PHYLOGENY LAB/ACTIVITY with BLAST

PART 1: Searching for Fossil Genes

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, meaning 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 US 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, which is capable of searching the GenBank databases. If you enter a nucleotide or amino acid sequence into BLAST, it will search for any known genes or proteins that are similar to the one 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 produce a functional GULO enzyme.

1. Go to https://blast.ncbi.nlm.nih.gov/Blast.cgi
2. Select Nucleotide Blast.
3. Copy the Mouse GULO nucleotide sequence from the file provided by your teacher into the box under "Enter Query Sequence".
4. Give your search a "Job Title" – like your last name and a number.
5. For the "Database" select the **Others** button, and choose **Nucleotide Collection (nr/nt).**
6. In the "Organism" box, type cow, or the taxonomic i.d. 9913.
7. Where it says "Optimize For" select the **Somewhat similar sequences (blastn)** button.
8. Double check all your settings as listed above, and then click on the BLAST button at the bottom.

READING YOUR RESULTS:

What you are seeing is the extent to which the nucleotide sequence for the MOUSE GULO GENE matches the nucleotide sequence of the COW GULO GENE. Red and pink 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.

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 or likelihood that the sequences matched up the way they did due to randomness. The *smaller* the E-value,

the more confidence you can have in your match. A good match should have a low E-value (red or pink line) and an alignment along nearly all of the sequence.

ENTERING YOUR RESULTS:

Complete your chart with BLAST results from a dog 9615, cow 9913, pig 9823, guinea pig 10140, human 9606 and chimp 9569.

PART 2: Does the human GULO gene produce a functional protein?

PROCEDURE:

1. Go to https://blast.ncbi.nlm.nih.gov/Blast.cgi
2. Select Protein Blast.
3. Copy the Mouse GULO nucleotide sequence from the file provided by your teacher into the box under "Enter Query Sequence".
4. Give your search a "Job Title" – like your last name and a number.
5. For the "Database" choose **Non-redundant protein sequences (nr).**
6. In the "Organism" box, type cow, or the taxonomic i.d. 9913.
7. The Algorithm should be **blastp** (protein-protein BLAST)

READING YOUR RESULTS:

When the search is complete, scroll down to view your results, both graphically, and numerically. 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 to random matching of some amino acid sequences from both proteins. An E-value of 0 means a statistically perfect match. There's no real consensus, but a good E-value is somewhere near 1×10^{-4}.

ENTERING YOUR RESULTS:

Complete your chart with BLAST results from a dog 9615, cow 9913, pig 9823, guinea pig 10140, human 9606 and chimp 9569.

CRITICAL THINKING QUESTIONS (to be answered in your LAB BOOK):

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 animal 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 to create your own PHYLOGENY

PROCEDURE:

1. Create a word-processing document called "Phylogeny by {your name}."
2. Choose a protein to explore from the list, or another protein in which you have a keen interest. The teacher can help you with some choices.
3. Go to www.uniprot.org, and type your protein of choice into the search bar at the top of the page.
4. In the "Entry" column, select the alpha-numeric code of the protein you want, and then click the FORMAT button at the top of the next screen. **Choose FASTA format.**
5. Copy that protein's sequence, including the title line, into your word-processing document.
6. From your initial protein search (done in #3 above), select **15 different organisms** to use in the construction of your phylogeny. Copy and paste each organism's FASTA protein sequence into your document, including the title line. **It is crucial that the sequences are all in FASTA format.**
7. Go to www.phylogeny.fr and choose "1-click" analysis.
8. Give your Phylogeny a name of your choice.
9. Copy and paste your ENTIRE collection of 15 organisms' protein sequences into the sequence box. *Try shortening the title lines of your sequences down to a simpler name that reflects only the genus and species, or common name.* Then click SUBMIT.
10. Choose the "Tree Rendering" tab above, which gives the same graphical results as the "Phylogeny" tab. Put your tree rendering into a pdf, and copy this image into your word-processing document.
11. Go to http://tcoffee.vital-it.ch/apps/tcoffee/do:regular and paste your 15 organisms' protein sequences into the sequence box provided and then click SUBMIT.
12. Copy and paste this sequence alignment into your word-processing document.

CRITICAL THINKING QUESTIONS (to be answered in your LAB BOOK):

1. What does your phylogeny suggest about conserved core processes and common ancestry of the organisms you selected?
2. What did your initial search results suggest about how widely distributed this gene was within and across the domains of life?
3. What can you tell about the degree of sequence conservation in the alignment that you generated in T-coffee? Does the alignment corroborate your tree results? Does anything appear out of sorts in your alignment?
4. What other data, either morphological, genetic, or both, could you add to your analysis that could improve or extend the phylogenetic tree?