Alignment Lab

In this lab we're going to try to look at the effects of different methods of DNA alignment. We'll try different settings in ClustalW. We'll go through how to convert the output from ClustalW to a Nexus file, and how to manipulate that Nexus file. We'll also use POY, which does alignment and tree searching together.

Get the sequences

Since a lot of you have not finished the Alignment Assignment, I've made this lab flexible depending on where you are.

If you have not finished the alignment assignment, go to http://ib.berkeley.edu/courses/ib200a/Alignment%20Assignment.html and complete the assignment. Just skip the section on *ClustalW* in this lab. Instead do the alignments on line. Then start with the Making Nexus Files section, which will give you more complete instructions on how to convert to Nexus files and run them in *PAUP** and continue from there. Use the *Conus* sequence file for all the parts of this lab. Incidentally I've modified this file, so you now read sequence names instead of accession numbers.

The rest of you have three options:

1) If you want you can get sequences from the organism that you will be working on for your project. Go to http://www.ncbi.nlm.nih.gov/ search for your organism in the nucleotide data base, and try to find about fifteen sequences of a gene from different organisms that will be useful for your study. Download them to your desktop in FASTA format.

2)If you already know that there is nothing good on *genbank* (or you're just in a rush), I've prepared an example FASTA file of COI sequences from about 1f Cephalopod species that you can download from **http://ib.berkeley.edu/courses/ib200a/cephalopod_COI.txt**.

3) For the really lazy you can just download the *Conus* sequence file from **http://ib.berkeley.edu/courses/ib200a/sequences.fasta**. I won't hold it against you, and as I said above I've modified this file so that it's prettier.

ClustalW

ClustalW is the most commonly used program for multiple sequence alignments. It works by first doing pairwise alignments of each pair of sequences to calculate distances between them. It uses those distances to derive a tree. That tree then guides the production of the multiple sequence alignment.

You've already had a chance to use *ClustalW* a couple of times. The first time was through the *Jalview* sequence editor and the second time was through the Kyoto University Bioinformatics Center website. This time we're going to go directly through the *ClustalW* web page of the European Bioinformatics Institute, where you have more control over the search parameters. Since ClustalW is the industry standard, we're going to use it for comparison to *POY*, and to see the effects that parameters can have on the alignment.

-Go to http://www.ebi.ac.uk/clustalw/#.

For the first pass let's just use the defaults.

-From the output **format menu** select **'aln wo/numbers'**. Hit the **Browse** button and select your FASTA file. Then hit **run**.

A web page will come up, letting you know that your job is in progress. It will automatically refresh, so that when your job is finished, a new web page will come up.

-Pay attention to how long it takes, and when the job is finished, make a note of the alignment score.

There will be four files available for you to download. The output file is a description of the alignment process. The .aln file is your actual sequence alignment in Clustal format. The .dnd file is the guide tree used to make the alignment. The input file is the file that you uploaded to *ClustalW*.

-Download the alignment and guide tree files and give them unique names.

-Repeat the same alignment only this time run a **'fast'** rather than a **'full'** alignment. Don't forget to select **'aln wo/numbers'**.

It probably won't make a difference for this alignment, because you don't have a very big matrix, but for more species or a longer sequence a fast alignment can lead to a big savings in time. The 'fast' alignment works by only considering a chunk of the sequence at a time. The options on the second line refer only to the 'fast' alignment and have to do with how big a chunk of the sequence you consider.

How long did it take? Did it seam faster to you?
Is your alignment score as good as it was last time?

-Save the alignment file again, but don't bother with the guide tree.

Let's do this one more time, but this time let's give it really unrealistic parameters to see how the parameters do effect the alignment.

-Set the **gap open penalty** to 1 and the **gap extension penalty** to 5. This will make it easier to start a gap then to extend it, a highly unrealistic situation.

-Run the same sequences and save the alignment file. Don't forget to select **'aln wo/numbers'**.

Making Nexus Files

In the next stage we are going to use *Mesquite* and *PAUP** to make Nexus files and trees out of your alignment files. I haven't introduced you to Mesquite yet, and we'll deal with it more in a future lab. It is really good for this type of matrix conversion, so I'm just going to introduce you to some of the real basics here.

-Open *Mesquite*: Applications>IB200>Mesquite Folder>Mesquite OSX

-Open Your First Alignment: **File>Open File**

-Navigate to your first alignment file, select it and hit **open**
-An 'Import File' window will pop up, where you can select the format of the file you are inputting. Select **Clustal (DNA/RNA)** and hit **OK**
-Now a 'Save Inport File as Nexus File' widow will pop up. *Mesquite* will automatically save any matrix you import as a Nexus File. You should give it a different name so that it won't overwrite your original alignment file. I'm just changing '.txt' to '.nex'. Hit **Save.**

-That's it. A nexus file with your aligned matrix should now appear in your original folder and a window should appear showing that same matrix.

-Now open the other alignment files. You can open the other alignments in *Mesquite* without closing this one, so that you can look at all of them at once.

-Scan over the alignments to compare them. Are they all the same? How is the third alignment with the bad parameters? Does it make sense, when compared to the other two?

Deriving Trees

-Now shut down *Mesquite* and open *PAUP** (it's in the same folder as all the rest).

-Open the Nexus file from your first alignment. **File>Open File**, then select your file and hit **open**.

-Run a full search (**Analysis>Exhaustive Search**), if your matrix is 10 taxa or less or a heuristic search (**Analysis>Heuristic Search**), if you have lot's of taxa.

-Note the number of characters, the number of parsimony-informative characters, the number of minimum length trees and the length of those trees.

-If you get multiple best tees compute a strict consensus tree and save it.
Trees>Compute Consensus. Then check the box next to **Output to treefile.** Pick a location to save the file, name it, hit **save** and **OK**
-Otherwise save your best tree: **Trees>Save Trees To File.** Pick a location to save the file, name it, and hit **save**.

-Repeat this procedure for your other alignments.

Compare Trees

Now let's compare those trees in *Mesquite*.

There are two ways to do this. Normally we would first open a data matrix to look at the distribution of that data over the tree. However, in this case, since we don't care about the distribution of data, but only the topology of the trees. We can open the tree file directly and let Mesquite create its own false matrix.

-Open *Mesquite.*

-Select **File>Open file**. Select the tree file *PAUP** just created from your first alignment, and open it.
-A widow will pop up asking you if you want to create a fake matrix. Hit **OK**.
-A fake data matrix should pop up. Pull down the **Taxa & Trees Menu** and select **New Tree Menu**. Then pick **Stored Trees** and hit **OK**.

-To open your other two trees pull down the **Taxa & Trees Menu** and select **New Tree Menu**. Then pick **Trees Directly from File** and hit **OK**. Select one of the other trees and hit **OK**. Repeat for all your other trees.

You should now have three tree windows that you can look at together and compare. Do their topologies differ? Are they compatible? Which is the best resolved? If you assume that the tree from the full alignment is the 'best', then how do the other two trees look?

If you want, you can view the guide tree too, but you have to convert it to a nexus file first by opening it in *Treeview*.

POY

POY is a program by Ward Wheeler, David Gladstein, and Jan De Laet that is freely available on the web (http://research.amnh.org/scicomp/projects/poy.php). But it is really a difficult program, and I wouldn't recommend it to you guys. It views the alignment as a problem, in which gaps and deletions are events that happen along the tree. It tries to minimize those events as well as base pair substitutions,

-Create a folder on your desktop called 'POY folder'. Drag the unaligned sequence file into this folder.

-Open up the 'terminal': **Go>Utilities>Terminal**.

-Type 'cd' and drag your 'POY folder' into the terminal window. Hit enter. This resets that folder as your default directory.

Now we are going to set up a line in the terminal window that tells it to run *POY* on your data file and gives some instructions about what we want it to do.

-Drag the program **POY.macOSX** from **Applications>200A** into the **terminal window**. Don't hit enter. This sets up that we are going to be running POY.

-Drag **your unaligned data file** into **the terminal window**. Don't hit enter. This tells *POY* to read that file.

-Now type ">POYtree –iafiles –printtree". >POYtree makes POY output a tree file to your default directory named POYtree (you could call it something else). –iafiles makes *POY* output an alignment based on that tree in a FASTA file. –printtree makes *POY* output a representation of the tree to the tree file. *POY* has lots of other options that control the heuristic search parameters and the criterion for evaluating trees, but we aren't going to deal with any of them.

-Now you can hit enter.

If you ran POY on a PC you would put the POY program in the POY folder also. Then you would open a command window, navigate to the folder and type "poy input file name *>POYtree –iafiles –printtree".*

You should have four files in your 'POY folder'. The POYtree and ia files should have an output of the tree that you can view in a tree viewer. The 'Your file name'.ia file should have an alignment in FASTA format. And the POY file will have a version of the tree that you can view with a text viewer.

-Open the 'Your file name'.ia file in a text editor. How does this tree compare to the trees you got from the *ClustalW* alignment? Are they at all alike?

-Open the *POY* alignment (don't forget it's in FASTA format) and the full *ClustalW* alignment (Clustal format) in *Mesquite.* How do they compare?

If you want (In other words you don't have to) you can try running the POY alignment in *PAUP**. Does it get the same tree as *POY*? As *ClustalW* and *PAUP**? What about the bootstrap support on all these different alignments?