

Laboratory Worksheet, Monday, Nov. 7.

I. Randomizing EM Training. This week we will take the randomized data we generated for the Casino problem and see whether we can improve performance by randomizing. I will just give you the most elementary version today, randomizing the inputs to the EM algorithm. So, in particular, this does not give you a Metropolis randomization during the calculation of the maximal log likelihood.

Look in Ctools for the new function script *samplestartEM.m*. This uses a special case of the *Dirichlet distribution*, or *Dirichlet prior*, on the set of all distributions on a single die. We will discuss that briefly before the lab.

So, the exercise is to take your 30 by 300 data set generated by *casinorandomizer.m* and use that as the data input for *samplestartEM.m*. As always, read the *samplestartEM.m* file to see what the inputs and outputs are, and the syntax of calling the function. Compare the results with your previous results using *dhmm_em.m*.

II. Protein Family Profiles III. Training Exercise Revisited. A quotation from two 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 HMMer program *hmmalign*. 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. Then use *hmmalign* 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.

Added this week, November 7, 2005: there is a command on ClustalW which will enforce the order of the sequences in the alignment, from top to bottom. Thus, at least that component of a comparison between the two results will be easier.

III. Protein Family Profiles IV. Searching with a Profile. The idea here is to use the HMM profile of I. or III. to search for family members against a

data base. We now have a very good database mounted locally, the Swissprot Database. You can now use the HMMer program *hmmsearch* to use the model you constructed in part I. above to search for family members in the Swissprot DB. Please use the documentation for HMMer to learn the command syntax. The safest thing to use as the name of the data base is the complete pathname, which will work whichever directory you are working in at the time: /usr/local/db/uniprot_sprot.fasta. This is the pathname on the lab linux machines. We will start distributing rewritable CDs this week, and this database will be something you can bring home to mount in your machine(s).

IV. 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 is to look at the directory of programs to get an idea of what is available. Jeff will show you how to get a good drawing of a tree and we will explore whether we can create different trees from our distance data from last time.

V. PHX Data. (*This is repeated from last week.*) 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-Mracek 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.