Biology 545 – Lab Exercise 1 – Alignment

This exercise has three goals. The first is to provide you with some experience navigating Genbank, the major database for molecular sequence data. The second is to provide you with some experience using ClustalW, a very widely used alignment program, and PAUP, a program for phylogenetic analysis. As we'll discuss in lecture, multiple sequence alignment operates via progressive alignment. The third goal is to demonstrate the influence that alignment parameters can have on the final alignment and subsequent phylogenetic analyses.

Here's what you'll need to do.

Get the sequence data.

Open your favorite search engine and go to the NCBI website www.ncbi.nlm.nih.gov/

Search for sequences by typing winchell mallatt under the Nucleotide database. Scroll to the bottom and set the display for 200 (from the default of 20).

Download the following 10 sequences:

AF212173, AF212174, AF342804, AF342803, AF342802,
AF342800, AF342799, AF342798, AF342797, AF342795.

Do this by checking the box to the left of each sequence. Then find *Send to* click on file and FASTA format. Save the file to your working directory. This will save all the sequences from these 10 organisms to a file that ClustalW can read.

Open ClustalW in your browser: https://www.genome.jp/tools-bin/clustalw

Load the sequences from the file you just created by clicking "Choose File" in ClustalW.

In ClustalW do the following:

> Create alignments with different costs and parameters. You will want to select the fast alignment option (Pairwise Alignment: FAST/APPROXIMATE). Record the parameter combinations you use below in question 1. Note that there are alignment parameters for both the pairwise alignment and the multiple alignment. Only the pairwise alignment parameters impact the results of this step. Before hitting the Execute Multiple Alignment command, make sure that you have selected FASTA as the output format. When ClustalW has finished, click on the You'll need to have 4 fasta files, one for each set of alignment parameters.

Save all 4 alignments to the same directory, and copy them to your RCDS directory by typing:
scp -r /"your local alignment directory" "your user ID"@ford.hpc.uidaho.edu:/mnt/ceph/"your RCDS directory"

Log into RCDS servers by typing: ssh "your user ID"@ford.hpc.uidaho.edu. Load the PAUP module by typing: module load paup, and launch paup by typing: paup. For usage in PAUP, we first need to convert our fasta files to nexus format. To do this, in the PAUP command line, type: ToNexus

fromFile=your_alignment.fasta toFile=your_alignment.nex format=FASTA. Load the nexus file into PAUP (exe your_alignment.nex), and do a quick neighbor-joining tree, just type nj.

Save the NJ tree by typing: Savetree file=Lab1.tre append=y;

Generate an NJ tree for the three other alignments and save each one to the same tree file (the append=y command keeps the program from overwriting the tree already in the file).

You'll have to edit the tree file and bracket out a lot of superfluous text. For PAUP files, they must begin with the text begin; and end with end;. Remove all intermediate begins and ends. You can also remove commented out text (indicated by either [] or >).

Next, read all of the trees into PAUP using: Gettrees file=Lab1.tre;

Once the trees are loaded, use treedist to generate a distribution of the tree-to-tree distances. You can also use ShowTrees to visualize all of the trees.

Now take each of your alignments to the GBLOCKS web server.

Go to the GBLOCKS server https://ngphylogeny.fr/tools/tool/276/form and upload one of your ClustalW fasta alignments. Choose a maximum number of contiguous non-conserved positions (smaller number=more stringent) and click submit.

When it's done, you can look at the conserved regions and output the cleaned alignment.

Do this for all of the alignments using the same stringency filters and, again use PAUP to compare the nj trees for each alignment. Are they more similar than they were before cleaning them? Compare your results with another student who used a different value for GBLOCKS stringency.

Questions

1. What were the 4 different alignment parameter combinations that you used?

2. How similar are the NJ trees from the first alignments?

3. After cleaning the alignments in GBLOCKS, were the NJ trees more similar? If not, do you think the alignment parameters or sequence cleaning may have contributed to those differences? How does the congruence between your trees compare to those generated using different stringency thresholds?