# Pre-Lab Questions Name: _____

1. Use the following data to construct a cladogram of the major plant groups.

**Table 1: Characteristics of Major Plant Groups**

| Organism | Vascular Tissue | Flowers | Seeds |
|---|---|---|---|
| Mosses | 0 | 0 | 0 |
| Pine trees | 1 | 0 | 1 |
| Flowering plants | 1 | 1 | 1 |
| Ferns | 1 | 0 | 0 |
| Total | 3 | 1 | 2 |

2. GAPDH (Glyceraldehyde 3-phosphate dehydrogenase) is an enzyme that catalyzes the sixth step in glycolysis, an important reaction that produces molecules used in cellular respiration. The following data table shows the percentage similarity of this gene and the protein it expresses in humans versus other species. For example, according to the table, the GAPDH gene in chimpanzees is 99.6% identical to the gene found in humans, while the protein is identical.

**Table 2. Percentage Similarity Between the GAPDH Gene and Protein in Humans and Other Species**

| Species | Gene Percentage Similarity | Protein Percentage Similarity |
|---|---|---|
| Chimpanzee (*Pan troglodytes*) | 99.6% | 100% |
| Dog (*Canis lupus familiaris*) | 91.3% | 95.2% |
| Fruit fly (*Drosophila melanogaster)* | (72.4% | 76.7% |
| Roundworm (*Caenorhabditis elegans*) | 68.2% | 74.3% |

a. Why is the percentage similarity in the gene always lower than the percentage similarity in the protein for each of the species?
b. In the space below, draw a cladogram depicting the evolutionary relationships among all five species (including humans) according to their percentage similarity in the GAPDH gene.

## Part I: Searching for Fossil Genes (adapted from David Form)

You are the manager of a new animal food supply company. You need to find out if vitamin C needs to be included in new animal foods designed for dogs, cows, cats, mice and guinea pigs. Based on your research on the GULO gene, you will be able to determine if you need to provide vitamin C in these foods.

Most mammals, such as mice, can produce their own vitamin C and therefore do not need a source of vitamin C in their diet. The GULO gene codes for an enzyme, L-gulonolactone oxidase, involved in vitamin C synthesis. The GULO gene is present in mice and most other mammals, but is either missing, or is nonfunctional, in some mammals. These animals cannot make their own vitamin C and must have this vitamin present in their diet. We will see if various mammals, including the primates, such as humans and chimps, have a functional GULO gene. A functional gene is able to produce a functional protein, in this case the GULO enzyme.

In some mammals, the GULO gene is an example of a pseudogene. Pseudogenes are vestigial genes. That is, they were once functional in an ancestral species, but since
they were no longer needed they accumulated mutations until they became non-functional. In many cases they evolve to the point where a protein can no longer be produced at all. Pseudogenes represent molecular evidence for evolution. As fossils are the remains of extinct organisms, pseudogenes are the remains of extinct genes.

**Procedure:**

The U. S. Government maintains a set of databases called GenBank, which contains nucleotide and amino acid sequences for those genes and proteins whose sequences have been determined. For your research we will use a computer program called **BLAST**. BLAST is able to search the GenBank databases. If you input a nucleotide

or amino acid sequence into BLAST, it will search for any known genes or proteins that are similar to the one that you entered. You will use BLAST to determine if the genomes of cows, pigs, humans and chimps contain functional GULO genes, or if they contain vestigial GULO pseudogenes, which do not result in production of the GULO enzyme.

Part A**. Is the GULO gene present in various mammals?**

1. Copy the nucleotide sequence below for the mouse GULO gene.

> mouse gulo gene CDS

ATGGTCCATGGGTACAAAGGGGTCCAGTTCCAAAACTGGGCGAAGACCTATGGCTGCAGTCCAG
AGATGTACTACCAGCCCACATCAGTGGGGGAGGTCAGAGAGGTGCTGGCCCTGGCCCGGCAGC
AGAACAAGAAAGTGAAGGTGGTGGGTGGCGGCCACTCGCCTTCAGACATCGCCTGCACCGATG
GCTTCATGATTCACATGGGCAAGATGAACCGGGTTCTCCAGGTGGACAAGGAGAAGAAGCAGG
TCACAGTGGAAGCCGGTATCCTCCTGACTGACCTGCACCCACAGCTGGACAAGCATGGCCTGGC
CCTGTCTAATCTGGGAGCCGTGTCTGATGTGACGGTTGGTGGCGTCATTGGGTCTGGAACACATA
ACACCGGGATCAAGCACGGTATCCTGGCCACCCAGGTGGTGGCCCTGACCCTGATGAAGGCTGT
GGAACAGTTCTGGAATGTTCTGAGTCAAGTAATGCAGATGTGTTCCAGGCTGCAAGGGTGCACC
TGGGCTGCCTGGGTGTTATCCTCACTGTCACCCTGCAGTGTGTGCCACAGTTCCACCTTCTGGAG
ACATCCTTTCCTTCGACCCTCAAGGAGGTCCTTGACAACCTGGACAGCCACCTGAAGAAGTCTG
AGTACTTCCGCTTCCTCTGGTTTCCTCACAGTGAGAACGTCAGCATCATCTACCAAGATCACACC
AACAAGGAGCCCTCCTCTGCATCTAACTGGTTTTGGGACTATGCCATTGGGTTCTACCTCCTGGA
ATTCTTGCTCTGGACCAGCACCTACCTGCCACGCCTCGTGGGCTGGATCAACCGCTTCTTCTTCT
GGCTGCTGTTCAACTGCAAGAAGGAGAGCAGCAACCTCAGCCACAAGATCTTCTCCTACGAGTG
TCGCTTCAAGCAGCATGTCCAAGACTGGGCCATCCCCAGGGAGAAGACCAAGGAGGCCCTGCTG
GAGCTAAAGGCCATGCTGGAGGCCCACCCCAAGGTGGTAGCCCACTACCCCGTGGAGGTGCGCT
TCACCCGAGGTGATGACATCCTGCTGAGCCCGTGCTTCCAGAGGGACAGCTGCTACATGAACAT
CATTATGTACAGGCCCTATGGGAAGGATGTGCCTCGGTTGGATTACTGGCTGGCCTATGAGACC
ATCATGAAGAAGTTTGGAGGCAGGCCCCACTGGGCAAAGGCCCACAATTGCACCAGGAAGGAC
TTTGAGAAAATGTACCCCGCCTTTCACAAGTTCTGTGACATCCGCGAGAAGCTGGACCCCACTG
GAATGTTCTTGAATTCGTACCTGGAAAAGGTTTTCTACTAA

2. On the internet, go to www.ncbi.nlm.nih.gov.
3. From the menu at the top of the page, select **BLAST**.
4. From the BLAST menu, select "**nucleotide BLAST"**. It is under "**Basic BLAST"**.
5. Copy the mouse GULO sequence into the box under "**Enter Query Sequence**".
6. Scroll down until you see "**Database**".
7. Check the "**Others**" box. Change the database setting in the box to "**nucleotide collection**".
8. In the "**Organism**" window, type in "**cow**". When you see cow appear in the box, select it.
9. Scroll to the bottom and find the "**BLAST"** button. Above the "**BLAST**" button, you will see the "**optimize for**" box. Select optimize for "**somewhat similar sequences (BLASTn)**"
10. Scroll back down and hit the "**BLAST**" button.
11. When **BLAST** is done with its search, you can scroll down and see a colorized diagram indicating the degree of similarity of the BLAST hits to your mouse GULO nucleotide sequence. Red and pink/purple mean a good match, while green, blue and black indicate a poor match. If the colored line spans the entire length of the window, then the "hit" sequence matches the inquiry sequence along its entire length. We want to see a high quality match along a majority of the inquiry sequence.

12. Below the colorized diagram is a "hit list" of your results. This shows the quality of matching as an E-value. An E-value is the chance that the matchup may be due to a random matching of a sequence of bases. The smaller the E-value, the more confidence you can have in your matching. A good match should have a low E- value (red or pink line) and an alignment along a large segment of the sequence.
13. Note your result in the chart below.
14. Now start again and do BLAST searches for pig (*Sus scrufa*), human (*Homo sapiens*), cow (Bos taurus), guinea pig (*Cavia porcellus*), and chimpanzee (*Pan troglodytes*) GULO genes. Record your data in the chart.

**GULO gene BLAST results**

| **Species** | **Present or Absent?** | **Query cover** | **% identity / E-value** |
|---|---|---|---|
| Mouse (*Mus musculus*) | Present | 100% | 100 / 0.0 |
| Dog (*Canis familiaris*) | | | |
| Cow (*Bos Taurus*) | | | |
| Pig (*Sus scrufa*) | | | |
| Guinea Pig (*Cavia porcellus*) | | | |
| Human (*Homo sapiens*) | | | |
| Chimp (*Pan troglodytes*) | | | |

**PART B. Does the human GULO gene produce a functional protein?**

We will now use protein **BLAST** to search for GULO proteins in cows, pigs, humans and chimpanzees.

1. Copy the mouse GULO protein below

> mouse GULO protein
MVHGYKGVQFQNWAKTYGCSPEMYYQPTSVGEVREVLALARQQNKKVKVVGGGHSPSDIACTDG
FMIHMGKMNRVLQVDKEKKQVTVEAGILLTDLHPQLDKHGLALSNLGAVSDVTVGGVIGSGTHNT
GIKHGILATQVVALTLMKADGTVLECSESSNADVFQAARVHLGCLGVILTVTLQCVPQFHLLETSFPS
TLKEVLDNLDSHLKKSEYFRFLWFPHSENVSIIYQDHTNKEPSSASNWFWDYAIGFYLLEFLLWTSTY
LPRLVGWINRFFFWLLFNCKKESSNLSHKIFSYECRFKQHVQDWAIPREKTKEALLELKAMLEAHPK
VVAHYPVEVRFTRGDDILLSPCFQRDSCYMNIIMYRPYGKDVPRLDYWLAYETIMKKFGGRPHWAK
AHNCTRKDFEKMYPAFHKFCDIREKLDPTGMFLNSYLEKVFY

2. Go to www.ncbi.nlm.nih.gov.
3. Select **BLAST** from the menu at the top of the page.
4. From the BLAST menu, select "**protein BLAST"**. It is under "**Basic BLAST"**.
5. Copy the mouse GULO sequence into the box under "**Enter Query Sequence**".
6. Scroll down until you see "**Organism**".
7. In the "**Organism"** window, type in "**cow**". When you see cow appear in the box, select it.
8. Scroll to the bottom and select the "**BLAST"** button.
9. When BLAST is done with its search, you can scroll down and see a chart of your results. Note your result in the chart below.
10. Now start again and do BLAST searches for pig (*Sus scrufa*), human (*Homo sapiens*), cow (Bos taurus), guinea pig (*Cavia porcellus*), and chimpanzee (*Pan troglodytes*) GULO genes. Record your data in the chart.
11. Look for proteins with the same name (, L-gulonolactone oxidase). If the GULO protein is not present, other, more distantly related proteins may come up. They will have a much lower score and a higher E-value. Note that the E-value represents the chance that the result is due a random matching of some amino acid sequences from both proteins. An E-value of 0 means a statistically perfect match. A good E-value should be much lower than $e^{-4}$.
12. Record your data in the chart.

**GULO Protein BLAST results**

| **Species** | **Present or Absent?** | **E value** | **Query cover** | **% identity** |
|---|---|---|---|---|
| Mouse<br><br>(Mus musculus) | Present | 0.0 | | |
| Dog<br><br>(*Bos taurus*) | | | | |
| Pig<br><br>(*Sus scrufa*) | | | | |
| Cow<br><br>(Bos Taurus) | | | | |
| Guinea Pig<br><br>(*Cavia porcellus*) | | | | |
| Human<br><br>(*Homo sapiens*) | | | | |
| Chimp<br>(Pan troglodytes) | | | | |

**Critical Thinking Exercises**

1. Why do you think that primates (monkeys, apes and humans) have lost the ability to produce vitamin C? (Hint: think about the diet of early primates).

2. Explain why the GULO gene in humans may be considered vestigial.

3. What can you infer about the GULO BLAST results between humans and chimps?

4. Your new pet food company is designing healthy foods for dogs, pigs, cows, mice and guinea pigs. From your results, to which types of feed will you suggest that the manufacturer add supplemental vitamin C? Justify your suggestion with a conclusive and specific result.

## Part 3: Search the databases and create your own phylogeny

Step 1. You will need to watch the pre-lab tutorial posted on iTunesU.
https://www.youtube.com/watch?v=iJGdhSqrTKI&list=HL1395263867
In this tutorial, I walk you through how to search a database (UniProt) and create a phylogenetic tree from selected protein sequences.

Step 2. Choose a protein (see suggestions in table below) that interests you or select one from the list provided below. Search for that protein name in UniProt.
www.uniprot.org

Step 3. Select a total of **15 different organisms** to use in the construction of your phylogeny. Copy and paste these protein sequences into a single document in FASTA format. It is crucial that the sequences are in **FASTA format**.

Step 4. Change all of the names for your 15 sequences to a simpler name that reflects only the genus and species or common name.

Step 5. Go to www.phylogeny.fr
and under "Phylogeny Analysis," select "one-click" analysis to create your tree. Feel free to name your analysis. Paste in all 15 sequences into the window and click "submit."

Step 6. Generate a pdf of your phylogeny and paste it into the space at the end of this document.

Step 7. Go to the T-coffee website: http://tcoffee.vital-it.ch/apps/tcoffee/do:regular
Paste your FASTA file of sequences into the "sequences to align" window and submit to generate a nice-looking alignment of the sequences. Take s screenshot(s) of the alignment and paste this into the bottom of this document as well.

Step 8. At the end of this document, write a one-page conclusion/interpretation of your phylogeny and sequence alignment that addresses the following questions:
-What does your phylogeny suggest about conserved core processes and what inferences can you make about the common ancestry of the organisms you selected to analyze?
-What did your initial search results suggest about how widely distributed this gene was within and across the domains of life?
-What can you tell about the degree of sequence conservation in the alignment that you generated in T-coffee? Does this alignment corroborate your tree results? Does anything stand out in the alignment as odd?
-What other data, either morphological, genetic, or both, could you add to your analyses that could improve or extend the phylogenetic tree?
-Elaborate on what else you learned or felt was most interesting about this lab.

**Suggested proteins to explore**

| | |
|---|---|
| Aquaporins | PITX1 |
| ATP synthase | MC1R |
| Catalase | Cytochrome C |
| Actin | Cytochrome B |
| GAPDH | Cellulose synthase (plants) |
| Keratin | Callose synthase (plants) |
| Myosin | Lipoxygenase |
| Pax1 | Superoxide dismutase |
| Hox genes | Peroxidase |
| Ubiquitin | rhodopsin |

Many others to choose from, so, be creative!

Name _____