Name: _____

Introduction to Bioinformatics – AS 250.265
Laboratory Assignment 6

In the last lab, you learned how to perform basic multiple sequence alignments. While useful in themselves for determining conserved residues and identifying similarities between proteins, multiple sequence alignments are quite often used in molecular evolution applications to develop a phylogenetic analysis for a group of proteins. When using multiple sequence alignments in this way, it is important that your alignments are correctly constructed, thus, the topics covered in the last chapter are all the more important here.

Once a multiple sequence alignment has been determined, it can be exported in any number of formats to software packages that can make phylogenetic trees like the ones shown in class. Some packages that can do this are PHYLIP (PHYLogeny Inference Package), and PAUP (Phylogenetic Analysis Using Parsimony). Since PHYLIP is non trivial for picking up quickly, and PAUP is expensive, we will use a third, easy to use package called MEGA3. In this lab, you will start with a multiple sequence alignment in MFASTA format and examine the different techniques MEGA3 provides for making phylogenetic trees.

This lab will constitute the first part of your assignment for the coming week. Your answers from this lab will count as 35 points toward assignment number nine. You may submit your answers in class on April 28th, or you may upload them as part of your solutions for assignment nine on the course webpage.

Part 1: MFASTA MEGA Warm Up

The protein we will use for this lab is ubiquitin. As you may recall from earlier assignments and exams, ubiquitin is a "ubiquitous" protein in eukaryotes that is commonly involved in protein degradation among other things. It is a small protein with a relatively simple topology of one helix and a beta sheet.

We will start with a multiple sequence alignment of several different ubiquitins, downloaded from the conserved domains database (entry cd01803). These ubiquitins come from a variety of simple eukaryotes, from yeast to amoebas to algae. The alignment is stored in an MFASTA format, a file format that stores multiple FASTA files in serial:

Box 1: An Example MFASTA File

```
>Sequence 1 Name
ACDEFGHIKLMNPQRSTVWYACDEFGHIKLMNPQRSTVWYACDEFGHIKLMNPQRSTVWY
...
ACDEFGHIKLMNPQRSTVWYACDEFGHIKLMNPQRSTVWYACDEFGHIKLMNPQRSTVWY
>Sequence 2 Name
...
>Sequence N Name
...
```

Of course, the ellipses above should be replaced with the remainder of the sequence 1, 2, N, etc. MFASTA files can store multiple sequence alignment results by using hyphens (-) as placeholders where there are gaps in the sequence alignment. Thus, the MFASTA file is a fairly convenient way to store multiple sequence alignment results.

a. Download the `ubiquitin.fasta` file from the laboratory six web page. Open this file up using a text editor. How many ubiquitin sequences are stored in the file? (3 points)

Under normal circumstances this would be a carefully constructed alignment, but in this case, we will assume that the people at NCBI know what they are doing and simply use it without knowing exactly what went into the alignment.

Using Virtual PC

MEGA3 (Molecular Evolutionary Genetics Analysis) is a freely available software package, but it is only available to use on Microsoft Windows. Since the machines in the lab are (non-Intel based) Macintosh computers, you will have to use a software packaged called "Virtual PC." This program emulates an x86 system so that Windows and DOS programs can run on a (PowerPC-based) Macintosh.

You can open Virtual PC on the Macintosh by opening the Application menu and double clicking the "Windows XP" icon. If Virtual PC asks you about installing an operating system, simply click "Cancel" to continue booting the machine you just selected. After a minute or so (it's not a terribly fast emulator), you will be presented with a Windows XP login screen.

Just like a real system, the virtual PC you will use can be booted and shut down. In addition, since the machine is virtual, you can stop it at any point and quit the Virtual PC program. After saving the memory contents to disk, you can then close the software and then open up your virtual computer later, resuming exactly where you left off. For the purposes of this class, however, it is best if you "turn off" the virtual computer when you are done with it.

To log in to the machine, click on the "User" account and use the password: `fitzkee` (in all lower case). Again, in another minute or so, you should have a standard Windows desktop in front of you.

One of the nice features that Microsoft incorporated in to the software package is the ability to drag and drop files from your Macintosh desktop to the Windows desktop. This drag and drop mechanism is how we will transfer our FASTA files back and forth between the two machines. It's important that, when you are done with your Virtual PC session, you move all your files back to your Macintosh desktop. Otherwise, anyone who logs in to the system will be able to see what you were doing (the Windows XP systems are open access to anyone who logs into that machine).

When you are comfortable with the Virtual PC interface, drag your ubiqtuitin MFASTA file onto the virtual PC desktop.

Of course, if you are doing this lab at home or elsewhere, you can simply download the MEGA software package at the following URL. It is free, and they don't seem to spam your email inbox.

`http://www.megasoftware.net/mega.html`

Starting MEGA

MEGA is a powerful software package, and this lab will only touch some of its basic functionality. If you wish to learn more about using the program, it is suggested that you do the tutorials within the help system.

For now, start MEGA by locating it in the Start Menu. When you start, you will be presented with the main MEGA menu. Since MEGA uses its own custom alignment format, it cannot read your FASTA file as a raw input. Instead, you must first convert it to a MEGA project by creating an alignment.

To do this, select "Alignment" from the menu. When prompted, tell the software you will import your alignment from a file. Browse to the `ubiquitin.fasta` file on your desktop and open it. You should see a colorful MSA appear in the sequence viewer.

Now that we have opened the alignment, we can save it as a MEGA project file. You can do this by Selecting the Data → Export → MEGA file from the Alignment Exporer menu. Save your MEGA project, give it a title, and then close the alignment exporer. You will be asked whether you want to open your alignment in MEGA, and you should answer "Yes."

When MEGA has loaded your project, a "Sequence Data Explorer" window will pop up. For now, let's ignore this window and focus on the original MEGA window that opened when you started the program. This window is your "starting point" for many of the tasks we will perform in the remainder of this lab. Notice at the bottom of the window it indicates that your project is currently active.

No take a look at the Data Explorer window again. You will see that it contains information about the sequence. Along the top is the sequence of the first species in your list (Saccharomyces). Whenever this sequence is duplicated in one of the others, a period is placed in the box, and only when a sequence differs is the amino acid identified. To display the entire multiple sequence alignment, you can click on the "TA" button near the top of the Window. You can also color all the residues by conserved columns (C), variable columns (V), columns that exhibit parsimony (Pi), and columns that are only different by one amino acid (S).

Calculating Evolutionary Distance

Recall that the distance-based methods for generating a phylogentic tree all use an intermediary distance matrix before the tree is generated. MEGA will allow you to calculate and observe the distance matrix using various distance metrics. The simplest distance metric, the p-distance metric, does not take into account that residues at a single position can mutate twice. Thus, it is

only a good metric for highly similar sequences (what we have here). Other metrics have been developed to take into account the fact that a single residue may "mutate back" over evolutionary time. You can get a summary of these distance measures at the following website, but we will only use the p-distance here:

`http://www.hku.hk/bruhk/gcgdoc/distances.html#algorithm`

From the main MEGA window, select Distances → Compute Pairwise. You will be presented with a list of options. Change the substitution model to Amino Acid → p-distance. Then, calculate the distance matrix by clicking "Compute."

b. Which two proteins are most similar? Which are most different? (3 points)

Part 2: Building Phylogenetic Trees

Building a Neighbor Joining (NJ) Tree

One of the distance-based methods we discussed in class was the neighbor-joining tree algorthm. This method forms a distance based tree by constructing a starting "star shaped" tree and reshaping the tree until the most similar sequences are joined as neighbors. This method, like the UPGMA method, will produce very good (but not always perfect) phylogenetic trees.

From the main MEGA window, click on the Phylogeny → Construct Phylogeny → Neighbor Joining option. You will be presented with a similar option menu as before. Again, make sure that p-distance is being used as the substitution distance metric, and compute the tree.

As you view the tree, you can click on and select different branches and nodes. When a branch or node is selected, you can perform several operations using the buttons on the left such as specify roots, flip branches, etc. You can also click on the *i* button at the top of a window to determine how long each branch is.

c. Using the information menu, determine the horizontal distance between the Giardia and Euplotes phyla. To do this, sum up all the horizontal distances between the two OTUs. Compare this to the value calculated by the distance matrix and explain. (5 points)

From within the tree explorer it is possible not only view the phylogram in the rooted view, but it is also possible to view the tree from a radiation perspective with no explicit root. To do this, click on the picture of the tree next to the *i* button near the top of the window. You should be able to select "radiation" as a tree format.

You can either save this tree in MEGA's proprietary format, or you can export the image as an extended metafile, which can be read into Microsoft Word or various image editors. MEGA's format allows you to edit the tree later, whereas the metafile allows you to use the tree in other programs. To save the tree, select File → Save at the top of the menu. Choose a good file name for your tree, just in case you need to view it again. Also, export an extended metafile by selecting Image → Save as Extended Metafile (EMF).

d. Upload your radiation tree in extended metafile format to the course website. (5 points)

Building a Maximum Parsimony (MP) Tree

The procedure for building a maximum parsimony tree is similar. Remember that this method simply examines the differences between sequences, so no implicit distance measure is needed. There are parameters that are of some importance, however: mainly, we will have to tell the program how to search for trees, as in general there are too many trees to search exhaustively and thus a heuristic search is needed.

From the main MEGA window, select Phylogeny → Construct Phylogeny → Maximum Parsimony Tree. At the top of the window, you will see a tab labeled "MP Tree Search Options." Click this tab. The options here will determine how the software will search for optimal MP trees. The three options are different methods that generate many trees and select the best. Since character based methods do not use distances, there may be many "optimal" trees—the first technique is guaranteed to find all the optimal trees, whereas the second and third are heuristic algorithms and may not find all the optimal trees. For now, leave the setting at "Close-Neighbor Interchange" with the default options. Then, construct the tree by selecting the "Options Summary" tab and clicking "Compute."

When you view the result, the default display is a cladogram. Change this to a phylogram by making sure the "Display Only Topology" toggle button is off. Then, you will be able to determine how many amino acids are different between each sequence by looking at the scale. We will see why the topology only button is on by default.

For heuristic methods of calculating trees, you will see that the "Tree #" option can be changed at the top of the window. This allows you to scroll through the solutions identified by the search.

e. Scroll through several of the solutions with the topology only option off, then scroll through the trees with the option on. Which setting gives the more consistent answer? What is the difference between a "topology only" tree and a phylogram? (3 points)

The advantage to using a *bona fide* phylogram is obvious: you can immediately grasp the "distance equivalent" between two sequences. Ideally, we would like to possess some quantifiable difference between proteins, and this can be done by "averaging" the many different solutions we calculated heuristically into a consensus tree. To calculate the consensus tree, select the right-most button at the top of the tree window. This will allow you to specify a consensus cutoff: Specifying a cutoff of 50% will average all the trees and list topologies that are consistent with more than 50% of the solutions.

By default, calculating the consensus tree is still not a phylogram: while you can have more confidence in the topology, the quantitative relations between taxa is still not displayed. To display a true consensus phylogram (realizing that it is likely not correct here), you must de-select the "Compute Condensed Tree" button and once again make sure the "Topology Only" button is toggled off. The numbers by each taxon are indicative of how many (what percentage) of the solutions support using that phylogenetic distance.

f. Submit a copy of your consensus phylogram with a 50% cutoff as an extended metafile. Compare this image to your (rectangular) NJ tree. (5 points)

We have calculated the tree using a heuristic method. Now, let's calculate the consensus tree using all of the optimal MP trees determined by the "Max-Mini Branch and Bound" method. This method of searching for optimal MP trees is guaranteed to find all of the (degenerate) optimal solutions without performing an exhaustive search of the possible trees.

g. Repeat the steps you completed above, this time using the "Max-Mini Branch and Bound" method for constructing the tree. Generate a rectangular consensus phylogram at the 50% cutoff as you did in part (f), and submit your image as an EMF file to the course website. Compare the three trees you have made so far. (3 points)

Part 3: Bootstrapping

Although we have determined the "optimal" solutions to our phylogenetic tree problem, we have not investigated how robust are trees are. That is, if some of our data were to be removed, would we still likely obtain the same tree, or are our trees held together by loose threads of homology?

To calculate this, we can perform a "bootstrap" analysis on our tree. This will systematically remove several columns from the alignment and examine what happens to the tree as a result. Similar to the method of calculating consensus trees, one can then estimate how often a particular branch occurs based on the number of times it shows up in the partially deleted data.

We will perform a bootstrap analysis on our branch and bound solutions. To do this, construct an MP tree as before and select the branch and bound search method. This time, however, select the "Test of Phylogeny" tab. Click on "Bootstrap," and use the default options, which will generate 500 different data sets using our original alignment. When you run the calculation, you will notice that it takes somewhat longer to compute the tree.

On your new tree window, you will notice a tab appears that allows you to view the "Bootstrap" consensus sequence. This tree can be viewed like any other, and you can turn on or off topology only, etc.

h. Submit a copy of your bootstrap consensus *phylogram* as an extended metafile. Given that a bootstrap cutoff of 70% or higher corresponds approximately to a *p*-value of 0.05 or above, what can you say about the statistical significance of this tree? (5 points)

i. How do you think you could improve the statistical significance of this tree? (3 points)

When you have finished the lab, save the files you wish to keep and copy them to your Macintosh desktop. Then, delete the files from the Windows XP machine, and turn it off using the Start Menu options.