reference number for this sequence, as noted with the red box, by shift-click highlighting it from beginning to end, and using Ctrl-C or **Edit>Copy** to copy the string into the clipboard.



1. In this step you will use one of the most important algorithms in biology: the Basic Local Alignment Search Tool (BLAST). The BLAST algorithm takes one sequence of DNA and finds matching DNA amongst billions of other sequences. BLAST was invented in 1990 by five people at NCBI and at the Computer Science departments at Penn State and the University of Arizona. You will run the BLAST algorithm on high-speed computers at NCBI. It will take a few minutes because you are sharing those computers with the world’s scientists. The algorithm will return to you the most closely matching sequences among the ~1 billion sequences in the NCBI database. Some of the database’s sequence are entire chromosomes, millions of base pairs long, but many are “expressed sequence tags,” the sequence of mRNA that has been made in the nucleus and gets exported to the ribosomes for translation into protein.
   1. In the MEGA Alignment Explorer, choose **Web** > **Do BLAST Search**. This will open another MEGA Web Browser window, already directed to the BLAST entry page at NCBI.



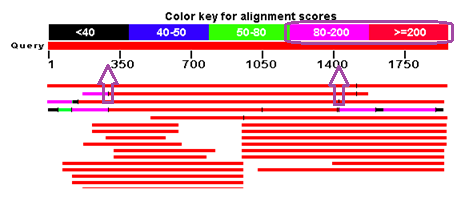
* 1. Choose tblastx as shown in ‘b’ in figure below. The tblastx option takes a DNA sequence you provide, translates it to protein, and uses the BLAST algorithm to search for protein sequences that align with the one you give, translated from DNA sequences in the database. You can set the translation code; we’ll search based upon the standard genetic code for nuclear DNA; mitochondria use a slightly different code.



* 1. As shown in ‘c’ in the figure above, paste the accession number from step 3f into the FASTA sequence window; BLAST will use the sequence corresponding to this accession number as your “query” sequence.
  2. Click the BLAST button at the bottom of the page. The algorithm is running on NCBI’s very fast computers, but it could take 5 or 10 minutes.



* 1. Take a screenshot of the Graphic Summary of your alignment results and paste it below the example here. The example here has three purple annotations used in the template below.

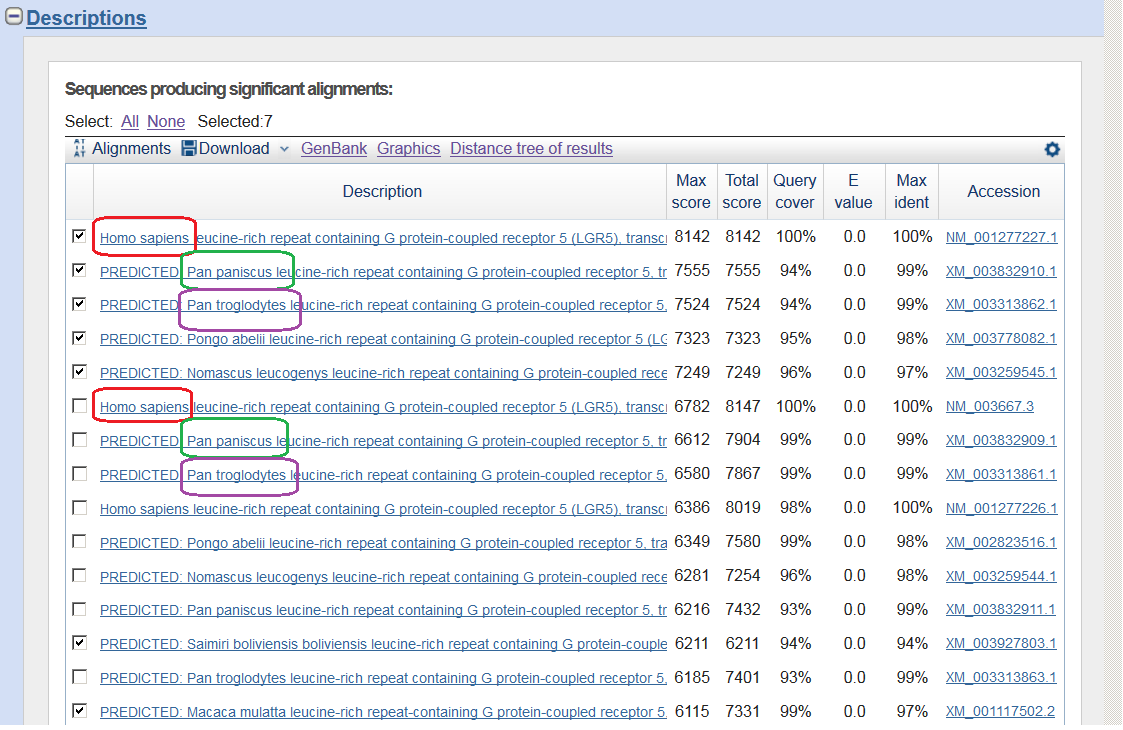


Each horizontal bar represents one search result subject. The color coding shows the extent and quality of the result’s alignment with your query DNA sequence; red regions have the best alignment. Pick one result from your screenshot and explain the representation of the horizontal bar to your partner, using the following example as a template:

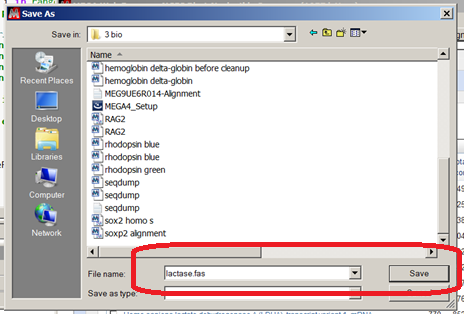
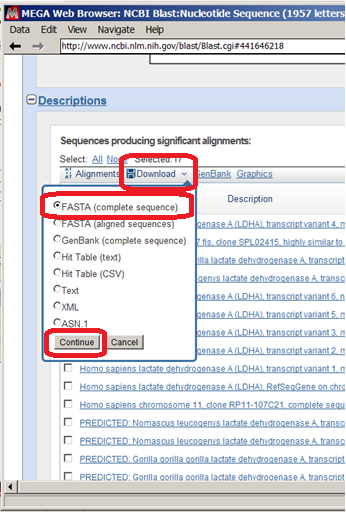
*“In the example results shown above, the fourth result aligns with the entire DNA sequence we provided, but only aligns well (score>=200) from base ~300 to base ~1450.”*

Your partner will pick another result to explain to you. Record both of your explanations along with your screen shot here.

* 1. Now look at the Descriptions section of the BLAST report. “Predicted” means that the scientist that submitted the DNA is only predicting what the protein does in the organism based on what a similar protein does in another organism. For each species that appears, select the checkbox of the highest scoring “mRNA” result. Do not worry about the “predicted” notation, since we’re only going to be using the fact that the DNA is in the organism.

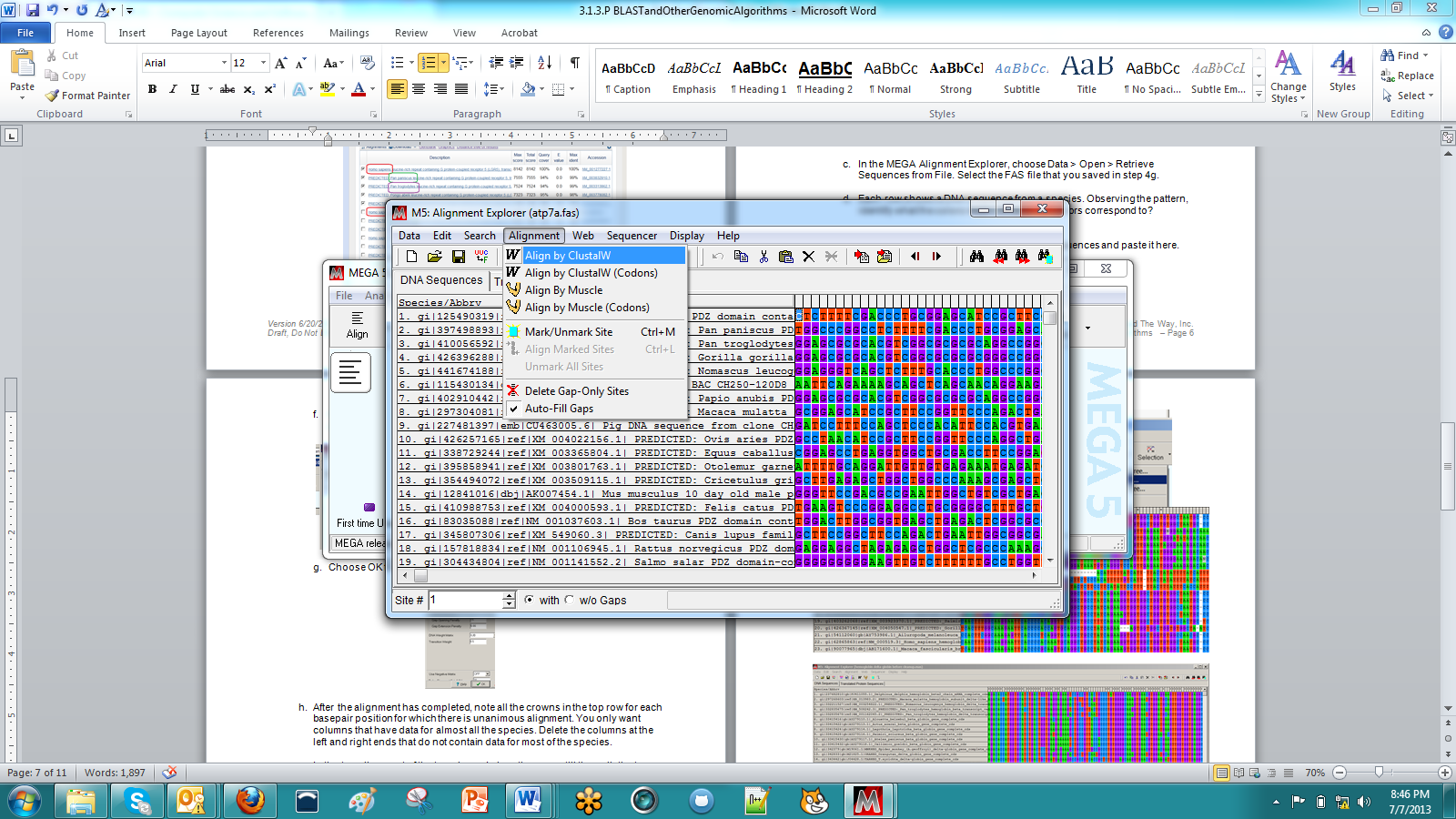
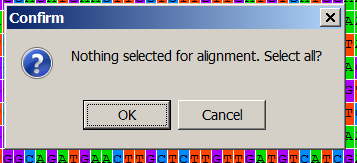


* 1. Download the selected results, using FASTA format for the complete sequence. Save the file using the name of your protein. Type .fas for the file extension.

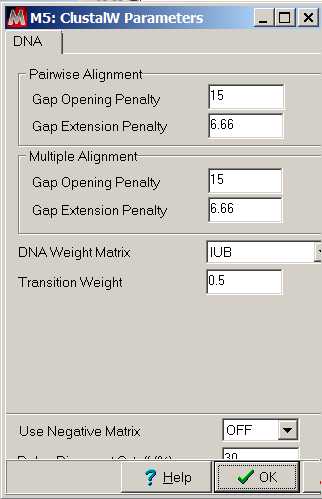




1. In this step, you will use your computer to align the sequences you downloaded. You will use the ClustalW algorithm, which is much slower but more accurate than BLAST.
   1. When you search for matches to a DNA sequence, why would it be impractical for NCBI computers to identify matching sequences in their database using the ClustalW algorithm?
   2. Why is it practical now for you to use ClustalW to align sequences to each other that you downloaded?
   3. In the MEGA Alignment Explorer, choose **Data** > **Open** > **Retrieve Sequences from File**. Select the FAS file that you saved in step 4g.
   4. Each row shows a DNA sequence from a species. Observing the pattern, identify what the colors mean. What do the colors correspond to?
   5. Take a screenshot of this unaligned list of sequences and paste it here.
   6. In the MEGA Alignment Explorer, choose **Alignment** > **Align by ClustalW**. Choose OK to select all when prompted.

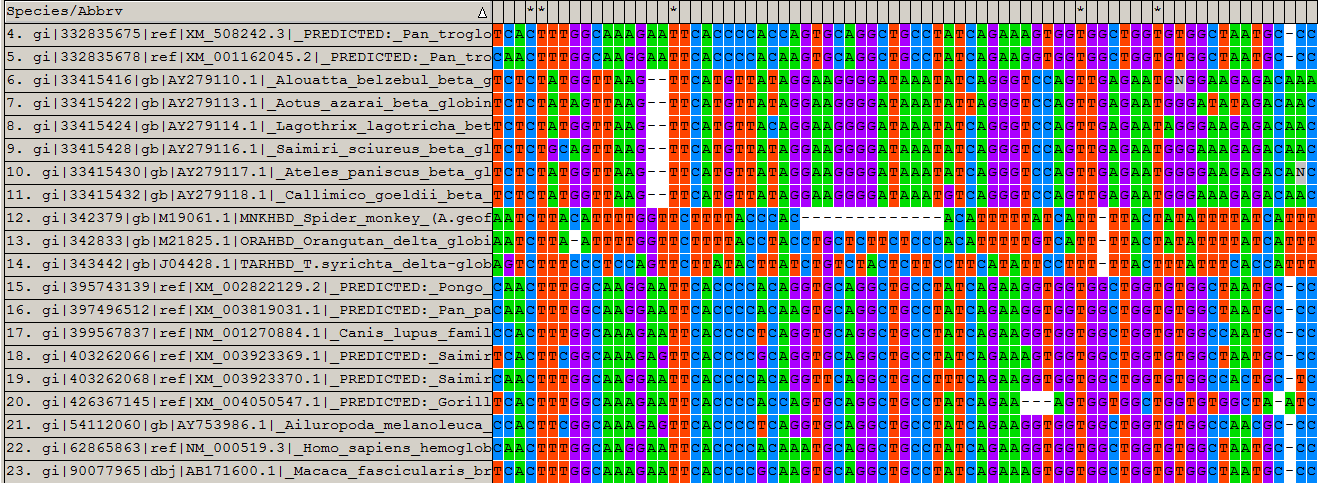
* 1. Choose **OK** to accept defaults for ClustalW parameters.

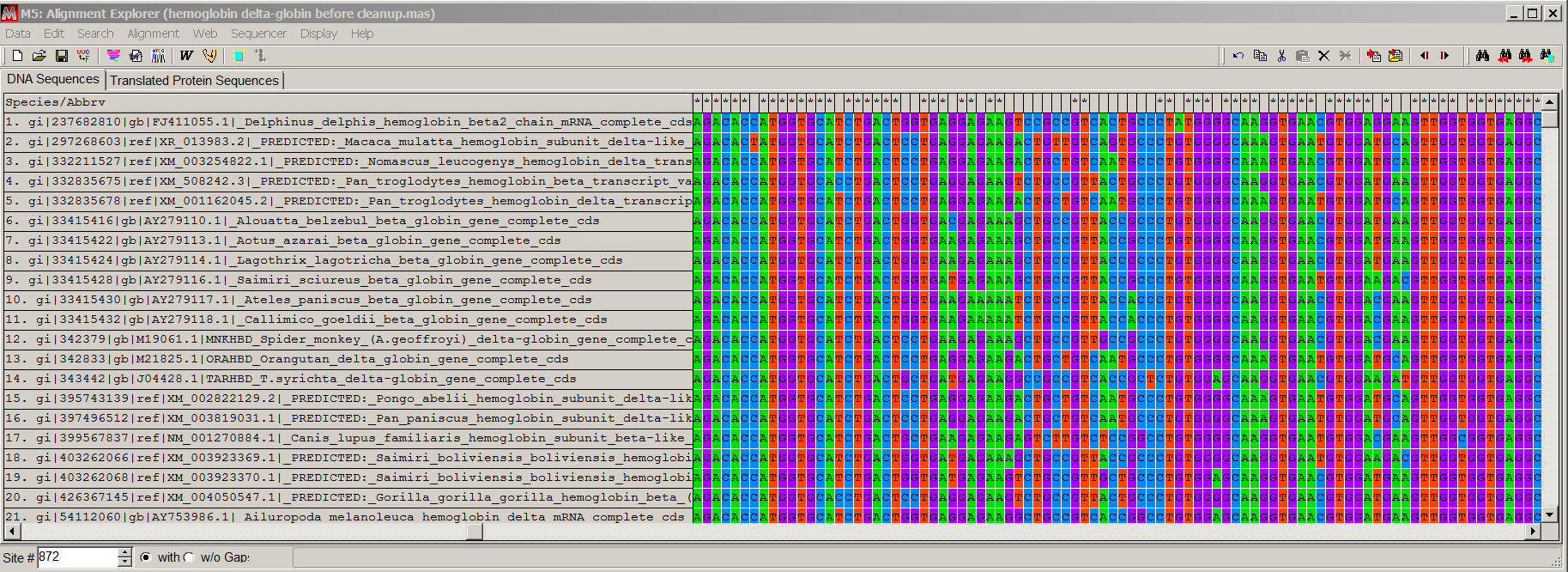


* 1. After the alignment has completed, note all the crowns in the top row for each basepair position for which there is unanimous alignment. You only want columns that have data for almost all the species. Delete the columns at the left and right ends that do not contain data for most of the species.

The two windows shown below are alignments of a group of sequences. Colorizing the sequence elements helps you see patterns and make adjustment by hand. In the top window, it looks like the group of sequences has four subgroups that each have an aligned pattern. Notice that the first and fourth subgroups are actually aligned with each other; the interface allows you to drag rows to help visually organize the alignment.

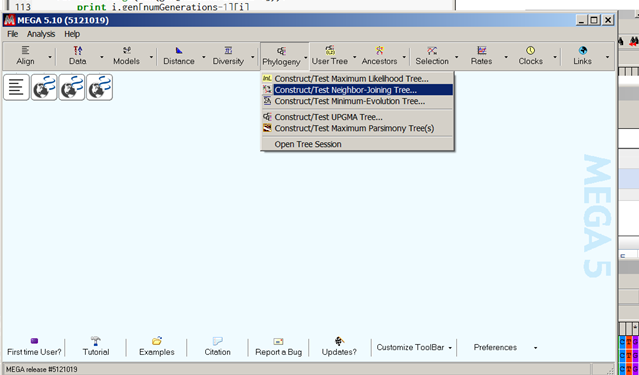
The second alignment window shown below is ready to be used to calculate a phylogenetic tree; you do not need the entire gene to be aligned. You should, however, be able to create a good alignment across all the species you used across the overlapping regions of alignment indicated in the Graphic Summary of BLAST results from Step 4e.



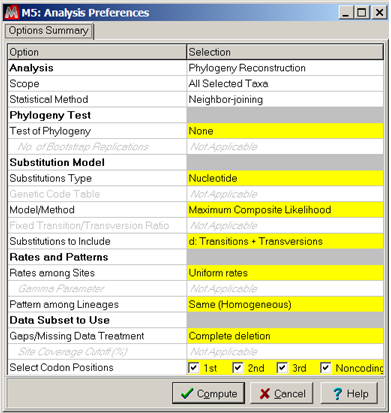


* 1. Export your alignment as a MEGA file so that you can use it with another algorithm by choosing **Data** > **Export Alignment** > **MEGA format**. Name the data by the protein you used. Select “**Yes**” when asked if these are protein-coding data.

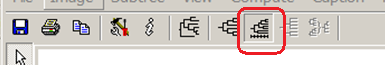
1. Now you will use another algorithm in MEGA: the nearest-neighbor joining algorithm for making a tree.
   1. In the MEGA main window, choose **Phylogeny** **> Construct/Test Neighbor-Joining Tree**. Select your file.



* 1. In the resulting dialog box, choose Compute and accept all of the default Analysis Preferences. This will open another window, the MEGA Tree Explorer.

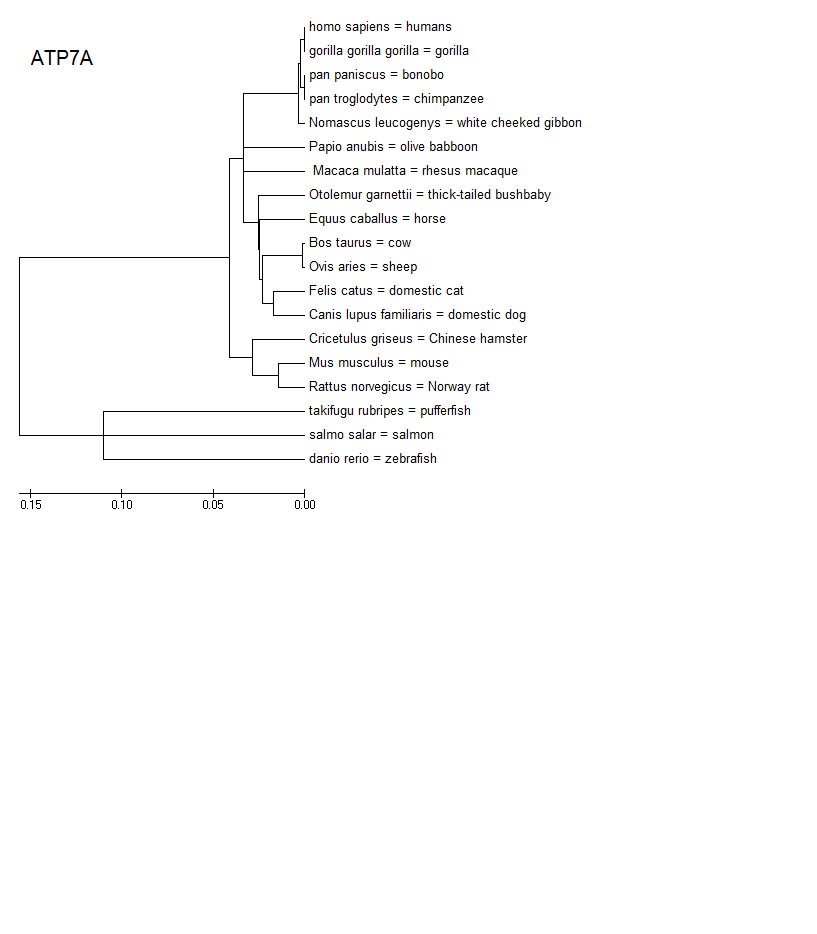


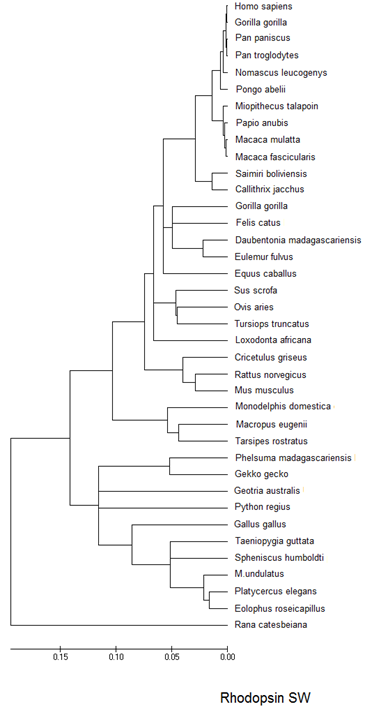
* 1. The **Compute Linearized Tree** button can be used to toggle whether horizontal lengths are proportional to genetic change.

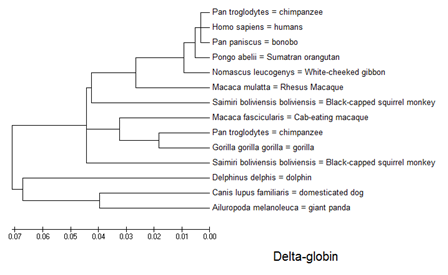


* 1. Using the snipping tool and paint, the tree can be annotated with species’ common names from the Project 3.2.6 Supplement B: List of Mammals.

Conclusion

1. In the introduction of this project, it was said that we now had the power to perform billions of calculations on billion-byte datasets. Describe the dataset(s) used in this project, and describe the type of algorithms that required the billions of calculations.
2. Generalize the relationships shown in your phylogenetic tree. What groupings and relationships between groupings do you observe? There is an additional tree shown here, and there are two more at the end of the conclusion questions. 
3. Exchange trees with one or more other groups. Explain your tree to the other groups, and have other groups explain their phylogenetic trees to you. What similarities and differences do you observe among the trees?
4. The tree shown here is for ATP7A, a protein used to transport copper in the body. The horizontal distance shown to the right of the division of cows and sheep is less than the horizontal distance shown after the division of cats and dogs. What does this distance indicate about the results of this calculation?
5. The collection of each sequence in the databank at NCBI requires a tremendous amount of laboratory work, typically several people working full time for a few months. How are published data an example of abstraction?
6. Biology is just one field that has been impacted by computing. Using the Internet, research and write about the impact of computation on another field.
7. Explain why the sort of discoveries you made in this project would not have been possible before the Internet.

****The following phylogenetic trees are provided as examples. You might find them useful in writing your responses to the preceding conclusion questions, or your teacher might find them useful for facilitating discussion.

****