

## Laboratory Worksheet, Monday, Nov. 21.

**II. Protein Family Profiles III. Training Exercise Revisited.** A quotation from three (?) weeks ago: “This exercise will be about constructing (“training”) a protein family profile HMM from real data. In this exercise, you will be given a sequence accession number (NP\_000671: alpha1 adrenergic receptor). You will pass through some relatively simple steps: BLAST your sequence. Choose a handful of the best hits, but don’t choose overlapping sequences (choose distinct species, if possible). Then submit these protein sequences to CLUSTAL for MSA (multiple sequence alignment). You may do this from the command line using the local installation. Then use the MSA of your “seed” sequences, running this through *hmmbuild*, the profile HMM construction program in the HMMer suite. Having done this, you can compare to what Pfam has made of your sequence and its relatives.”

This week do the same thing initially except that you will now use the SAM program *buildmodel*. So, this time you will save half of your sequences which BLAST found (i.e., top twenty) and use half to train the HMM as above. Jeff will display the exact line commands you need to get started. Then use *hmmalign*, or the SAM equivalent, to align the remaining sequences to the HMM model. This will give you a larger alignment. Does this alignment compare well to an alignment of all 20 sequences done by ClustalW? Does this alignment depend on the sequences used in the “seed”? Notice that your seed is smaller than it was last week.

**II. PHX Data.** (*This is repeated from the last month!*) You should find PHX project data in the 548 Resources page, or on Ctools (depending on how far Dr De Wet and I have gotten with webDAV!). Please begin checking this for completeness. This is hand gathered data: you may want to improve it. Compare it to the Karlin-Mrceck paper if possible to see how complete any of these classes is compared to what was used in KM. The final point is to choose two organisms we can use in our project, and so you should be evaluating whether you want to choose organisms already begun or start over in data collection yourselves. We should ideally form two groups of two each to do this project. You probably won’t finish this this afternoon.

**III. PHYLIP.** PHYLIP is the name of the suite of phylogeny programs developed by Joseph Felsenstein of the University of Washington. In lab we saw some of these programs mounted at the Institut Pasteur in Paris. Please check the “mothership” at: <http://evolution.genetics.washington.edu/phylip.html>. This is also now linked to the Web Resources page, as well as to the Lab Worksheets page where you found this Worksheet. The “exercise” for now is simply to read the documentation for the bootstrap and for DNAML before next Monday. We will find the relevant pages in the lab.