Question: How can databases be used to investigate gene sequences and to construct cladograms that provide evidence for evolutionary relatedness among species?

BACKGROUND
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 *E. coli*. The complete location and sequence of the genes in each of these species is 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 and cure genetic diseases. Many of our genes are identical or similar to those found in other species, so learning about the sequence of genes in other species helps us to understand evolutionary relationships among all organisms.

Suppose you identify a single gene that is responsible for a particular disease in fruit flies. What if there is a similar gene in humans? Would the gene cause the same disease in humans? Without a sophisticated technological method, it would take you 9.5 years to read through the entire human genome to try and locate the same sequence of bases (1) as in the fruit fly. This definitely isn’t practical.

Bioinformatics is a field that combines statistics and computer science to analyze biological data. Using bioinformatics, entire genomes can be quickly compared in order to detect genetic similarities. The most powerful bioinformatics tool currently in use is **BLAST** which stands for **Basic Local Alignment Search Tool**. 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 laboratory investigation, you will use BLAST to analyze several genes and use the information to construct a cladogram. A **cladogram** (also called a phylogenetic tree) is a visualization of the evolutionary relatedness between species. The adjacent is a simple cladogram:

Note that the cladogram is tree-like, with the endpoints of each branch representing a specific species. The closer two species are located to each other, the more recently they share a common ancestor. For example, a spikemoss and a quillworth share a more recent common ancestor than a spikemoss and a clubmoss.

In a properly scaled cladogram, branch lengths are proportional to length of time and the intersection between two branches represents the common ancestor.
The following cladogram includes additional details such as the evolution of particular physical structures. These structures are referred to as **shared derived characters**. Note that the placement of the derived characters corresponds with when that character evolved and that every species above the character label possesses that structure. For example, tigers and gorillas have hair, but lampreys, sharks, salamanders, and lizards do not have hair.

---

**COMPLEX CLADOGRAM**

What organisms have lungs?

What three structures do lizards possess?

According to the cladogram, which structure evolved first, dry skin or hair?

Historically, physical structures were used to create cladograms, however, modern day cladistics relies more heavily on genetic evidence.

---

Chimpanzees and humans share 96% (?) of their DNA which would place them close together on a cladogram. Humans and fruit flies share approximately 60% of their DNA which would place them farther apart on a cladogram.

In the space provided, draw a complex cladogram that depicts the evolutionary relationship between humans, chimpanzees, and fruit flies.
LEARNING OBJECTIVES

- To create cladograms that depict evolutionary relationships among organisms
- To analyze biological data with a sophisticated bioinformatics online tool
- To connect and apply concepts pertaining to genetics and evolution

PART 1: GETTING READY

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

<table>
<thead>
<tr>
<th>Organisms</th>
<th>Vascular Tissue</th>
<th>Flowers</th>
<th>Seeds</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mosses</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pine trees</td>
<td>+</td>
<td>+</td>
<td>0</td>
</tr>
<tr>
<td>Flowering plants</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Ferns</td>
<td>+</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

2. GAPDH (glyceraldehyde 3-phosphate dehydrogenase) is an enzyme that catalyzes the sixth step in glycolysis, an important reaction in the process of cellular respiration. The following data table shows the % 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.

<table>
<thead>
<tr>
<th>Species</th>
<th>Gene</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chimpanzee</td>
<td>99.6%</td>
<td>100%</td>
</tr>
<tr>
<td>Dog</td>
<td>91.3%</td>
<td>95.2%</td>
</tr>
<tr>
<td>Fruit fly</td>
<td>72.4%</td>
<td>76.7%</td>
</tr>
<tr>
<td>C. elegans (a roundworm)</td>
<td>68.2%</td>
<td>74.3%</td>
</tr>
</tbody>
</table>

a) Why is the % similarity in the protein always higher than the % similarity in the gene for each of the species? (hint: recall how a gene is expressed to produce a protein)

__________________________________________________________________________________________

b) Draw a cladogram depicting the evolutionary relationships between all four species according to their % similarity in the GAPDH gene in the box above.

You can also prepare for the lab by working through the following online activities:

“The Evolution of Flight in Birds” (provides a real-world example of how cladograms are used to understand evolutionary relationships): http://www.ucmp.berkeley.edu/education/explorations/reslab/flight/main.htm


PART 2: THE INVESTIGATION
A team of scientists have uncovered the following fossil specimen near Hells Creek, Montana. Make some general observations about the morphology (physical structure) of the fossil and record your observations in the space provided.

**FOSSIL SPECIMEN**

Little is known about the fossil other than it appears to be a new species. Upon careful examination of the fossil, small amounts of soft tissue have been discovered. Normally, soft tissue does not survive fossilization, however, on rare occasions, the proper conditions exist so that these tissues remain remarkably intact. The scientists were able to extract proteins from the tissue and use the information to sequence several genes. Your task is to use BLAST to analyze these genes and determine the most likely placement of the fossil species on the following cladogram:

**FOSSIL CLADOGRAM**

Form an initial hypothesis as to where you believe the specimen should be placed on the cladogram based on the morphological observations you made earlier.

**DRAW** your hypothesis on the cladogram.
STEP 1: LOCATING AND DOWNLOADING GENE FILES
There are four gene files (Gene 1 - Gene 4) for you to download to your flash drive. These files are located online: You will need to save them from home as they are blocked by DCSS.

Note that these files will not open on your computer. They only work when opened on the BLAST website.

STEP 2: INSTRUCTIONS FOR BLAST QUERIES
Upload the gene sequences into BLAST by following these directions:
2. Click on “saved strategies” from the menu on the top of the page.
3. Click “file upload” and locate one of the gene files you saved onto your computer.
4. Click “view”
5. A screen will appear with the parameters for your query already configured. NOTE: Do not alter any of the parameters. Scroll down the page and click on the BLAST button at the bottom.
6. After collecting and analyzing all of the data for that particular gene (see instructions below), repeat this procedure for the other three gene sequences.

STEP 3: INSTRUCTIONS FOR ANALYZING BLAST QUERIES
1. Scroll down to the section titled, “Sequences producing significant alignments”. The list of organisms that appear below this section are those with sequences identical to or most similar to the gene of interest. The most similar sequences are listed first and as you move down the list, the sequences become less similar to your gene of interest.

2. If you click on a particular species listed, you’ll get a full report that includes the species’ classification scheme, the research journal the gene was first reported in, and the sequence of bases that appear to align with your gene of interest.

3. If you click on the link titled “Distance tree of results,” you will see a cladogram of the species with similar sequences to your gene of interest placed on the cladogram according to how closely their matched gene aligns with your gene of interest.

STEP 4: ANALYSIS OF RESULTS
Recall that species share similar genes because of common ancestry. 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 a cladogram.

As you collect information from BLAST for each of the gene files explain whether the data supports or refutes your original hypothesis and whether the data supports or rejects your original placement of the fossil species on the cladogram. __________________________________________________________

______________________________________________________________________________

______________________________________________________________________________
For each BLAST query, consider the following:
What species has the most similar gene sequence as the gene of interest? __________________________

_________________________________________________________________________________________

Where is that species located on the cladogram? ________________________________________________

_________________________________________________________________________________________

How similar is that gene sequence? ___________________________________________________________

_________________________________________________________________________________________

What species has the least similar gene sequence as the gene interest? __________________________

_________________________________________________________________________________________

STEP 5: DRAWING THE CLADOGRAM
Redraw the original cladogram and include your final placement of the fossil species.

PART 3: EVALUATION OF RESULTS
Discuss with your classmates the position each person chose to place the fossil specimen on the cladogram. Does everyone agree? If not, discuss the evidence collected and debate the proper placement on the cladogram.

_________________________________________________________________________________________

__________________________________________________________

What assumptions about the quality of the sample must be made in order to have confidence in the gene analysis?

_________________________________________________________________________________________

6
On the main page of BLAST click on the link “list all genomic databases.” Note that there are currently only around 1500 genomes sequenced thus far (as of 2010). How does this limitation impact the proper analysis of the gene data used in this lab?

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

PART 4: WHERE CAN YOU GO FROM HERE?
BLAST your own genes of interest!
Now that you’ve completed the first part of the investigation, you should feel more comfortable using BLAST. The next step is to learn how to find and BLAST your own genes of interest. To locate a gene, go to the Entrez Gene website (http://www.ncbi.nlm.nih.gov/gene) and search for the gene. Once you have found the gene on the site, copy the gene sequence and input it into a BLAST query. (HINT: Why not BLAST your DNA to Disorder Project gene)

Here is an example of how to do this:
1. On the Entrez Gene website search “human actin”
2. Click on the first link that appears and scroll down to the section “NCBI Reference Sequences”
3. Under “mRNA and Proteins” click on the first file name “NM 001100.3”
4. Just below the gene title click on “FASTA”
5. The nucleotide sequence displayed is that of the actin gene in humans.
7. Click on “nucleotide blast” under the Basic BLAST menu.
8. Paste the sequence into the box where it says “Enter Query Sequence.”
9. Give the query a title in the box provided if you plan on saving it for later.
10. Under “Choose Search Set” select whether you want to search the human genome only, mouse genome only, or all genomes available.
11. Under “Program Selection” choose whether or not you want highly similar sequences or somewhat similar sequences. Choosing somewhat similar sequences will provide you with more results.
12. Click BLAST.
13. In humans, what is the importance of the gene you chose? Would you expect to find that gene in all organisms? Why or why not?

Some gene suggestions you could try out:

<table>
<thead>
<tr>
<th>Actin</th>
<th>ATP synthase</th>
<th>Catalase</th>
<th>GAPDH</th>
<th>Keratin</th>
</tr>
</thead>
<tbody>
<tr>
<td>Myosin</td>
<td>Pax I</td>
<td>Ubiquitin</td>
<td>Zinc finger</td>
<td></td>
</tr>
</tbody>
</table>