# LAB 21 – Using Bioinformatics to Investigate Evolutionary Relationships; Have a BLAST!

## Introduction:

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 – 25,000 genes that define a human being. The project also successfully mapped the genomes of other species, including the fruit fly, mouse and *Escherichia coli*. The location and complete sequence of the genes in each of these species are available for anyone in the world to access via the Internet.

Why is this information important? Being able to identify the precise location and sequence of human genes will allow us to better understand genetic diseases. In addition, learning about the sequence of genes in other species helps us understand evolutionary relationships among organisms. Many of your genes are identical of similar to those found in other species.

Suppose you identify a single gene that is responsible for a particular disease in fruit flies. Is that same gene found in humans? Does it cause a similar disease? It would take you nearly 10 years to read through the entire human genome to try to locate the same sequence of bases as that in fruit flies. This definitely isn't practical, so a sophisticated technological method is needed.

**Bioinformatics** is a field that combines statistics, mathematical modeling, and computer science to analyze biological data. Using bioinformatics methods, entire genomes can be quickly compared in order to detect genetic similarities and differences. An extremely powerful bioinformatics tool is **BLAST**, which stands for **B**asic **L**ocal **A**lignment **S**earch **T**ool. Using BLAST, you can input a gene sequence of interest and search entire genomic libraries for identical or similar sequences in a matter of seconds.

In this investigation, you will use BLAST to compare several genes, and then use the information to construct a **cladogram**. A cladogram (also called a phylogenetic tree) is a visualization of the evolutionary relatedness of species. **Figure 1** to the right is a simple cladogram.

Note that the cladogram is treelike, with the endpoints of each branch representing a specific species. The closer the two species are located to each other, the more recently they share a common ancestor.

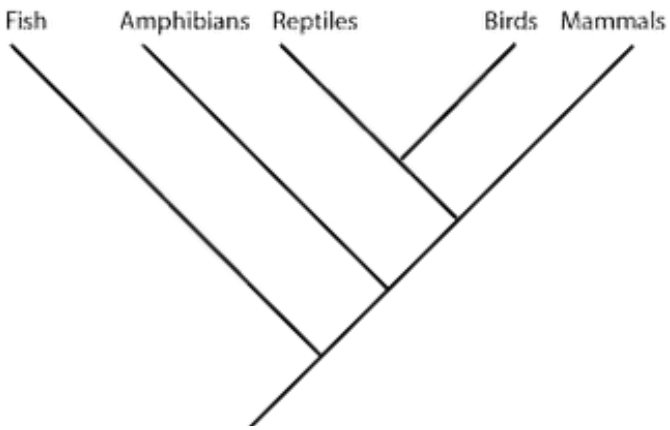



**Figure 1: A simple cladogram of the Phylum Chordata.**

The second cladogram in **Figure 2** below includes additional details, such as the evolution on particular physical structures called synapomorphies (shared derived characters/traits). Note that the placement of the derived traits corresponds to *when* (in general, not a specific, sense) *that character evolved*; every species above the character label possesses that structure. For example, mice and chimps have hair but salamanders and perch do not.

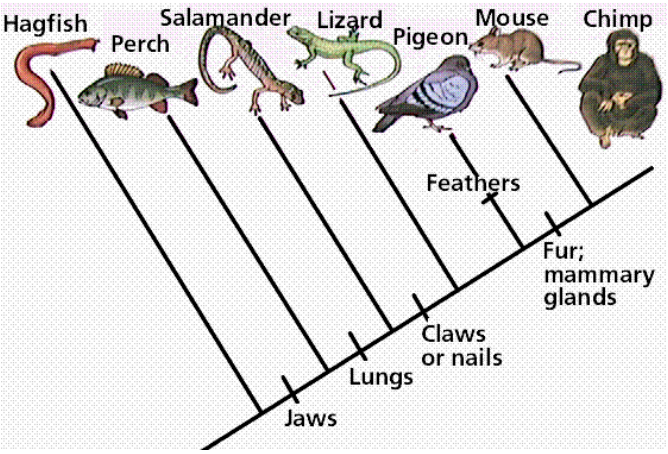



**Figure 2: A simple cladogram showing synapomorphies - shared dervied traits.**

Historically, only physical structures were used to create cladograms; however modern-day cladistics relies heavily on genetic evidence as well. Chimpanzees and humans share over 95% of their DNA, which would place them closely together on a cladogram. Humans and fruit flies share approximately 60% of their DNA, which would place them farther apart on a cladogram.

### Example 1:

Use the following data to construct a cladogram for the major plant groups below the table.

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

| Plant Group | Vascular Tissue | Flowers | Seeds |
|---|---|---|---|
| **mosses** | 0 | 0 | 0 |
| **pine trees** | + | 0 | + |
| **flowering plants** | + | + | + |
| **ferns** | + | 0 | 0 |

**Example 2:**
GAPDH (glyceraldehydes 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 primary sequence of the corresponding protein is identical.

**Table 2: Percentage Similarity of the GAPDH Gene and Protein with *Homo Sapiens***

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

**Questions:**

1) Why is the percentage similarity in the gene always lower than the percentage similarity in the protein for each of the species? (Hint: Recall how a gene is expressed to produce a protein.)

_____

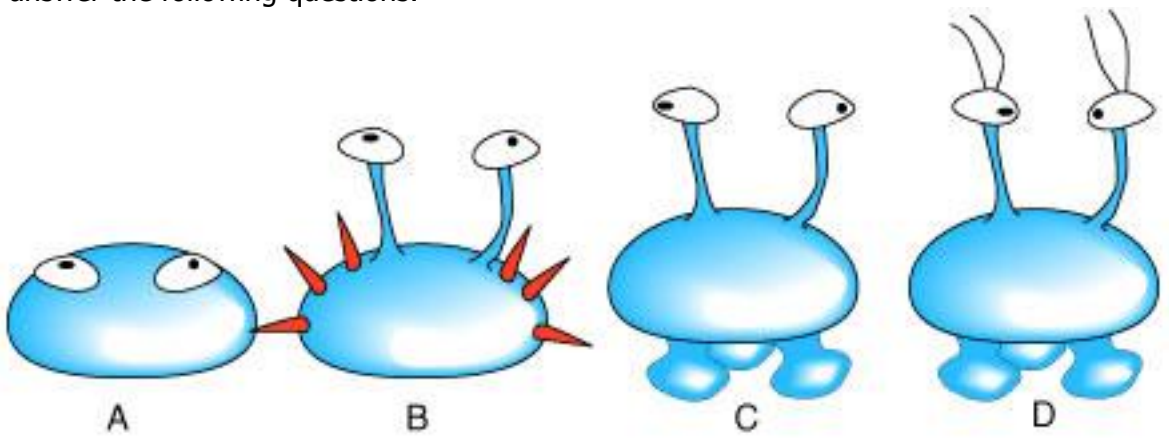
_____

_____

_____

2) Draw a cladogram depicting the evolutionary relationships among all five species (including Humans) according to their percentage similarity in the GAPDH gene.

3) The following four alien species were discovered, and it was realized that species "A" is very primitive – therefore it is the outgroup for the bunch. Use the pictures of each species to answer the following questions.




a. Fill out the following character matrix with a "O" for the ancestral trait (outgroup), and a "+" for a derived trait (different than outgroup).

| **Physical Characteristic** | **SPECIES** | | | |
|---|---|---|---|---|
| | **A** | **B** | **C** | **D** |
| eyes present | | | | |
| spines present | | | | |
| eyes on stalks | | | | |
| "feet" present | | | | |
| antennae present | | | | |

b. Generate a cladogram from the above matrix in the space below. **Label** the position of the synapomorphies.

## PART I – Using BLAST

A team of scientists has uncovered the fossil specimen in the photo to the right (**Figure 3)** near Liaoning Province, China. You should make some preliminary observations about this fossil based on its morphological features.

Little is known about the fossil. It appears to be a new species. Upon careful examination, small amounts of soft tissue have been discovered. Normally, soft tissue does not survive fossilization; however, rare situations of such preservation do occur. Scientists were able to extract DNA nucleotides from the tissue and use the information to sequence several genes.

You will use bioinformatic methods to support a hypothesis as to the evolutionary relationship between modern organisms and a fossilized organism using DNA collected from soft tissues. You will use BLAST to analyze the information from several genes and determine the most likely placement of the fossil species on the figure on the right.

**Figure 3: Newly discovered chordate speices found near Liaoning Province, China.**

## Procdure:

1. Form an initial hypothesis as to where you believe the fossil specimen should be place on the cladogram based on the morphological observations you made earlier. *Draw your hypothesis on the cladogram to the right* **(Figure 4)**.

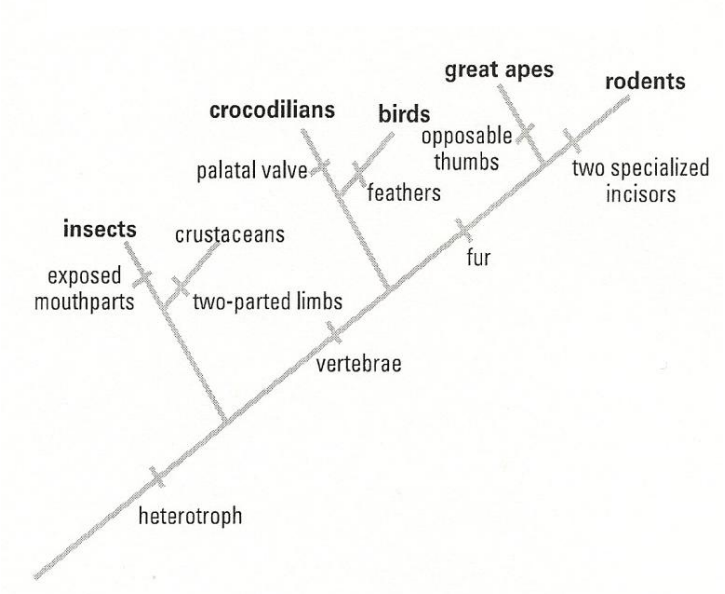



**Figure 4: Cladogram of some related groups to the newly discovered fossil found near Liaoning Province, China.**

2. Locate and download gene files 1, 3, and 4 obtained by sequencing the DNA collected from the soft tissue in the fossil. Download 3 files from:
   **http://blogging4biology.edublogs.org/2010/08/28/college-board-lab-files/**

   (When you download these files, rename them to "gene1", "gene3", and "gene4" to your account. You will not be able to view them but you will be able to upload them into the BLAST site. These might also be sent to you via email…)
3. Upload each gene sequence into BLAST by following the directions below. You will be doing this for each sequence *separately*.
   a. Go to the BLAST site: **http://blast.ncbi.nlm.nih.gov/Blast.cgi**
   b. Click on "Saved Strategies" from the menu at the **top** of the page. This will allow you to upload your first file.
   c. Under "Upload Search Strategy," click on "Browse" and locate your first file in your file directory.
   d. Click on "View."
   e. A screen will appear with the parameters for your query already configured. NOTE: Look but do not alter any of the parameters. Scroll down the page and click on the "BLAST" button at the bottom of the page.
   f. Collect and analyze the information from your first gene sequence (according to the instructions on the next page). Then do this with the other gene sequences.

## Analysis of Sequences – Graphic Summary

The chart below (**Figure 5**) is a graphical summary of your first sequence. The first line represents the most similar sequence that BLAST was able to pull from the data bases it searched. If you hover over the line, you will see the species of organism that this sequence was derived from. Other, less similar sequences are included from top (most similar) to bottom (less similar).

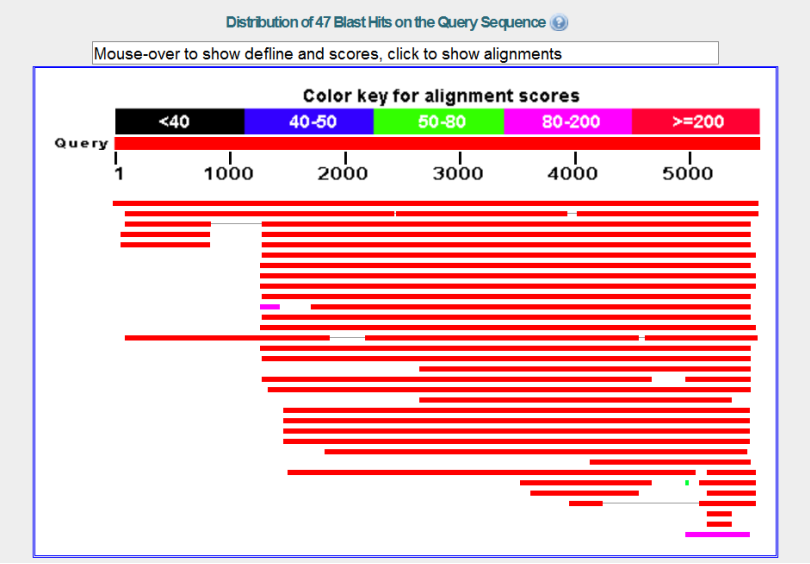



**Figure 5: Graphical summary from BLAST.**

## Analysis of Sequences – Descriptive Summary

**Table 3** below lists from top to bottom the information about each of the organisms that were represented in the graphic summary. The species in the list are those with sequences identical to or most similar to the gene of interest. The most similar sequences are first, and as you move down the list, the sequences become less similar to your gene of interest.

NOTE: 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 the cladogram.

**Table 3: Descriptive Summary from BLAST**

Descriptions

Legend for links to other resources: U UniGene E GEO G Gene S Structure M Map Viewer PubChem BioAssay

Sequences producing significant alignments:

| Accession | Description | Max score | Total score | Query coverage | E value | Max ident | Links |
|---|---|---|---|---|---|---|---|
| NM_204790.1 | Gallus gallus collagen, type V, alpha 1 (COL5A1), mRNA >gb\|AF13727 | 10288 | 10288 | 100% | 0.0 | 100% | U E G M |
| XR_118465.1 | PREDICTED: Meleagris gallopavo collagen alpha-1(V) chain-like (LOC1 | 3917 | 8848 | 96% | 0.0 | 97% | U G M |
| XM_003757475.1 | PREDICTED: Sarcophilus harrisii collagen, type V, alpha 1 (COL5A1), r | 3487 | 4077 | 89% | 0.0 | 82% | G |
| XM_001506246.2 | PREDICTED: Ornithorhynchus anatinus collagen, type V, alpha 1 (COL5 | 3476 | 3476 | 75% | 0.0 | 82% | G M |
| XM_001372383.2 | PREDICTED: Monodelphis domestica collagen, type V, alpha 1 (COL5A | 3465 | 3465 | 75% | 0.0 | 82% | G M |
| XM_003794863.1 | PREDICTED: Otolemur garnettii collagen, type V, alpha 1 (COL5A1), m | 3227 | 3227 | 76% | 0.0 | 80% | G |
| XR_131578.1 | PREDICTED: Equus caballus collagen alpha-1(V) chain-like (LOC10006 | 3221 | 3221 | 75% | 0.0 | 80% | G M |
| NM_000093.3 | Homo sapiens collagen, type V, alpha 1 (COL5A1), mRNA | 3177 | 3177 | 76% | 0.0 | 80% | U E G M |

***Max(imum) Score:*** the highest alignment score of a set of aligned segments from the same subject (database) sequence. This normally gives the same sorting order as the E Value. *The higher the max score, the closer the alignment.*

***E(xpect) Value:*** the number of alignments expected by chance with a particular score or better. It is sort of like a *control* for your hypothesis. *The lower the e value, the closer the alignment. Sequences with e values less than 1 e-04(1x10-4) can be considered related with an error rate of less than 0.01%.*

***Accession:*** If you click on the **Accession number** for a particular species listed, you will get a full report (**Figure 6**) that include the classification scheme of the species, the research journal in which the gene was first reported, and the sequence of bases that appear to align with your gene of interest. It will identify the gene, in this case it is **collagen**, that you are working with and you will also see the common name of the organism if it has one!

**Gallus gallus collagen, type V, alpha 1 (COL5A1), mRNA**

NCBI Reference Sequence: NM_204790.1

FASTA Graphics

Go to:

```
LOCUS       NM_204790               5575 bp    mRNA    linear   VRT 25-AUG-2012
DEFINITION  Gallus gallus collagen, type V, alpha 1 (COL5A1), mRNA.
ACCESSION   NM_204790 XM_419250
VERSION     NM_204790.1  GI:46048884
KEYWORDS    .
SOURCE      Gallus gallus (chicken)
  ORGANISM  Gallus gallus
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Archosauria; Dinosauria; Saurischia; Theropoda; Coelurosauria;
            Aves; Neognathae; Galliformes; Phasianidae; Phasianinae; Gallus.
REFERENCE   1  (bases 1 to 5575)
  AUTHORS   Gordon,M.K., Marchant,J.K., Foley,J.W., Igoe,F., Gibney,E.P.,
            Nah,H.D., Barembaum,M., Myers,J.C., Rodriguez,E., Dublet,B., van
            der Rest,M., Linsenmayer,T.F., Upholt,W.B. and Birk,D.E.
  TITLE     Complete primary structure of the chicken alpha1(V) collagen chain
  JOURNAL   Matrix Biol. 18 (5), 481-486 (1999)
```

**Figure 6: Summary report of a gene from BLAST.**

4. Scroll back up to the top of the analysis page. Right above the graphics section, find "Other reports". Click on "Distance tree of results" to see how this gene aligns with other species.

## Alignments

**Figure 7** shows the actual nucleotide comparisons between your unknown or "query" sequence and the most common sequence belonging to *Gallus gallus.* The results are tedious at best but there is a quick way for you to build a **cladogram** for the sequences. Click on "Distance tree of results" *above* the chart. You can click on the tabs above the tree and switch from rectangle, to slanted, to radial and force diagrams.

If you find that your tree seems a bit complicated, go back to BLAST, download the file a second time and go to "Optimize for" and select "Somewhat similar sequences." The resulting tree will be less complex. If you want to save a picture of your tree, *right click* on the tree and save it as a .png You can then insert it into a document.

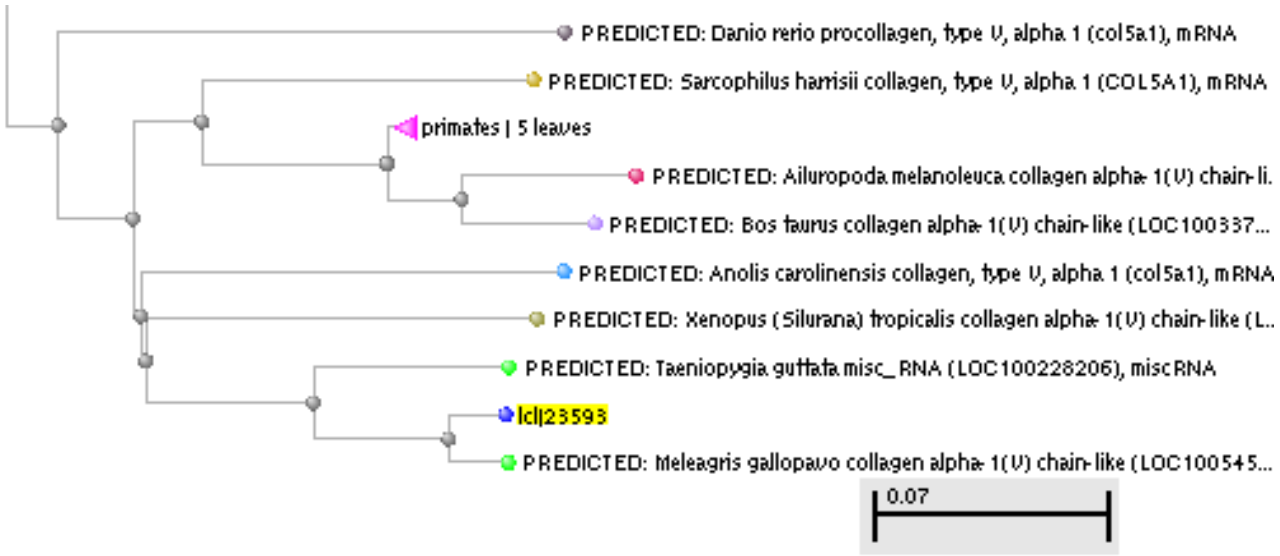



**Figure 7: Cladogram generated from BLAST showing how the unknown sequence aligns with a chosen group of species.**

## PART I Questions:

1) Once you have loaded each of your sequences into BLAST, and you have your results, indicate which species in the results has the most similar gene sequence to each gene of interest.

_____

_____

_____

_____

2) How similar is the gene sequence (for each gene)?

_____

_____

_____

3) Where is that species located on your cladogram? (Describe the nearby branches – for each gene.)

_____

_____

_____

_____

_____

_____

4) Did the analysis of each gene support or refute your original hypothesis? Did you have to redraw your original cladogram based on your results?

_____

_____

_____

_____

5) On the main page of BLAST, click on the link "List All Genomic Databases." How many genomes are currently available for making comparisons using BLAST? How does this limitation impact the proper analysis of the gene data used in this lab?

_____

_____

_____

_____

6) What other data could be collected from the fossil specimen to help properly identify its evolutionary history?

_____

_____

_____

_____

## PART II – Designing Your Own Investigation

Now that you have completed this investigation, you should feel more comfortable using BLAST. The next step is to learn how to find and BLAST your own genes of interest. If you select a human gene, BLAST will compare that gene sequence to any similar sequences in the databases. You will be researching the closest match to your human gene.
Hypothesize as to the species of organisms that might have the most similar genes to humans.

1. To locate a gene, you will go to the **Entrez** (IIPA: /ɑ̃tʁe/) http://www.ncbi.nlm.nih.gov/sites/gquery and select "gene".
2. Search for a gene that you know is present in humans. For example, you might use human actin, myosin, catalase, keratin or ubiquitin – or another one that you can pick that we mentioned in lecture/extra credit/etc. You can use gene names like Pax1, SRY1. Enzymes are also proteins so you might search "human ATP synthase". You can decide which gene you want to use. Type the name of the human gene in the search bar and click on "search."
3. In the previous assignment, you were given gene sequences to put into BLAST. In this assignment, you have used Entrez to find a segment you want to search for. The page you are on now, lists the results of your search. You should be able to select a reasonable sequence by clicking on the heading.
4. This next page has a lot of information about your gene but scroll *all the way* down to, "NCBI Reference Sequences (RefSeq)." One of the things you can do is to select a gene sequence FASTA file but, try something new. Select a file under "mRNA and Proteins." You can then select a sequence that codes for the mRNA. Click on one of the selections. In the next page, click on "FASTA".

NOTE: **FASTA format** is a text-based format for representing ***either nucleotide sequences or peptide sequences***, in which nucleotides or amino acids are represented using single-letter codes – ACGT for nucleotides; ACDEFGHIKLMNPQRSTVWY for amino acids.

5. You will see a sequence of DNA nucleotides complimentary to the mRNA for the protein you have selected. Copy this sequence (only the nucleotides), and return to BLAST http://blast.ncbi.nlm.nih.gov/Blast.cgi
6. Select "Nucleotide BLAST" and, instead of selecting "Saved strategies" just paste your sequences into the rectangle box labeled "Enter Query Sequences." You will be pasting a FASTA sequence into the box.
7. You can change some of the parameters of your analysis at this point. For example, under "Choose Search Set." You can select whether you want to search the human genome only, the mouse genome only, or all genomes available. Under "Program Selection," you can choose whether or not you want highly similar sequences or somewhat similar sequences. Choosing "Somewhat similar sequences" will provide you with more results but your tree will be much more complex.
8. Click on BLAST and you will get the analysis of your gene of interest.
9. Produce a cladogram based on your gene of interest, copy and paste it into a WORD file. Title the cladogram with the name of the gene of interest and print the page to be turned in with the rest of the questions.

## PART II Questions:

7) What is the *function* in humans of the protein produced from the gene you selected?

_____

_____

_____

_____

8) Would you expect to find the same protein in other organisms? If so, which ones – why? Which other organisms had gene sequences most similar to the human gene you selected?

_____

_____

_____

_____

9) Is it possible to find the same gene in two different kinds of organisms but not find the protein that is produced from that gene? Why might this happen?

_____

_____

_____

_____

10) If you found the same gene in **all** organisms you test, what does this suggest about the evolution of this gene in the history of life on earth?

_____

_____

_____

_____

_____

_____