EXERCISE 10 name _____

Using Biological Databases on the Internet

Objectives

After completing this exercise, you should be able to:

- Locate biological databases of DNA and protein sequences, as well as scientific publications on the Internet.
- Retrieve and compare sequence information from databases.
- Compare evolutionary relatedness and draw phylogenetic trees from sequence comparisons.
- Study protein structure and function from database comparisons.

Prelab

Before you come to lab, read this entire exercise.

One of the goals of the Human Genome Project is to sequence the entire human **genome**: all 22 somatic chromosomes, along with an X and a Y chromosomes. Since there are 3.2 billion base pairs in the human genome, this is a daunting goal. Sequencing began in 1990 and by February, 2001, the sequencing was over 90% complete. This milestone came 4 years earlier than the target date. The only sequences not covered at thatpoint were highly repetitive regions such as the centromere and telomere regions of the chromosomes, regions that resist cloning by the host and vector systems that are currently used. Another goal of the Human Genome Project is to sequence the genomes of other species of interest, such as model organisms used by biologists, pathogens of medical importance, and crop plants of agricultural importance. This goal has also been exceeded, and today the genomes of over 900 species have been at least partially sequenced.

This large international project has not only collected enormous databanks of DNA sequence information from the genomes of dozens of species, it has also promoted the development of highly automated strategies for studying nucleic acids and proteins. Much of the reason for the success of the Human Genome Project comes from the introduction of new **"high throughput"** technologies for DNA sequencing that can use automation and robotics to complete.

The Human Genome Project has been a catalyst for change in the way biologists approach the study of living things. Biologists today using the sophisticated laboratory technology for sequencing DNA are collecting data faster than they can interpret it. A new field called **bioinformatics** is developing for the storage and management of the data stored in these rapidly growing databases, as well as for the use of a computer as a general tool for discovering how living things work.

For example of how this can work, when a scientist sequences a new stretch of DNA, this information becomes meaningful when a new gene can be discovered in the sequence. Often these genes are hard to spot, especially in eukaryotic genomes, since the majority of DNA in these organisms does not code for genes---in the human genome genes account for less than 2% of the DNA sequence. A computer program can search the sequence for tell-tale signs of a gene from sequence data, by searching for DNA consensus sequences for translational start and stop codons, a ribosome binding site, intron splicing sites, and a promoter. In eukaryotes, a region of high guanosine and cytosine (G and C) content is frequently found near clusters of genes, so mapping GC content along a chromosome can also help to locate the presence of a gene in the DNA sequence. Since gene sequences can be highly conserved between different species, an especially powerful approach for identification of a gene in sequence data is to search databases of DNA sequences looking for **sequence similarities**. Due to these so-called **"in silico" (**computer program), tools and the dramatic growth of DNA sequence databases, the rate of gene discovery has increased exponentially.

The power of bioinformatic approach for the discovery of genes has been proven with the completion of the yeast genomic sequencing project in 1996. Once fully sequenced, the bioinformatic approaches for identification of genes scanned the genome for genes. The genes found this way could be compared with the large number of genes that had already been discovered through more classical molecular genetic techniques. The results were remarkable. Before the yeast genome was sequenced in 1996, an international collaboration of scientists studying the genetics of this model organism had identified an impressive 2,000 genes by conventional genetic analysis. When the yeast genome sequencing was completed, bioinformatic searches for similarities of DNA sequences from other organisms were able to locate an additional 2,000 genes. This means in less than one year, a single laboratory using DNA sequencing and computer searches of sequence data could both duplicate and double the gene discovery of a 20-year international effort.

Once a gene has been identified, many new questions can be asked: what kind of protein does it code for and what is its function? How does it interact with other biomolecules of the cell? Is it expressed at all times as a so-called "housekeeping gene", or is it a developmentally regulated gene? Is its expression tissue-specific? Is it expressed in response to an environmental factor? These questions are the same questions that have been asked by molecular and cell biologists for decades, usually by studying one protein and its gene at a time. With the enormous information coming from the genomics project, however, biologists can ask the same questions about more complex systems. Instead of asking about one protein at a time, biologists can now ask questions about hundreds of proteins at a time, looking for patterns of structure and patterns of expression. Looking at the proteins on a genomic scale is a new field now called **proteomics.**

When a new gene has a sequence that has been found to be homologous to a gene in a database that has already been characterized, sometimes many of these questions about protein structure and function can be answered quickly by the bioinformatic approach. For example, the 2,000 new genes discovered by the yeast genomic sequencing project, discussed above, matched genes whose function had already been determined.

Bioinformatics approaches are playing an increasing role in protein structure studies. Although the final conformation of a protein is determined by the amino acid sequence of that protein, we have yet to model the correct folding of a protein by its amino acid sequence by computer. There is progress, however, in achieving this so-called "holy grail" of proteomics. As our database of proteins whose crystal structure has been solved, though, the more often we can predict protein structures by their sequence similarities. Also, we have discovered by analysis of sequence databases that there are certain conserved protein families with high sequence homology in part, if not all of the amino acid sequence. Computer program can currently predict protein structures by homology modeling when the sequence homology is as low as 25%. This means that if the amino acid sequences agree by more than 25%, the computer program can accurately predict the secondary and tertiary structure of the amino acid sequences.

Questions about how genes are regulated are rapidly being answered by a new technology **called "DNA chips" or a "microarrays"**. These chips are designed to allow many hybridization experiments to be performed in parallel. Oligonucleotides are synthesized on a glass surface similar to a microscope slide. Borrowing from the photolithography technology used to etch semiconductor circuits into silicon for chips by the computer industry, arrays of oligonucleotides can be laid out at a density of up to one million different oligonucleotides per square centimeter. By judicious selection of oligonucleotide sequences, complementary DNA for of all the genes expressed in an organism can be assigned at specific positions on a given microarray. A microarray can easily assay for which genes are being expressed in a cell by harvesting the mRNA from the cell, labeling the mRNA by covalent linkage to a colored molecule, and allowing the labeled mRNA to hybridize with the oligonucleotides on the microarray. The microarray is said to be "interrogated" in this way by the labeled mRNA. The genes that are actively being transcribed into mRNA by the cell are then determined by viewing the microarray under the microscope to see which oligonucleotides were hybridized with the labeled mRNA. Although there are very variations in the exact way that the microarrays are designed and in the exact way that they are hybridized with labeled nucleic acids, there is one thing that they all have in common: massive amounts of data from single experiments, requiring computer-assisted analysis and archiving of the results.

The power of the microarray and bioinformatics approach is having a major impact in medical research. For example, a biotechnology company called Sagres Discovery in Davis, California has recently announced that it has identified over 1000 new oncogenes in the mouse genome after just one year of research using this approach.

The practical applications from database information and the new bioinformatics tools are far-reaching. For example, with the discovery of a new oncogene and study of its structure and function comes the possibility of a new anti-cancer drug or strategy. The discovery of disease genes can lead to diagnostics for inherited diseases. Plant geneticists are using the detailed information coming from genomics to identify DNA markers to speed the breeding of new traits in our crop plants. With the discovery of DNA sequence **polymorphisms** (variations in allele frequencies within populations), comes DNA fingerprinting strategies for identity testing in forensics and the judicial system.

In this exercise, you will use the computer to access GenBank, the database repository of all DNA and protein sequences housed at the National Center for Biotechnology Information (NCBI) at the National Institutes of Health (NIH).

In Section I you will compare amino acid sequences of proteins from different organisms to study their evolutionary relatedness. In Section II you will use a DNA sequence to find a protein translational sequence **(ORF for "open reading frame")** of a plant defensin and study the function of this protein by finding homologous sequences in the protein database. In Section III you will use databases of the biological literature available online through the National Library of Medicine to discover what researchers are reporting for the structures and functions of plant defensin proteins.

Lab Procedures

I. Determine the evolutionary relatedness of species through comparisons of amino acid sequences of α- and β-hemoglobin

A bat looks much like a rodent until it flies, at which point it looks much like a bird. So what are bats more closely related to: birds or mammals? You will be able to discover for yourself the answer to this question by exploring the protein databases available to you. Since hemoglobin amino acid sequences have been studied extensively in a wide range of species, these proteins make a good candidate for comparing evolutionary relatedness. There are more sequences available for the alpha chain than there are for the beta chain, but you will be able to use either in this exercise.

1. Decide whether you want to do your work with α-hemoglobin or with β-hemoglobin. You may want to collaborate with a partner and do companion searches, one searching with α-hemoglobin, and one doing searching with β-hemoglobin. If this is the case, you will want to search for the same species of mammals, bats, and birds, and at the end of the exercise, you can compare your results with each other to determine whether your different proteins showed the same evolutionary relationships.

2. Go to the web site: http://expasy.ch/sprot/sprot-top.html.

3. Go about 40% of the way down the page to the area "Access to Swiss-Prot and TrEMBL". Click on the link "by description or identification"

4. In "Enter search key work" type "alpha hemoglobin" and click "submit". The results of this search will come up on your screen. How many protein sequences were reported to you from this query?

   _____

5. You may scroll down and look through this long list of α-hemoglobin sequence for one from a bat species, but it may be faster to narrow your search. Go back to the "Enter search key work" and type "bat alpha hemoglobin" and click "submit". When you get the results of this search, how many sequences of alpha hemoglobin did you get for bat species?

   _____

   ⇨ NOTE: Check the species names and common names for each of the α-hemoglobins that came in this sequence report to make sure that that are, in fact, bat sequences. Sometimes a search won't recognize the difference, for example, between a "bat" and some other word, such as "wombat"!

6. Select a bat α-hemoglobin sequence to save to a floppy disc by clicking on the color-highlighted and underlined **accession code** for that protein sequence. An accession code is how protein sequences are identified and archived in databases. In the case of α-hemoglobin sequences, this accession code will start with the letters "HBA". The symbols for all α-hemoglobins will begin with "HBA".

   ⇨ NOTE: If you are working with a partner who is doing a companion study with β-hemoglobin, you will have to collaborate on your selection of which bat α-hemoglobin sequences to save.

7. The page that opens will contain information about the sequence such as the taxonomy of the organism that it came from. At the bottom of the page you will see the protein sequence written with single-letter designations of the amino acids. To the right at the bottom of the page, you will see a link "FASTA format". This is the best way to save sequence information on your floppy, because it is a sequence format that all computer search programs can understand. Click the link and this will bring up a page with the sequence on it.

8. Copy the amino acid sequence to a floppy. To do this,
   a. Highlight the amino acid sequence,
   b. copy it (using the "Edit" pull-own menu),
   c. open Notepad (use your mouse to go from "START" in the lower left corner of your computer screen to "PROGRAMS" to "ACCESSORIES" to "NOTEPAD"), and
   d. paste it into Notepad.
   e. name your Notepad file and
   f. save it to your floppy disc (drive A).

   ⇨ NOTE: It is OK to copy and save your FASTA-formatted sequence using a word processor program rather than Notepad if you feel that this more convenient.

9. Return to the web page with the list of bat alpha hemoglobin sequences (2 "back" clicks on the web pages will get you there), identify another sequence for a bat α-hemoglobin and repeat the process of highlighting the FASTA formatted amino acid sequence to your Notepad file. Save all your FASTA formatted α-hemoglobin sequences together in one file on your floppy.

10. When you have saved two α-hemoglobin sequences from two bat species, repeat steps 3-9 to get 2 sequences from bird species and 2 sequences from mammalian species. It doesn't matter which species you choose, as long as 2 are birds and 2 are mammals. You may want to choose species that you predict are related to bats. (If you are collaborating with a partner searching for β-hemoglobin sequences, your partner should search for the same species that you have chosen.)

    => IMPORTANT: Be aware that if you are limiting your search for bird α-hemoglobin sequences with the keyword "bird", the search will only locate protein entries where the name "bird" appears. If the entry was archived under other descriptions such as "hawk" or "eagle" or "penguin, you will find entries under these categories.

11. When you have saved six α-hemoglobin sequences to your floppy disc (two from bats, two from birds, and two from mammals), go to **http://clustalw.genome.ad.jp.** CLUSTALW is a computer program that you can use to search for sequence similarities between many sequences at a time and display regions of alignment.

12. Copy your entire file of sequences saved on your floppy disc into the textbox and click "Submit" . Note that the sequence descriptions preceded by the ">" mark will be copied in with the protein sequences. This will not be a problem with your search. Without changing any of the default settings on your search, click on the blue colored "Execute Multiple Alignment" bar.

13. The page that will come up next will show the alignment of amino acid sequences for the 6 proteins that you have retrieved from the SWISSPROT database, using the single-letter designations for amino acids. An asterisk will appear along the bottom row of amino acid alignment at positions where there is an amino acid that is found in all 6 proteins. These amino acids are said to be "**highly conserved"**, since they haven't changed since these species diverged from a common ancestor.
    a. How many of the amino acids are found to be the same in all of the 6 α-hemoglobin sequences in your alignment?

       _____

    b. What percentage of all the α-hemoglobin amino acids are conserved in all 6 proteins? (You will have to count the number of conserved amino acids by hand.)

       _____

    c. Examine the regions of conserved amino acid sequences. Are there any specific regions of the α-hemoglobin sequences that are especially conserved? Is one end of the molecule more conserved than the other? Describe your observations.

       _____

       _____

d. Do you see any amino acids that appear more frequently in conserved regions of the protein than in the nonconserved regions? If so, which amino acids are they? (Go to the table at the end of this Lab Exercise to decode the single-letter designation for amino acids.)

_____

_____

e. If you did find amino acids that were more frequently conserved in your alignment report, were they ones with side groups that were nonpolar, polar, or charged?

_____

14. At the top of your CLUSTALW report, you will find the exact percentages of amino acids in the sequence alignment that are identical when comparing only two sequences at a time. For example, if your report says "Sequences (1:2) Aligned. Score: 87.2", this means that when the first two sequences saved on your floppy were aligned, 87.2% of the amino acids were identical in both sequences. Transfer these percentages into a table format, in which the species whose sequences you have aligned are headers for both the columns and the rows. Your table should similar to this:

Table 1: Percent identity in amino acid alignment for α-hemoglobins

SPECIES	(bat #1 name)	(bat #2 name)	(bird #1 name)	(bird #2 name)	(mammal #1 name)	(mammal #2 name)
(bat #1 name)	100%	(87.2% for example)	XXX	XXX	XXX	XXX
(bat #2 name)		100%	XXX	XXX	XXX	XXX
(bird #1 name)			100%	XXX	XXX	XXX
(bird #2 name)				100%	XXX	XXX
(mammal #1 name)					100%	XXX
(mammal #2 name)						100%

Notice that you need not fill out both halves of this table since the information is redundant. From this table, can you see whether the α-hemoglobin sequences are more similar for bats and birds, compared with bats and mammals? What does this suggest about the evolutionary relatedness of these species? Which species diverged from each other the most recently and have the most recent common ancestor? Which species diverged from each other the most long ago and have the most ancient common ancestor? **From the information in this you should be able to predict that bats are more closely related to either birds or mammals.**

_____

_____

_____

_____

15. A phylogenetic tree can present the relatedness of species from sequence similarity data such as your Table 1. These trees link species that are more closely related in "branches", and the length of the branches is their evolutionary distance. You can draw a phylogenetic tree from your amino acid alignment report by pairing species who have the most sequence similarities to make short branches, and branches. Species who have less sequence similarities will branch from each other on the tree farther apart.

    The CLUSTALW on the page that your report appears on will automatically draw a phylogenetic tree for you. At the bottom of the page, click on either the blue-colored "NJ-tree" or "Rooted Dendrogram". Print out the tree that appears on the screen. Does this tree agree with your analysis above of the "Percent identity in amino acid alignment for α-hemoglobins" table? Explain.

    _____

    _____

16. One way to test the validity of the phylogenetic tree that you drew for bats, birds, and mammals is to compare it with trees constructed from sequences of other proteins. Compare your tree with a tree constructed by your partner searching for β-hemoglobin sequences. Does your tree derived from β-hemoglobin sequences agree with one drawn from β-hemoglobin sequences? Are the length of the branches the same?

    _____

    _____

17. (OPTIONAL) Repeat the comparisons that you made (steps 1-16 above) with other species such as the following:
    a. Compare whales to mammals and fish.
    b. Compare reptiles to birds and mammals

II. Investigating proteins from DNA sequences

It is possible for you to discover new genes and to find suggestions of what a protein's structure and function might be from your home computer just by accessing the enormous databases of sequence information available at GenBank. In this exercise, we will explore a class of genes that code for "plant defensins". When a DNA sequence is known but the gene product has not been studied to determine where genes lie in the nucleotide sequence, you can look for start codons followed by stop codons in frame with the start codons and in the correct frame with a start codon. This approach is pretty much guesswork, though so a better approach is to query GenBank with the DNA sequence, looking for similarities with other DNA sequences in the database. When a match is found and the match has been characterized, you have strong evidence for the identity of a gene in the nucleotide sequence.

In this exercise you will do just that, search for a homologous nucleotide sequences to a patented plant defensin DNA sequence in order to identify the protein coded for in the nucleotide sequence. You will discover what other classes of proteins have a similar amino acid sequence and be able to explore not only the probable structure of the protein plant defensin protein, but also find suggestions for how this plant defensin protein might function in the plant.

1. Go to the web site http://www.ncbi.nlm.nih.gov.

2. The "Entrez" menu allows you to do searches based on the type of protein that you are interested in. Type "plant defensins" in the search box, and click on "Go".

3. The page that comes up will ask you which database you wish to search in. Select ""Nucleotide sequence" database. In the search report that you get, how many plant defensin DNA sequences have been deposited in GenBank so far?

   _____

4. The most recently submitted sequences are reported at the top of the list. Scroll down to the end of the list to the first sequences that were submitted for plant defensins. There will be a series of sequences that were submitted at the same time that are listed separately. You can access more information about these sequences by clicking on the colored identifier for the sequence. The identifier number is called an **accession code**, and is assigned by NCBI when a DNA sequence is submitted, for the purposes of archiving the information. In the case of these plant defensin proteins, the accession codes should look like this: **AX046743.**

5. When you click on the accession code for a sequence, a page will come up that gives a more complete description of the sequence. It will contain information such as who submitted the sequence, the type of organism that it came from, whether the sequence was reported in a scientific journal or not, and some information about the sequence itself. Click on a few of the earliest plant defensins and answer the following questions.

   a. Who submitted these sequences? _____

   b. When were these sequences submitted? _____

   c. Most scientific journals and the U.S. Patent Office now required that newly sequenced DNA that is being reported must be submitted to GenBank at the time the journal article is published or the patent is granted. Are the first plant defensins submitted to GenBank linked to a journal article or to a patent?

      _____

6. The nucleotide sequence appears at the bottom of the page that comes up when you click on the accession code. The DNA sequence is numbered from the 5' to the 3' end of the sequence. If the start codon for translation of the nucleotide sequence is known, that information will be given under the category "Features" and will be given as "CDS" for "coding sequence". When the coding sequence is known, the amino acid sequence that is translated from the nucleotide sequence will be given in the single-letter amino acid designation.

7. Are any the coding sequences known for the first plant defensins that were submitted to GenBank?

_____

8. This first submission of plant defensin sequence data included sequences from four different species: *Dimorphotheca sinuata, Picramnia pentandra, Parthenium argetatum, and Nicotiana benthamiana*. Click on the accession codes until you have located a sequences for a plant defensin from each of these species.

9. Copy four nucleotide sequences from this earliest submission of plant defensins, one from each of the four species, to a floppy disc. Since you will want to use this data for searches, save the sequence in FASTA format. Do this one at a time by
   a. clicking on the colored accession codes "AX046743"
   b. highlighting the nucleotide sequences that appear at the bottom of the page that comes up,
   c. use the pull-down menu at the top of the page under "Summary" to select "FASTA"
   d. copy the sequence (click on "Copy" from the "Edit" pull-own menu),
   e. open Notepad (use your mouse to go from "START" in the lower left corner of your computer screen to "PROGRAMS" to "ACCESSORIES" to "NOTEPAD"), and
   f. paste it into Notepad. Name your Notepad file and save it to your floppy disc (Drive A).

   ⇨ NOTE: It is OK to copy and save your FASTA-formatted sequence using a word processor program rather than Notepad.

   Repeat steps a-f for three other accession codes: "AX046747", "AX046751", and "AX046767"

10. Since there was no journal article to read about these plant defensin proteins, we don't know anything about them except their name, where they were found, and what their DNA sequence is. If someone else has sequenced a homologous gene and reported the function of the gene product, you may be able to deduce some information about the plant defensins that you have selected.

    To search GenBank for a DNA sequence that is homologous to the four sequences that you have saved, click the "Back" button until you have returned to the NCBI page that you began with, or go to http://www.ncbi.nlm.nih.gov again. Click on **"BLAST"** at the top of the page and select the blue-colored "Standard nucleotide-nucleotide BLAST". Copy one of your nucleotide sequences from your disc to the text box of BLASTN by highlighting the sequence and pasting it in. Click on "BLAST!" and on the next page click on "Format".

    ⇨ There is an excellent and easy to follow tutorial on how to perform BLAST searches of GenBank at this site: www.geospiza.com. Click on the "Education" button.

    GenBank is an enormous database, so this search may take a few minutes to complete, especially during busy times of the day. When your BLASTN report comes up, it will show a "Distribution of Hits on the Query Sequence" that displays in a figure the length of sequence that matched or **"hit"** the sequence that you did your search on (your **"query" sequence**). The alignment is scored by color code for how well the sequences aligned: red is the best agreement of sequences. If you mouse-over these color-coded lines, information about the sequence represented in the line will appear in the information box above the figure.

    Below the figure will appear an alignment report that includes some information about the species from which the sequence was taken and a score for how well the sequence aligned with your query sequence. This score is based on a statistical calculation, where a high score indicates a high level of agreement. An "E value" is also listed for each sequence in the report. This is another statistical calculation that estimates the probability that the sequences matched just due to chance. The shorter the stretch of sequence that aligned, the higher the E value, since it is more likely that a short sequence aligned due to chance.

Conduct a BLASTN search for each of your 4 saved plant defensin nucleotide sequences one at a time. If a search using these DNA sequences of plant defensins saved on your floppy disc do not result in a high level of sequence identity, discontinue any future searches with these. You can tell whether your search has found a good "hit" either visually by looking at the color-coded figure that shows regions of sequence homologies, or by reading the score and E values below this figure. If your search results in a long stretch of sequence homology, continue on and answer the following questions.

a. What is the name of the gene that gave you a good BLASTN hit (i.e. align well in a long stretch of sequence homology and has a Score of >100 and an E value of less than e^{-20})? What is the species and genus that this sequence came from?

_____

b. What is the score and E-value for the similarity between this sequence and your query sequence?

_____ _____

c. Scroll down below the color-coded alignment figure and below the blue-coded sequence alignment report. Look for a green "Alignments" header on the right side of the screen and look at the first aligned sequence. The first sequence alignment in this list will be the sequence that gave you the best hit for your query sequence. Of the 142 residues in your query sequence, how many are identical in the best hit sequence?

_____

d. Click on the color-coded accession code for this "best hitting" nucleotide sequence. The page that comes up will give you information about this DNA sequence. This sequence was published in a journal. What was the title of the journal article?

_____

If this article has been archived in the PubMed or Medline database of scientific journals, there will be a blue-colored number to the right of Medline or PubMed that will link you to these databases. If you click one of these numbers, you find more information about the journal article in which this DNA sequence was first reported. Frequently an abstract of the article is archived here. Could you find an abstract of this article in Medline or Pubmed?

_____

The page with information about the "best hitting" nucleotide sequence will give the nucleotide sequence at the bottom of the page, and the amino acid sequence higher on the page. **Find the amino acid sequence for this protein and copy it to your floppy disc**. (Follow the same procedure above for saving sequences to Notepad first in order to create this file.) You will use this file to search GenBank for matching protein sequences in the next step of this exercise.

11. Our "data mining" for information about plant defensins similar to the ones patented by DuPont has not been very fruitful so far, but there are other strategies that we can try. To expand your search for homologous proteins, you can switch from nucleotide searches to a protein search, using the protein sequence that you found to be homologous to some of the DuPont plant defensins. Nucleotide sequences of homologous genes aren't as highly conserved as amino acid sequences of the proteins that they code for, so sometimes a protein sequence search yields more sequence matches than a DNA search.

    Go back to the NCBI page at http://www.ncbi.nlm.nih.gov again and select **"BLAST".** From the BLAST page, select the blue-colored "protein-protein BLAST". Copy the amino acid sequence that you copied on your floppy disc into the query box and click on "BLAST!". Click on "format" when the next page comes up and wait for your protein search results.

On your BLASTP results page, you will notice a figure showing a red-colored bar designating a conserved domain. Click on "Gamma-thionin" to investigate the amino acid sequence that is conserved in a number of proteins.

Conserved domains are amino acid sequences that have been found in a number of protein sequences that have been submitted to GenBank. A conserved domain is annotated with a **P_{FAM} number** (for "protein family") to cross reference the proteins that have this conserved domain. What is the P_{FAM} number and the name for the conserved domain of your BLASTP search?

_____

a. Click on this P_{FAM} number to get an alignment report for this conserved domain. How many amino acid sequences are there in this alignment report, including your query sequence?

_____

b. At the top of the alignment report is a **"consensus sequence"** that shows the sequence of amino acids that are most often found in the alignment. How many amino acids are there in this consensus sequence?

_____

c. The alignment report is color-coded according to how well the amino acids are conserved in the consensus sequence. Red is the most highly conserved residues. To determine which positions are **absolutely conserved** (found 100% of the time in the alignment), we need to make our color code more stringent. Do this by selecting in the "View Alignment" row a pull-down menu for "color at" and select "identity". How many amino acids in the consensus sequence are absolutely conserved?

_____

d. The conserved amino acids are cysteine (C), glutamic acid (E), and glycine (G). How many times do each of these amino acids appear in the absolutely conserved consensus sequence?

Cysteine _____ glutamic acid _____ glycine _____

e. You can download a graphics program in order to see the 3-dimensional structure of these conserved proteins. You can download this program directly from this GenBank page by selecting the "download Cn3D" and follow the instructions. (If you can't download this "Cn3D" program from NCBI, you can find a tutorial to help at www.geospiza.com (select their "education" button) or you can aske your lab instructor for assistance.)

f. Once the Cn3D software had been downloaded, return to your pFAM00304 consensus sequence report and click on "View 3D Structure" in the gray box. A colorful graphic of the 3-dimensional structure of the consensus protein will appear in the upper left side of the screen.

The sticks in the structure that will appear on your screen represent the polypeptide backbone and 4 disulfide bridges between 8 cysteines. Notice that if you click and drag on this structure, it will rotate. Click on the "Style" pull-down menu to select "Rendering Shortcuts" for "Worms". This will change the appearance of your structure to show an alpha helix wrapped around a nail and beta-sheets represented by flat arrows. How many strands of beta helix are present in these conserved proteins?

_____

You can change the color coding on this 3-dimensional protein structure to display the degree of hydrophobicity of the amino acid residues. Click on the "Style" pull-down menu to select "Coloring shortcuts" for "Hydrophobicity". The blue color that appear on your structure represent amino acids that are hydrophilic and the red color represents amino acids that are hydrophobic. Click on the structure and rotate it so that you can see where the hydrophilic

and the hydrophobic residues lie in the 3-dimensional structure. Do you see any asymmetry in the placement of these residues, so that one side of the molecule is clustered with hydrophilic residues and the other side is clustered with hydrophobic residues? Which side of the molecule would you predict to be associated with other proteins or biological membranes?

g. To get an idea of how this gamma-thionin protein family works, go back to the Conserved Domain Database page and click on the "Proteins: Click here for CDART summary of Proteins…" bar. The next page will show that there are two other types of protein domains that overlap the consensus sequence. What are they called?

h. Click on the conserved domain with the longest stretch of similar sequences. What types of proteins are there with this conserved domain?

What is the mechanism of action of some of these toxins?

III. Searching databases for scientific literature

Although we have inferred a lot about the plant defensins by simply searching for sequence homologies, we still don't have a clear picture of what plant defensins are doing in plants, which sorts of plants make them, whether their expression is tissue-specific or developmentally regulated, etc. You can use online databases of the scientific literature to search for answers to these and other questions. We will continue our investigation of plant defensins by searching the database called MedPub maintained at NCBI by the National Library of Medicine.

1. Go to the site http://www.ncbi.nlm.nih.gov again, and click on PubMed. Type "plant defensins" into the query box and click "Go". How many "hits" did you get for this query?

   _____

2. It would probably be more convenient to narrow our search to get meaningful information more efficiently. You will notice that many of the articles listed are not in English. To limit your search to articles written in English, go back to your query page and click on "Limits" at the top of the page. On the next page, click on the "Language Types" pull-down menu to select "English". Click on go again. How many "hits" did you get for your query this time?

   _____

3. You can peruse the abstracts of the listed articles by simply clicking on them. Select some titles that look informative and search the abstracts for general information about what plant defensins are, what they are good for, how they work, and where they are found in nature. Not all authors include this general information in their abstracts, but some do, so don't get discouraged if the articles are too detailed. You will notice that some of the articles on the list have a colored box saying beside them saying "free in PMC" to notify you that the full length article is available for your viewing from PubMed Central without cost to you. Take advantage of the online availability of these articles by clicking on the "free in PMC" box. You will notice that most authors begin the introduction of their articles with the general information that we are interested in. General information often appears in the Conclusions section of scientific articles.

   How many full-length articles were you able to find in you PubMed search for protein defensins?

   _____

4. (OPTIONAL) Write a short report discussing what you have learned from the literature about plant defensins about their biological role in plants. Include what you have learned in Part II (above) about the patented plant defensins that were patented by DuPont, and suggest a reason that DuPont might have gone to the expense of patenting them. Of what interest are plant defensins to the biotechnology industry?

References

1. Brown, T.A. *Genomes.* Wiley-Liss. 1999.
2. Gibas, C. & Jambeck, P. *Developing Bioinformatics Computer Skills.* O'Reilly. 2001
3. Puterbaugh, M.N. & Burleigh, J.G. Investigating Evolutionary Questions Using Molecular Databases. *Am. Biol. Teacher.* 63:422-431. 2001

Amino Acid Designations

Nonpolar side groups		**Polar side groups**		**Charged side groups**	
Alanine	A or Ala	Asparagine	N or Asn	Arginine	R or Arg (+)
Isoleucine	I or Ilu	Cysteine	C or Cys	Aspartic acid	D or Asp (-)
Leucine	L or Leu	Glutamine	Q or Gln	Glutamic acid	E or Glu (-)
Methionine	M or Met	Glycine	G or Gly	Histidine	H or His (+)
Phenylalanine	F or Phe	Serine	S or Ser	Lysine	K or Lys (+)
Proline	P or Pro	Threonine	T or Thr		
Tryptophan	W or Trp	Tyrosine	Y or Tyr		
Valine	V or Val				