Elaborate: Bioinformatics

BACKGROUND

A molecular biologist or geneticist that has isolated a gene for an unknown protein needs to "characterize" or figure out what the gene does. One way to do this is to compare the newly discovered gene to genes in other organisms to see if there is a very similar gene that's already been studied. If there are closely related genes to be found in the existing gene databanks, it's a very good possibility that the unknown protein coded for by the gene has a similar function to the proteins coded by the other similar genes.

In this activity, a gene for an unknown protein in corn will be used to search for related genes in other plants, and then by looking at the functions of those genes in the other plants, it may be possible to figure out something about the function of the gene in the corn plant. In addition, a table and/or phylogenetic tree can be created to show degrees of relation between the genes that are examined depending on how similar the DNA sequences are for these related genes.

PROCESS AND PROCEDURES

1. In this activity, you will be using two web sites simultaneously. The first site (called the **Biology Workbench**) has a bunch of tools that let you look at and compare the genes of many different organisms. The second site (The National Center for Biotechnology Information) is a database run by the federal government with the codes for many different genes from many different organisms.

   Open a browser and go to the first site at this URL:
   http://workbench.sdsc.edu

   This is the **Biology Workbench**, and it will allow you to access a number of Bioinformatics tools from all over the web. If this is your first time to the **Biology Workbench**, create a new account by following the instructions for first time users.

2. Once you've entered the **Biology Workbench**, scroll down and click on the **[Session Tools]** button.

3. Highlight "**Start New Session**" by clicking on it with your mouse, then click **[Run]**.

4. In the **Session Description** box, enter: *Unknown Corn Protein* or some other meaningful label to identify this activity, then click **[Start New Session]**.

5. Click on the **[Nucleic Tools]** button.

6. You're going to leave the **Biology Workbench** temporarily and open a second browser window or tab to access the following site:
   http://www.ncbi.nlm.nih.gov
   This is the National Center for Biotechnology Information.

7. Make sure the drop down menu to the left of the **Search** box says "**Nucleotide**". If not, click on the down arrow and choose "**Nucleotide**". In the search box you can enter the name of a particular protein, the name of a species, or even words like "unknown" or "uncharacterized". You could also enter a combination of these words or terms. Or you can enter the "accession number" for a particular gene. When people discover new genes, they enter the codes into a huge database, and they're given an "accession number" by the government. It's sort of like the Dewey decimal system for books in the library. Our "unknown" corn protein has the accession number: **AF152600**. Enter this number into the **Search** box and then press the **[Search]** button. You will see a page come up with "*Zea mays*" at the top. *Zea mays* is the scientific name for corn.

8. ***Optional:*** To see more information on a gene, click on "Display Settings" at the upper left of the screen. For example, if you choose "**Genbank**", then click the **[Apply]** button, you can see the DNA code for the gene and info such as the journal the gene was first published in.

9. Either click on "Display Settings" again, choose "**FASTA**" and click the **[Apply]** button, or just click on the "**FASTA**" shortcut under the species name top left.

10. To import the sequence into the **Biology Workbench**, click and drag the mouse to highlight all the information beginning with the "**>**" sign and including all of the DNA bases listed. **Copy** the highlighted section.

11. Return to the **Workbench** window. Click on "**Add New Nucleic Sequence**", the click **[Run]**.

12. In the "**Label:** " box (***not*** the box by the **[Browse...]** button), type something that will describe the sequence and help differentiate between the different sequences. (Ex: *Corn unknown mRNA*). Then **Paste** the copied FASTA sequence into the "**Sequence:** " box. Delete the FASTA identification line (all the info just after the "**>**" sign on the first line – be careful not to delete any of the DNA data or the "**>**" sign) and replace it with the same label you just typed in the label box. Then click **[Save]**. (You may have to scroll up or down to see the **[Save]** button.)

13. Click on the box to the left of the gene sequence that was just entered. A check mark should appear. Scroll through the Nucleic Tools listed in the box and select "**BLASTN - Compare a NS to a NS DB**" and click **[Run]**.

    Where it says "**Choose from 1 up to 16 Databases:**" scroll down within the window and highlight **GenBank Plant Sequences** (all 3 parts), then scroll the entire page down and click the **[Submit]** button. (You will have to use shift-click or control-click to highlight all three Plant Sequences.) Do not worry about changing ***any*** other settings on this page.

    This will give you a list of genes within this database that match your gene to varying degrees. The closest matches (with the highest scores and lowest "e" values) are found at the top. As the list continues, the genes are less and less similar to the gene you're checking. Along with the sequences and information on the genes, you will also find the accession numbers. Rather than weeding through all the genes listed there, we'll use a sample table of accession numbers found previously by doing BLASTN searches for the unknown corn protein that we will use for our lab.

14. Click the **[Return]** button (you may have to scroll up or down to find it), and then go back to the NCBI site and type in the an accession numbers (starting with the first one below) from the table below in the **Search** box and press the **[Search]** button. Repeat steps **8** through **12** for each number listed below until all the genes have been entered and saved.

BLASTN homologous genes		
Accession #	**Plant**	**Scientific Name**
AB037106	Satsuma Orange(/Mandarin)	*Citrus unshiu*
AF009338	Cotton	*Gossypium hirsutum*
X96785	Spinach	*Spinacia oleracea*
X92117	Mouse-Ear Cress	*Arabidopsis thaliana*
X92118	Iceplant (common)	*Mesembryathemum crystallinum*
U84268	Barley	*Hordeum vulgare*
AF165939	Lemon	*Citrus limon*

You're now going to compare these sequences for similarity as well as create a phylogenetic tree using the data you've entered.

15. Click on the box to the left of each gene that was entered. Scroll down in the Nucleic Tools box, select "**CLUSTALW - Multiple Sequence Alignment**" and click **[Run]**. After that tool has run and the screen has changed, click **[Submit]**. When the page is done loading, you can scroll down and see how much of the DNA from your different samples match each other.

16. After you're sure the previous job has finished loading, click the **[Import Alignments]** button. This takes you out of the Nucleic Tools section and activates the **[Alignment Tools]** section of **Biology Workbench**. Check the box next to CLUSTALW and the genes that have been entered. Scroll down in the Alignment Tools box to "**CLUSTALDIST - Generate Distance Matrix with Clustal W**" and click **[Run]**. Then click **[Submit]**. Scroll down to the Clustal Distance Matrix. The closer a number is to zero, the more similar the DNA from those two organisms are and possibly the more closely related to each other. Copy this table into your notebook and answer Analysis question #3.

17. After the matrix data has been copied into your notebook, click on the **[Return]** button. Scroll down in the Alignment Tools box and select "**DRAWGRAM - Draw Rooted Phylogenetic Tree from Alignment**" and click **[Run]**. Then click **[Submit]**. Scroll down to see the phylogenetic tree.

18. While the phylogenetic tree and matrix are great tools for comparing how closely related genes from two organisms are, it still doesn't tell us what those genes do in those other organisms. For this, we need to go back and look at the actual journal articles for one (or all) of those other genes we examined.

There is a very nice characterization of this gene given for the lemon (Moshe, 2000). The sequence codes for the gene of one of the proteins involved in pumping acid into vacuoles in the cells. In citrus fruits, this has been tied to the amount of citric acid concentration in the juice – or to put it simply: how sour the fruit tastes.

ANALYSIS

Answer the following questions in your notebook using complete sentences or well labeled diagrams.

1. Create a phylogenetic tree showing how closely related you think the following animals are: dog, cat, wolf, squirrel, butterfly, moth, ant, spider.

2. Using the images shown in class, draw a tree showing how you think the plants in the images might be related to each other (spinach, lemon, barley, satsuma orange, cotton, corn, mouse ear cress, iceplant).

3. Using the matrix/table generated in step 16 in the Process/Procedures section, redraw your phylogenetic tree to show how closely related the different plants are.

4. How is the computer deciding how closely related different organisms are? What is it comparing?

5. Compare the final tree drawn by the computer in step 17 of the Process/Procedures with your trees drawn above. Which plants were more closely related than you thougt?

6. The article on the lemon says this gene helps citrus plants taste sour. Do you think that is what the gene is doing in corn plants?

7. Are the two citrus fruits the most closely related plants to corn on the phylogenetic tree? If you wanted to look at an article describing what the gene does in a more closely related plant to the corn, which plant would you pick?