PLTW COMPUTER SCIENCE

Project 3.2.6

Genomic Data

Introduction

We now have the power to perform billions of calculations on billion-byte datasets. Why 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 uncover relationships of those mechanisms among different forms of life. So much data are becoming available, so quickly, that high school students can make new 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

Resources

Supplement A: Biology Vocabulary

Supplement B: List of Proteins

Supplement C: Standard Genetic Code

Supplement D: List of Mammals

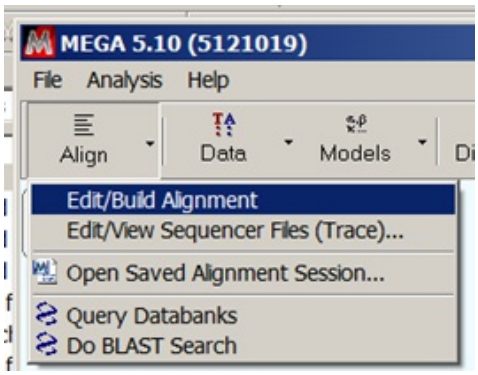
Procedure

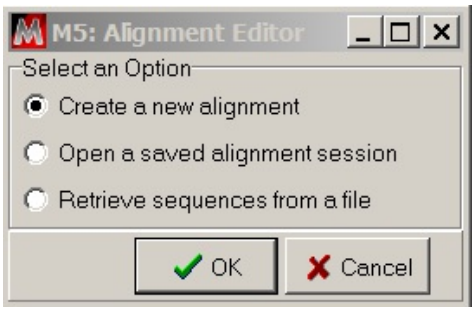
In this project you will learn some biology vocabulary and use computational power to:

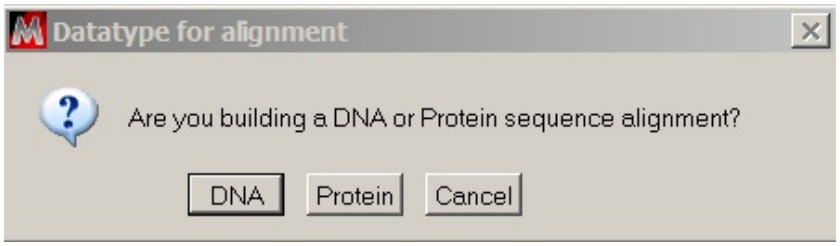
1. Search a protein database for keywords and retrieve the encoding human (*Homo sapiens*) DNA sequence
2. Find DNA sequences in other species that are similar to *Homo sapiens* DNA sequence for that protein
3. Align the various species' DNA sequences
4. Build a tree that shows relationships between species with similar DNA sequences

1. Form pairs as directed by your teacher.
2. Become familiar with some biology terms and vocabulary.**Supplement A: Biology Vocabulary**
3. Launch MEGA.

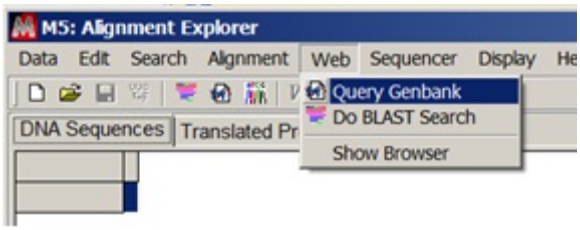
4. Search a protein database for keyword to retrieve a human DNA sequence, as follows.
    - In MEGA, Choose **Align > Edit/Build Alignment**. Choose to **Create a new alignment** in the dialog box, and select DNA for the datatype. This will open a second MEGA window, the MEGA Alignment Explorer.





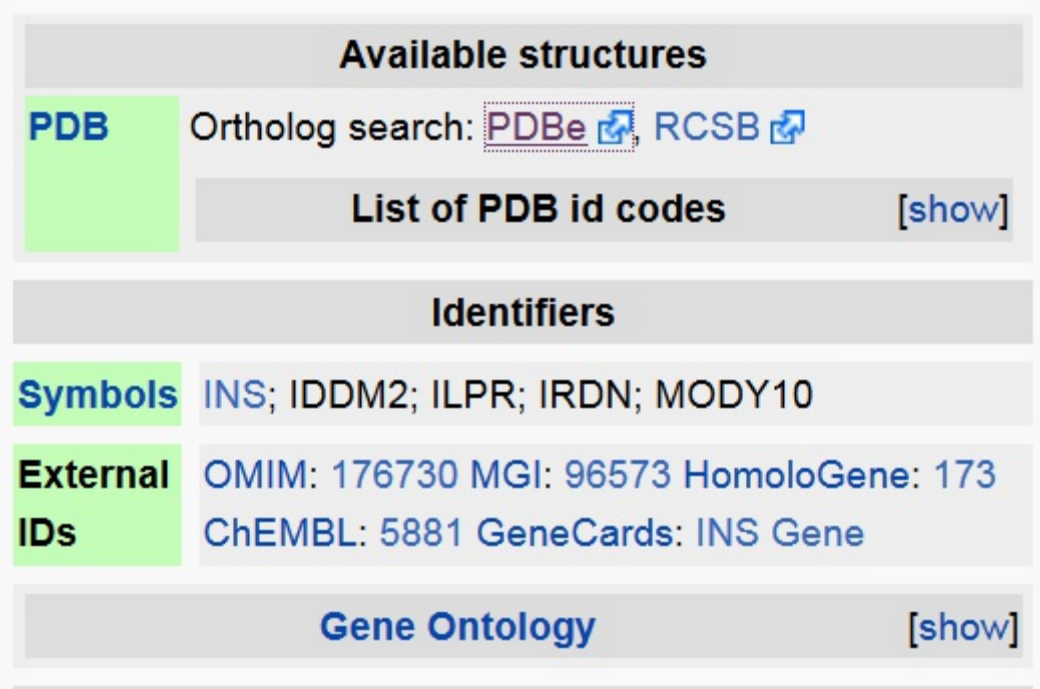


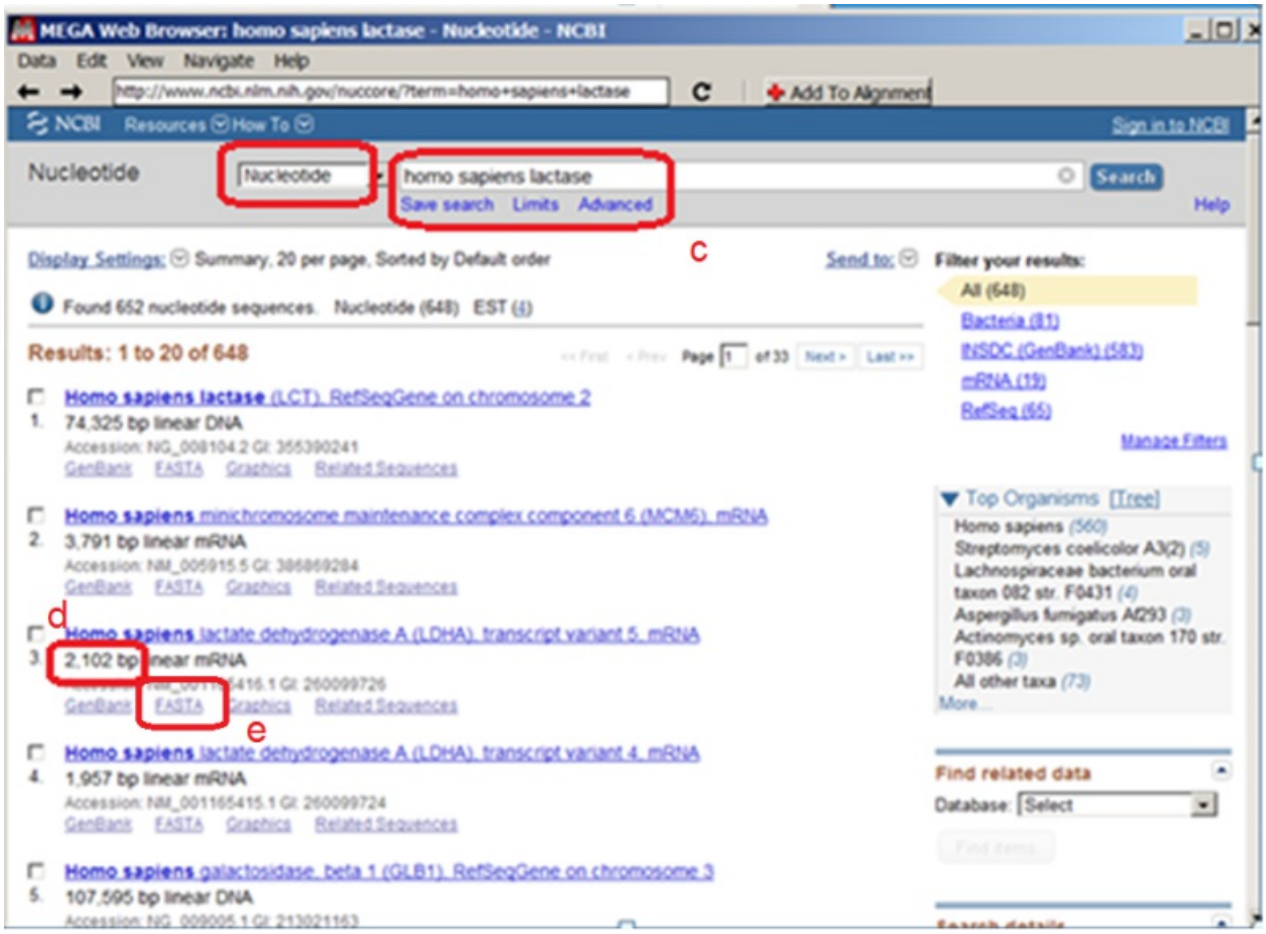
    - In the MEGA Alignment Explorer, choose **Web > Query Genbank**. This will open a third MEGA window, the MEGA Web Browser. The browser will already be directed to a search page at the National Center for Biotechnology Information (NCBI), an agency of the U.S. government within the National Institutes 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.



    - 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 B: Human Proteins**. Use the all capital-letter symbol for the protein from the protein’s wikipedia page as shown in the figure below.

Supplement B: List of Proteins

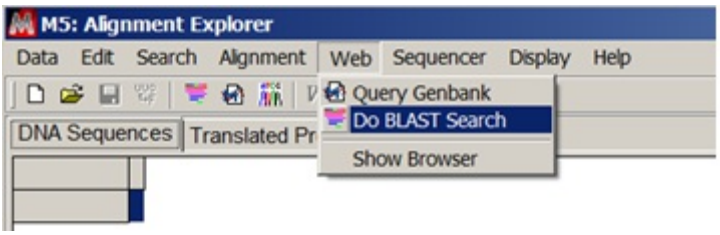




- If one of the sequences seems like the protein you were looking for, check the length of the DNA sequence in basepairs, as shown in ‘d’ in the figure above. You want a result that is around 500-1000 basepairs; shorter sequences will match too many sequences from other species, and longer sequences will take too much computation to align with other sequences in the time you have. You also want a result that is mRNA.

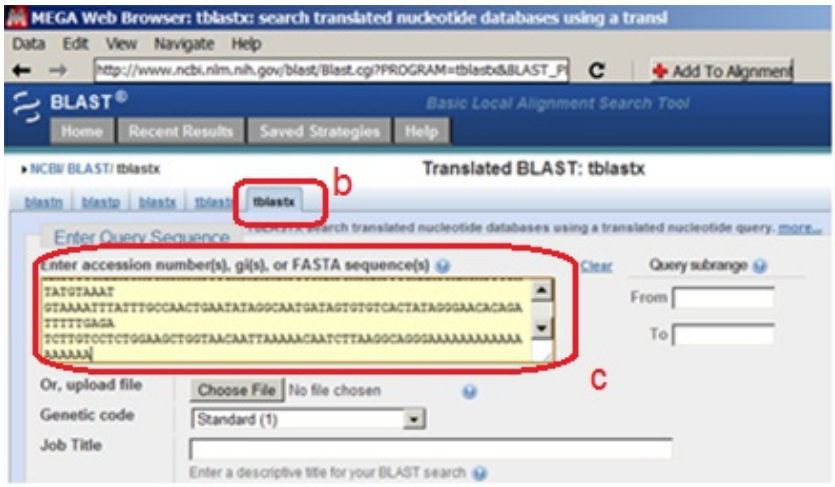
- Select the FASTA file for your sequence. See ‘e’ in figure above.
- The MEGA Web browser will navigate to the webpage for this DNA sequence that has been uploaded by the scientist(s) who collected the DNA sequence data. Copy the DNA sequence by click-and-drag or by shift-click highlighting it from beginning to end, and using Ctrl-C or **Edit > Copy** to copy the string into the clipboard.

5. 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 millions 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 160 million sequences in the NCBI database. Some of the database’s sequence are entire chromosomes, millions of base pairs long, but most 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.
   - 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.



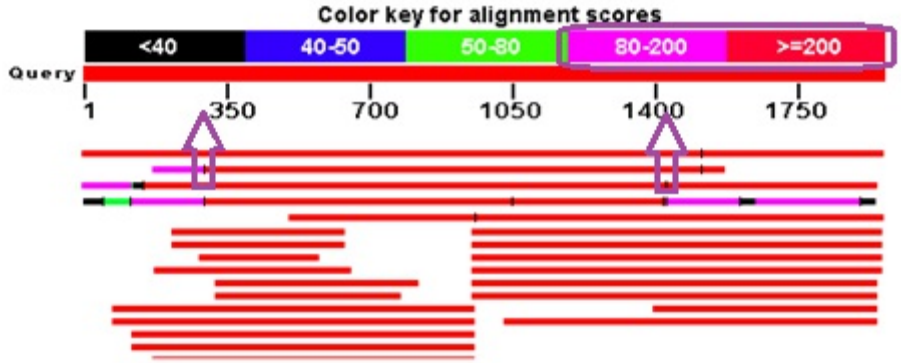
   - Choose tblastx as shown in ‘b’ in figure below. The tblastx option takes a DNA sequence you provide, translates it to protein, and uses the BLAST algorithm to search for protein sequences that align with the one you give, translated from DNA sequences in the database. You can set the translation code;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

     Supplement C: Standard Genetic Code



- As shown in ‘c’ in the figure above, paste the DNA from step 4f into the FASTA sequence window; BLAST will use this as your “query” sequence.
- 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.

- 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.

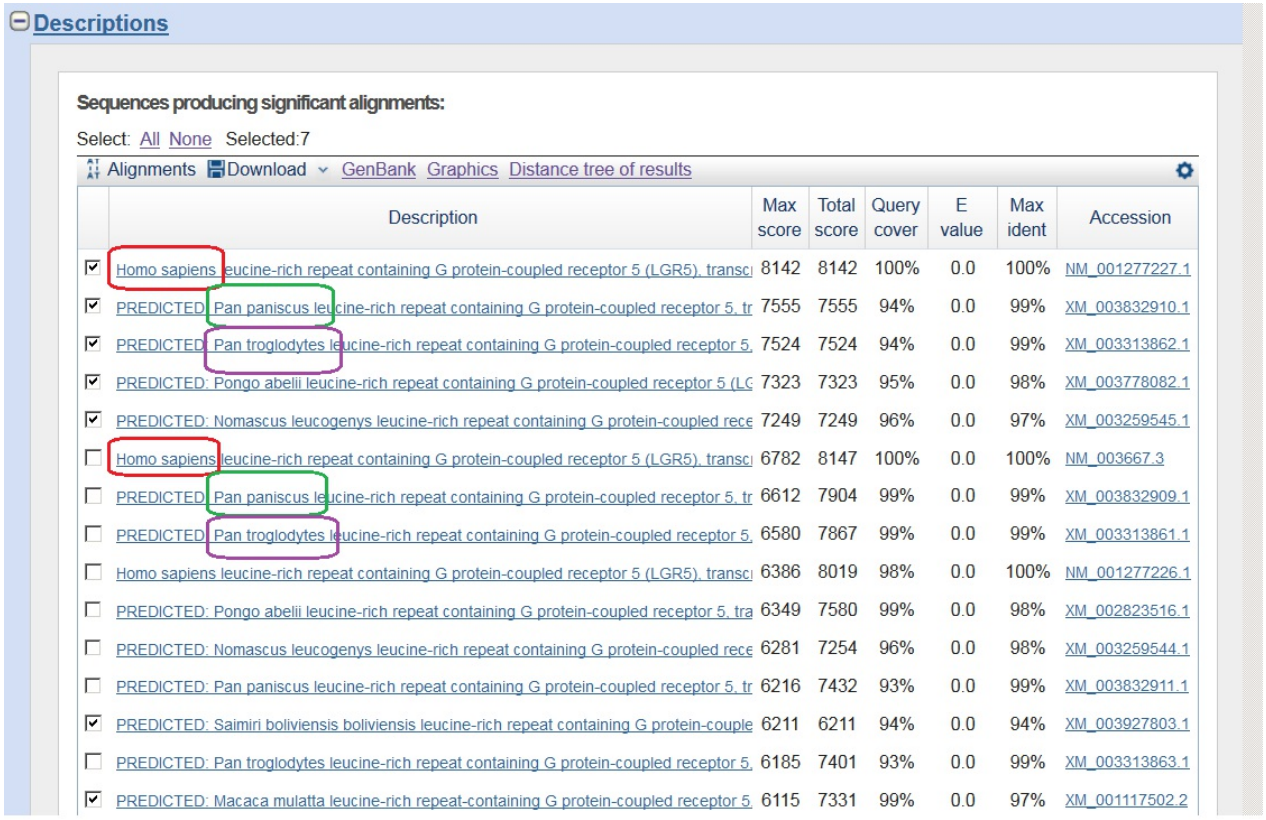


Each horizontal bar represents one search result subject. The color coding shows the extent and quality of the result’s alignment with your query DNA sequence; red regions have the best alignment. Pick one result from your screenshot and explain the representation of the horizontal bar to your partner,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

- 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.



Descriptions

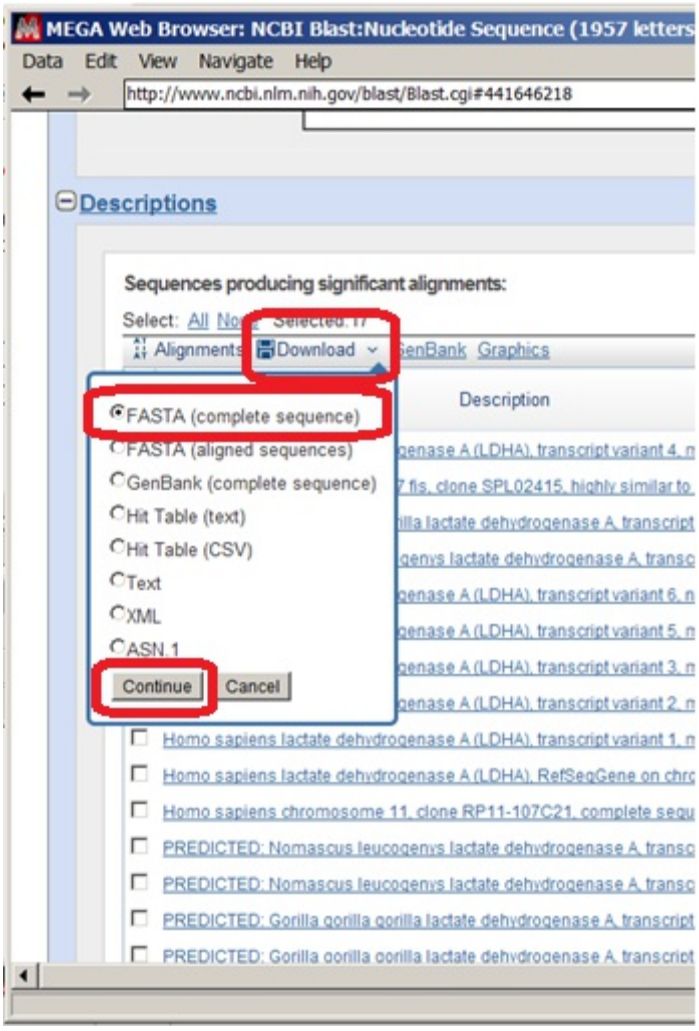
Sequences producing significant alignments:

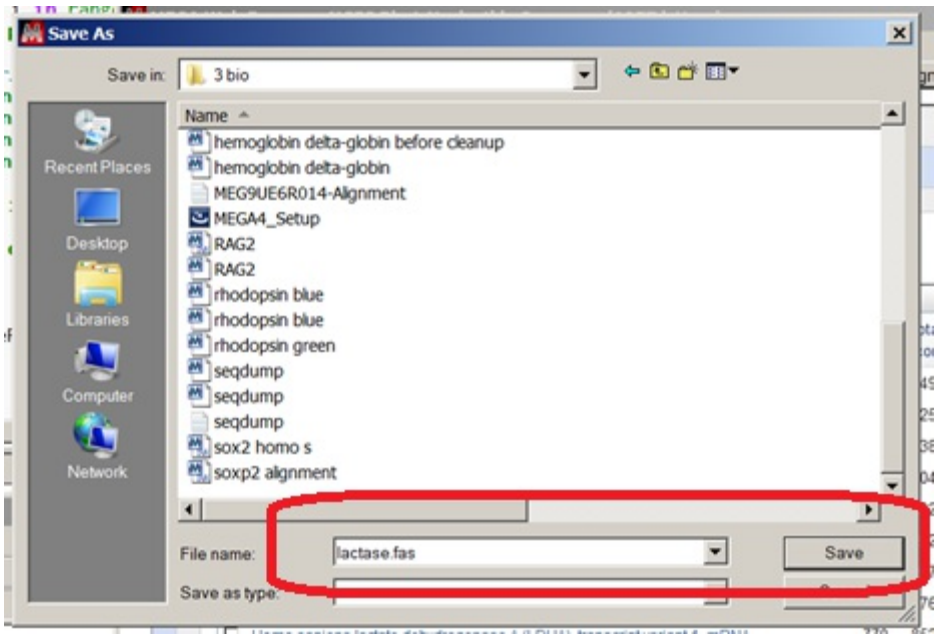
Select: All None Selected:7

Alignments Download GenBank Graphics Distance tree of results

	Description	Max score	Total score	Query cover	E value	Max ident	Accession
☑	Homo sapiens leucine-rich repeat containing G protein-coupled receptor 5 (LGR5), transc	8142	8142	100%	0.0	100%	NM_001277227.1
☑	PREDICTED: Pan paniscus leucine-rich repeat containing G protein-coupled receptor 5, tr	7555	7555	94%	0.0	99%	XM_003832910.1
☑	PREDICTED: Pan troglodytes leucine-rich repeat containing G protein-coupled receptor 5,	7524	7524	94%	0.0	99%	XM_003313862.1
☑	PREDICTED: Pongo abelii leucine-rich repeat containing G protein-coupled receptor 5 (LC	7323	7323	95%	0.0	98%	XM_003778082.1
☑	PREDICTED: Nomascus leucogenys leucine-rich repeat containing G protein-coupled rece	7249	7249	96%	0.0	97%	XM_003259545.1
☐	Homo sapiens leucine-rich repeat containing G protein-coupled receptor 5 (LGR5), transc	6782	8147	100%	0.0	100%	NM_003667.3
☐	PREDICTED: Pan paniscus leucine-rich repeat containing G protein-coupled receptor 5, tr	6612	7904	99%	0.0	99%	XM_003832909.1
☐	PREDICTED: Pan troglodytes leucine-rich repeat containing G protein-coupled receptor 5,	6580	7867	99%	0.0	99%	XM_003313861.1
☐	Homo sapiens leucine-rich repeat containing G protein-coupled receptor 5 (LGR5), transc	6386	8019	98%	0.0	100%	NM_001277226.1
☐	PREDICTED: Pongo abelii leucine-rich repeat containing G protein-coupled receptor 5, tra	6349	7580	99%	0.0	98%	XM_002823516.1
☐	PREDICTED: Nomascus leucogenys leucine-rich repeat containing G protein-coupled rece	6281	7254	96%	0.0	98%	XM_003259544.1
☐	PREDICTED: Pan paniscus leucine-rich repeat containing G protein-coupled receptor 5, tr	6216	7432	93%	0.0	99%	XM_003832911.1
☑	PREDICTED: Saimiri boliviensis boliviensis leucine-rich repeat containing G protein-couple	6211	6211	94%	0.0	94%	XM_003927803.1
☐	PREDICTED: Pan troglodytes leucine-rich repeat containing G protein-coupled receptor 5,	6185	7401	93%	0.0	99%	XM_003313863.1
☑	PREDICTED: Macaca mulatta leucine-rich repeat-containing G protein-coupled receptor 5	6115	7331	99%	0.0	97%	XM_001117502.2

- 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.





6. 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.
    - 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?

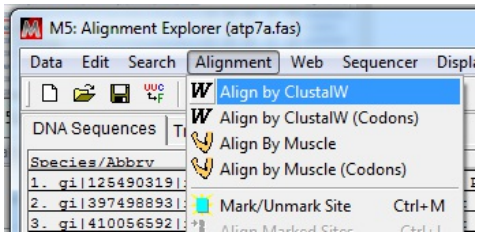
    - Why is it practical now for you to use ClustalW to align sequences to each other that you downloaded?

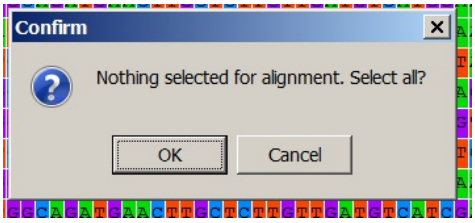
    - In the MEGA Alignment Explorer, choose **Data > Open > Retrieve Sequences from File**. Select the FAS file that you saved in step 5g.

    - Each row shows a DNA sequence from a species. Observing the pattern, identify what the colors mean. What do the colors correspond to?

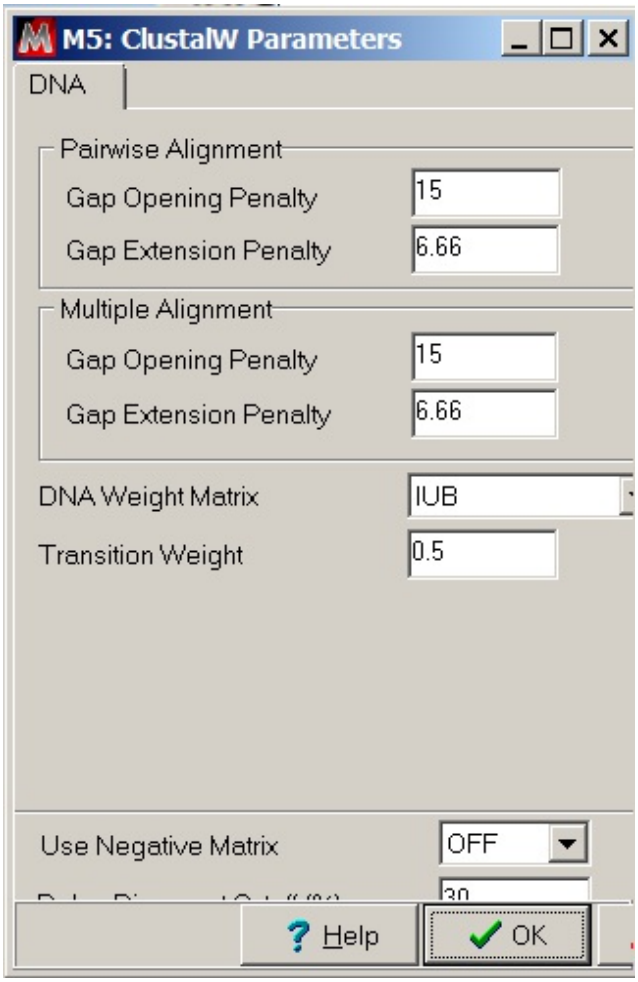
    - Take a screenshot of this unaligned list of sequences and paste it here.

    - In the MEGA Alignment Explorer, choose **Alignment > Align by ClustalW**. Choose OK to select all when prompted.



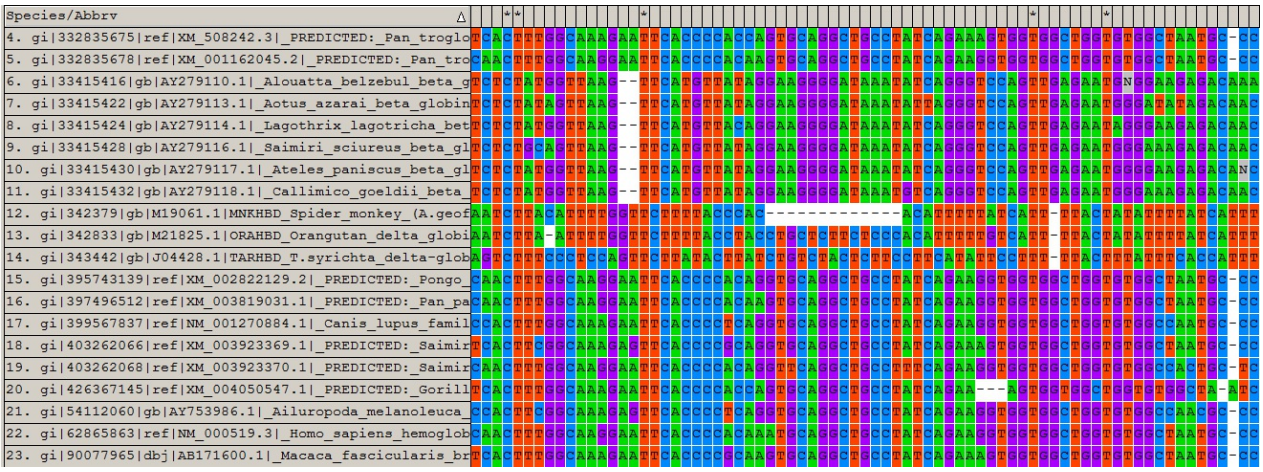


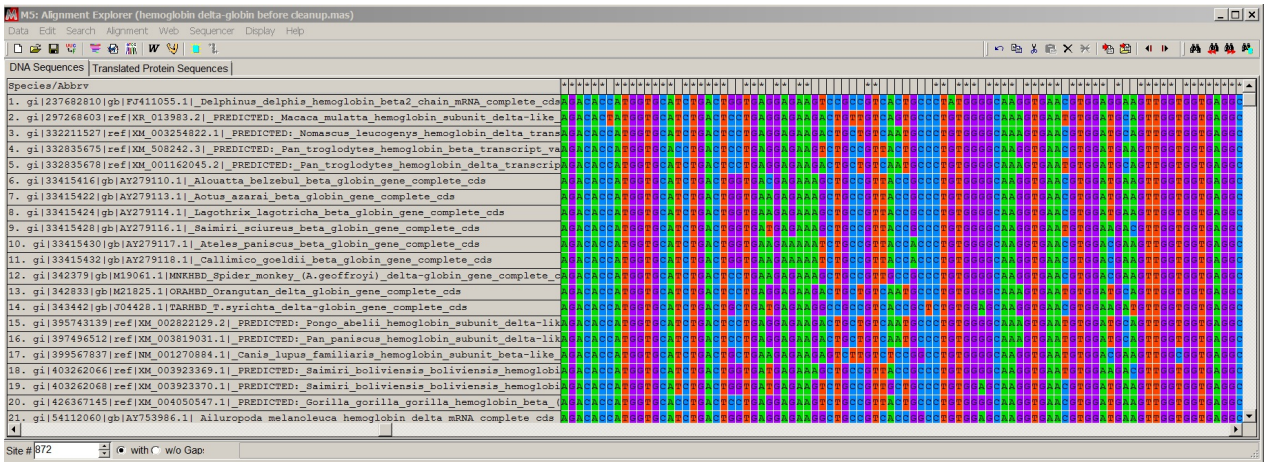
    - Choose **OK** to accept defaults for ClustalW parameters.



- 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.

  In the top alignment of the two shown below, there are still three distinct groups that do not align with each other. The bottom of the two alignments 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 5e.

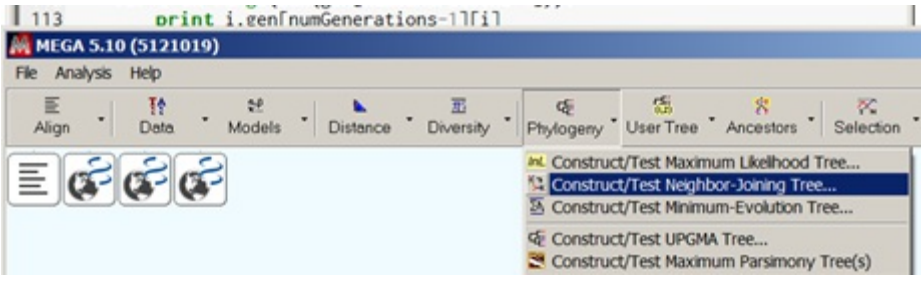




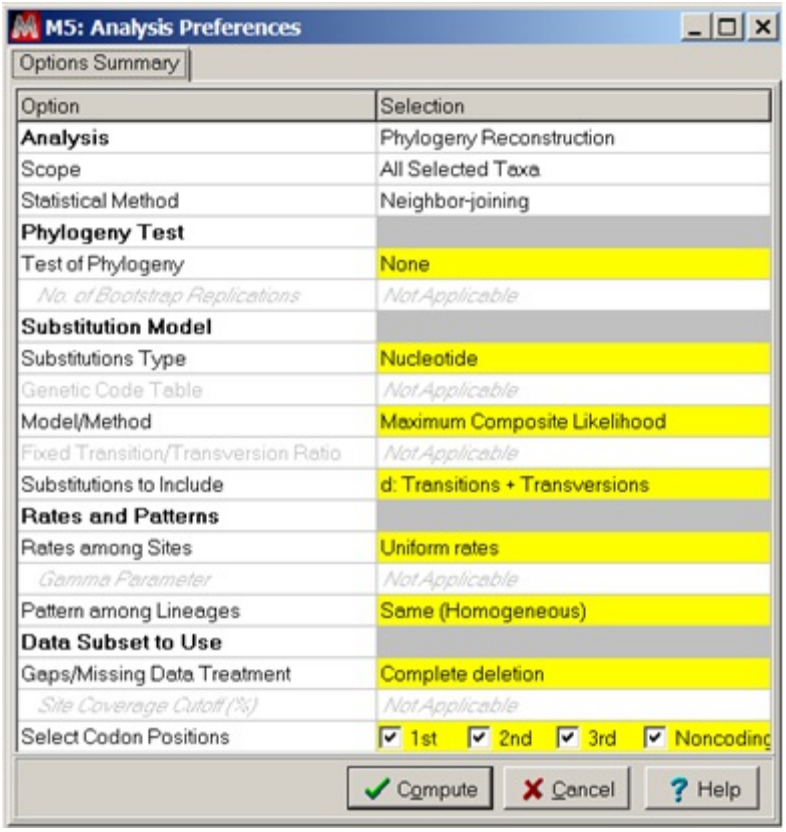
- Export your alignment as a MEGA file so that you can use it with another algorithm by choosing **Data > Export Alignment > MEGA format**. Name the data by the protein you used. Select "**Yes**" when asked if these are protein-coding data.

7. Now you will use another algorithm in MEGA: the nearest-neighbor joining algorithm for making a tree.

   - In the MEGA main window, choose **Phylogeny > Construct/Test Neighbor-Joining Tree**. Select your file.



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



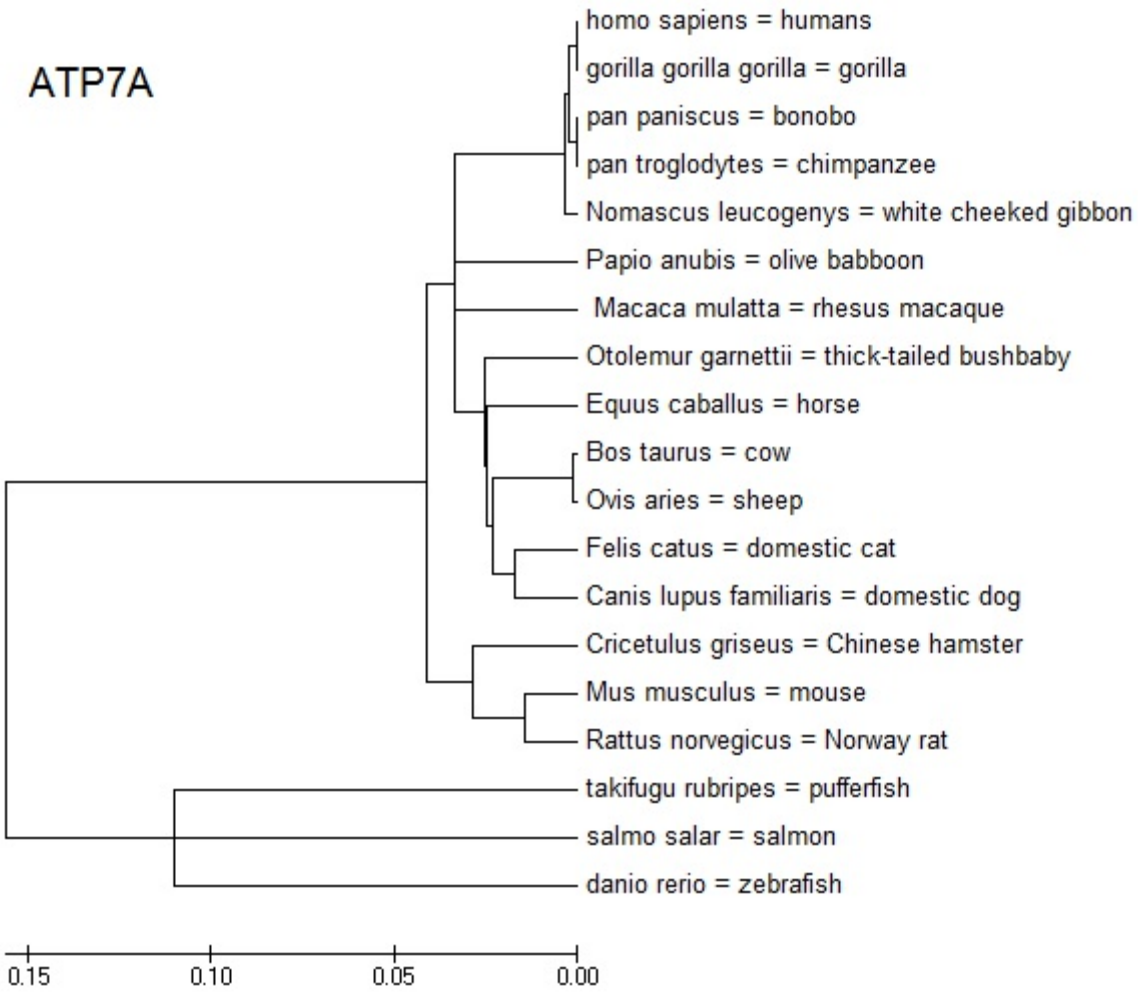
- The **Compute Linearized Tree** button can be used to toggle whether horizontal lengths are proportional to genetic change.

- Using the snipping tool and paint, the tree can be annotated with species' common names from the **Supplement D: List of Mammals**.

  Supplement D: List of Mammals

Conclusion

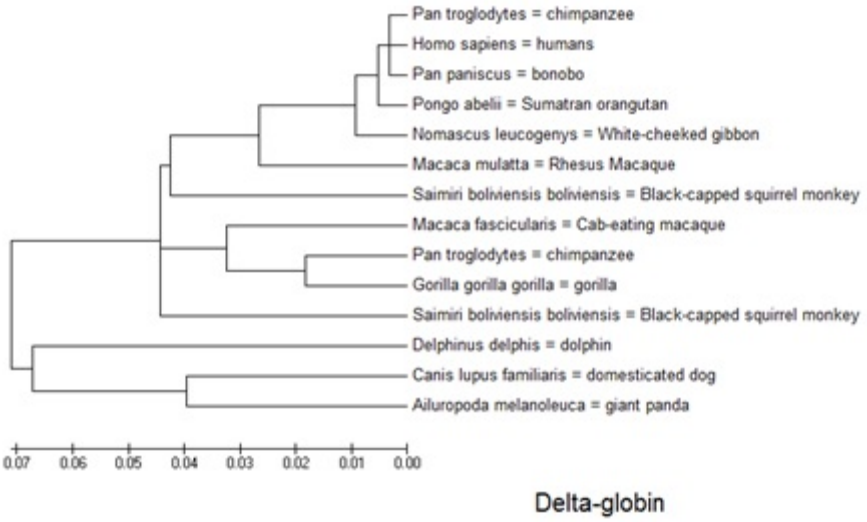
1. In the introduction of this project, it was said that we now had the power to perform billions of calculations on billion-byte datasets. Describe the dataset(s) used in this project, and describe the type of algorithms that required the billions of calculations.
2. Generalize the relationships shown 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.



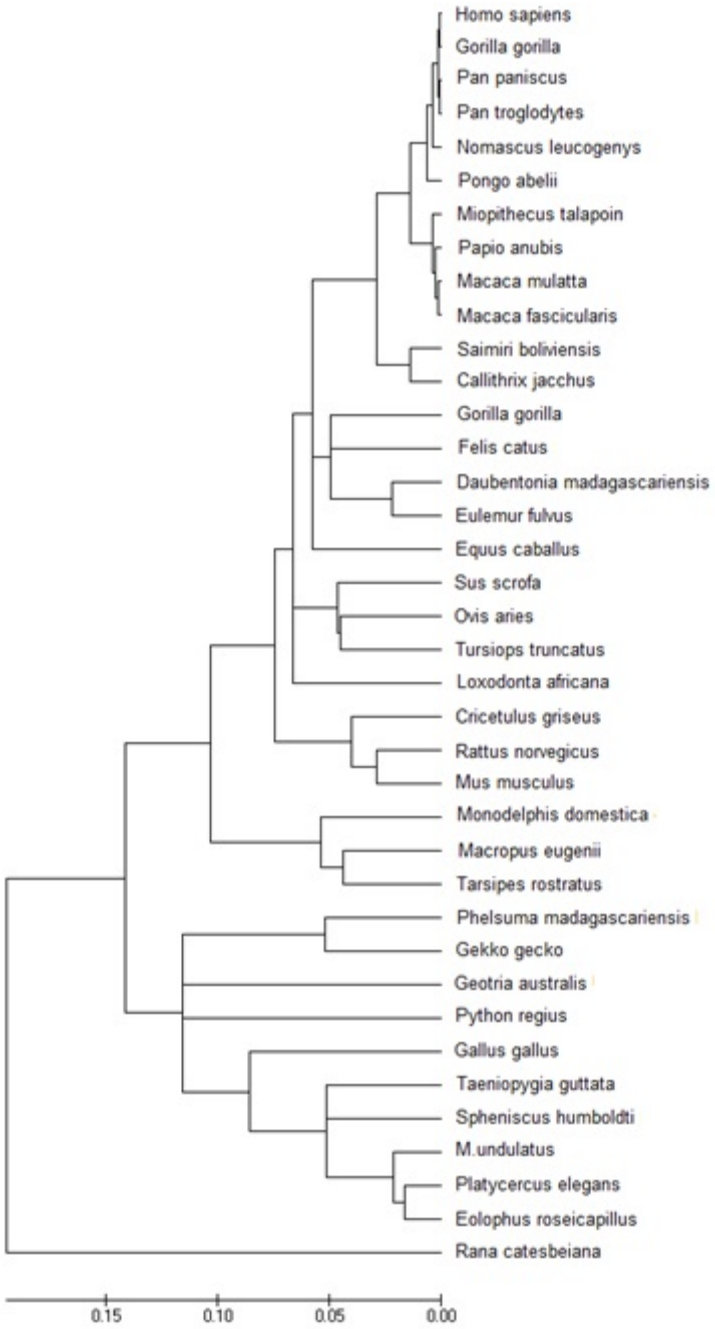
3. Exchange trees with one or more other groups. Explain your tree to the other groups, and have other groups explain their phylogenetic trees to you. What similarities and differences do you observe among the trees?
4. The tree shown here is for ATP7A, a protein used to transport copper in the body. The horizontal distance shown to the right of the division of cows and sheep is greater 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. Describe the meaning of the prefix “meta” in the statement that metadata creates new knowledge from the data at NCBI. What are the relationships here between information and knowledge?
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.



Delta-globin



Rhodopsin SW

Project 3.2.6 Supplement C: Biology Vocabulary

Term	Definition
Amino Acid	The building block of proteins, there are 20 distinct “standard” amino acids, meaning that they are encoded by DNA.
Complementary DNA	In double stranded DNA, the strands are complementary, pairing G-C and A-T. The nucleotides are read in opposite directions, so AAATGC is complementary to GCATTT.
DNA	Deoxyribonucleic acid is a long molecule created from a chain of DNA nucleotides.
Mutation	A mutation is a change in a DNA sequence. It can be caused by a chemical that damages the DNA, in which case the chemical is called a mutagen. A point mutation is a change in which a single nucleotide is changed to another nucleotide. Mutations also occur when nucleotide are inserted or deleted.
Nucleotide	A nucleotide is a molecule of roughly 40 atoms that is the building block of either RNA or DNA, depending on whether an oxygen atom is present in a particular location in the chemical structure. The single “letters” in DNA or RNA each stand for a specific nucleotide: G (guanine), C (cytosine), A (adenine), T (thymine), and U (uracil).
Phylogenetic Tree	A diagram grouping organisms into families based upon genetic similarity.
Protein	Large molecules that perform most actions in organisms, from digestion to muscle contraction, control of cell growth, and sensory perception. Proteins are made from one or more long chains of amino acids.
RNA	Ribonucleic Acid is a long molecule containing a sequence of RNA nucleotides (G, C, A, and U).
Transcription	DNA is transcribed to RNA in the nucleus.
Translation	RNA is translated to protein outside the nucleus at the ribosomes.

Project 3.2.6 Supplement B: List of Proteins

Introduction

In Project 3.2.6 you will pick a protein as a starting point for searching a database. There are approximately 20,000 proteins encoded by human DNA. Here are a few that might be interesting or familiar to you:

hemoglobin-alpha, hemoglobin-beta	subunits of hemoglobin, which carries oxygen in red blood cells
insulin-INS, somatostatin-SST	hormones released by the pancreas, important in diabetes and blood pressure
growth hormone-GH2, prolactin-PRL,	protein hormones secreted by pituitary gland
vasopressin-AVP, oxytocin OXT	protein hormones secreted by pituitary gland that control maternal behavior and pair bonding
Estrogen receptor-ER2	protein that responds to estrogen
atrial-naturetic peptide, atrial naturetic factor	hormones released by the heart
cholecystokinin, gastrin	hormones released by the gastrointestinal tract
allatostatin, proctolin	neuropeptides that regulate food intake and growth
leptin	hormone released by adipose (fat) tissue
CCAP	neuropeptide that regulates heart rate
Foxp2	protein regulating certain genes, associated with language development
prothrombin	important in forming blood clots
Galanin, Enkephalin, Neuropeptide Y	proteins associated with norepinephrine and noradrenaline
Somatostatin, Cholecystokinin	proteins associated with inhibitory neurotransmitter GABA
VIP, Substance P	proteins associated with neurotransmitter acetylcholine
	protein associated with neurotransmitter

neurotensin	dopamine
enkephalin	proteins associated with serotonin
dynorphin, cocaine-and amphetamine regulated transcript (CART)	proteins associated with oxytocin
collagen, elastin	structural proteins in cartilage
keratin	protein that forms hair, nails, hooves
actin, tubulin	create cytoskeleton
myosin, kinesin, dynein	motor proteins creating forces in muscles
alcohol dehydrogenase	Protein that metabolizes alcohol
rhodopsin medium wavelength	Senses green light in the eye

Project 3.2.6 Supplement D: Standard Genetic Code

In Project 3.2.5 you use DNA sequences that encode proteins. A simple understanding of biochemistry will help you understand the data in this lesson. Although simplified, the following explanation is a good start.

The material inside organisms falls mostly into four groups:

- Carbohydrates (sugars, starches, celluloses) – for short term energy storage
- Lipids (fats and oils) – for long-term energy storage
- Proteins – for doing everything from digestion to muscles to thinking
- DNA/RNA – for storing information about how to make proteins

The building blocks for proteins are 20 different types of amino acids, and these amino acids are strung together one after another when a protein is built. The instructions for building each particular protein is encoded in DNA in the cell nucleus. The instructions are transcribed from DNA into RNA, which then leaves the nucleus and travels to the ribosome where the instructions are used by translating the code into protein. The code of DNA/RNA nucleotides come in sets of three bases called a codon. Most of these codons are translated to an amino acid, but a few of the codons signal for the ribosome to let go of the growing protein, thus stopping translation.

All known life on earth shares essentially the same genetic code. All life uses the system in which three nucleotides are used to encode each amino acid. Which amino acids are encoded by which three-nucleotide sequences has only very slight variation across the kingdoms of life. The system below is the standard coding table, which describes the pattern by which RNA transcribed from DNA is translated to amino acids. Mitochondrial DNA uses a slightly different coding table.

In the table below, the amino acids are given by name, by their standard three-letter code, and by their standard single-letter symbol.

1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	(Phe/F) phenylalanine	UCU	(Ser/S) serine	UAU	(tyr/Y) tyrosine	UGU	(Cys/C) cysteine	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) leucine	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG		UGG	(Trp/W) tryptophan	G
C	CUU	(Leu/L) leucine	CCU	(Pro/P) proline	CAU	(His/H) histidine	CGU	(Arg/R) arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I) isoleucine	ACU	(Thr/T) threonine	AAU	(Asn/N) asparagine	AGU	(Ser/S) serine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) lysine	AGA	(Arg/R) arginine	A
	AUG	(Met/M) methionine	ACG		AAG		AGG		G
G	GUU	(Val/V) valine	GCU	(Ala/A) alanine	GAU	(Asp/D) aspartic acid	GGU	(Gly/G) glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) glutamic acid	GGA		A
	GUG		GCG		GAG		GGG		G

Project 3.2.6 Supplement A: List of Mammals

Introduction

How and why does our new power to perform huge numbers of computations on huge datasets 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 using computers to uncover the basis for life. So much data are now available that high school students can discover new things by creating new metadata that scientists have not yet had the chance to create.

You will access the genomes of the following mammals. Because the names are the *Genus species* name of the animal, you might not recognize them, so here's a list with common names. Where only one word is given, it is the genus.

- *Acipenser* = sturgeon
- *Bos taurus* = cow
- *Canis lupus* = wolf
- *Canis lupus familiaris* = dog
- *Capra* = goat
- *Cavia* = Guinea pig
- *Chinchilla* = chinchilla
- *Equus* = horse
- *Felis catus* = housecat
- *Gallus gallus* = chicken
- *Gorilla* = gorilla
- *Homo sapiens* = human
- *Lemur catta* = lemur
- *Loxodonta* = elephant
- *Mus musculus* = house mouse
- *Macaca mulatta* = macaque
- *Oryctolagus* = rabbit
- *Otolemur* = gelago
- *Ovis aries* = sheep
- *Pan paniscus* = bonobo
- *Pan troglodyte* = chimp
- *Panthera leo* = lion
- *Panthera tigris* = tiger
- *Papio* = baboon
- *Pongo* = orangutan
- *Rattus norvegicus* = Norway rat
- *Sus scrofa* = pig
- *Ursus* = bear

- **Non-mammals: ~1000 genomes sequenced include**
- *Danio* = zebra fish
- *Rana* = frog
- *Salmo* = salmon
- *Takifugu* = pufferfish

- *Xenopus* = African cloud frog

The tree relating similarities among these mammals varies, depending on which protein is used for the calculations. A consensus tree is shown below from Song, Liu, Edwards, and Wu. (2012). *Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model.* PNAS, 109:37, pp.14942-14947.