

Laboratory Worksheet, Friday, Jan. 30.

Preliminary Comments. Lab Procedure. I would recommend that you keep an online lab notebook. It might be good to keep things organized both by date, a file where you could dump whatever was worked on during a given lab session, as well as files appropriate for the different projects. Thus you might start one on codon bias.

I. Familiarity With Web Resources. The two sequence examples from last week were membrane proteins and were not represented in the PDB. One that will be is the potassium ion channel protein whose structure was solved by R. MacKinnon and coworkers. That should be enough of a description for you to find this protein in the various servers visited last time. [MacKinnon won the Nobel prize this year for this work.] See part IV below for comments about viewing the protein structures.

Another general type of protein you should try to compare across the three databases is cytochrome *c*. This is a protein used to transport electrons. It is a more standard water soluble protein which was solved some time ago. How does it operate? [This question does not have to be answered this week; you should be able to find where to learn this from the database examination this week.]

Finally, you should download the sequence data from GenBank locus ECORRD (the locus is another way to refer to *some* files in GenBank). How is this different from the previous examples? What kind of sequence is it? Save this to a file.

II. Pairwise Sequence Comparison. As a first example, compare the sequence of ECORRD in part I. above to the sequences in the file EctRNAdat in the 548 directory. (You should pull that file out and save it alongside your ECORRD file.)

The point of this exercise will be eventually to try to see whether we can detect any evolutionary relationship between these various sequences. This will be difficult and perhaps even inconclusive. There are reasons for this, and we can try to learn that, too. The ideas behind this are related to the so-called RNA World hypothesis about the biochemical precursor to the “modern world” of proteins. What I would like you to do – and it may not be possible yet, due to a lost application file – is to use the Waterman program for pairwise sequence comparison at USC (use the link in the “Web Resources” page to the USC server, then link to the alignment server), together with a report of *p*-value from a simulation estimate of the distribution. If you go to the online manual pages for Waterman’s SeqAln suite of programs, you can check for the command *pvlocal*. I do not believe the program is running yet locally, so for now, we can do one of two things: run Smith-Waterman from the USC site using the web server. This will only calculate the optimal score and alignment. The web server will not report the *p*-value. An alternative is to go to BLAST and use <http://www.ncbi.nlm.nih.gov/blast/b12seq/b12.html>, the pairwise BLAST function. This will use BLAST and will report the *E*-value. In either case (Waterman or BLAST), use linear gap penalties, using scoring parameters $s(a, a) = 1$, $s(a, b) = -0.9$, and $s(a, -) = -2.1$. In the lab, it will suffice to make a few of

these comparisons. Be sure to cover the case of cysteine in the EctRNAdata file and at least one other.

III. Putative Highly Expressed Genes (PHX). To carry out the Karlin-Mr'azek calculations, we will have to build a pipeline for importing data, code for making the basic calculations for each gene sequence and for storing the results.

a) As a first step, write a perl script which takes a protein encoding gene sequence (nucleic acid data), and computes the frequency distribution for alanine codons.

Eventually, we will need the frequency distributions for all amino acid residues. A script converting n.a. data to a.a. data for all codons is in the 548 resources directory. It is contained in the *Beginning Perl for Bioinformatics* folder.

b) Begin thinking about where we would get the data to make the PHX computations. Where would you look to begin finding it?

IV. Looking Ahead: Molecular Viewers. When we discuss structure and its relation to sequence we will want to be able to view and manipulate x-ray crystallography images of molecules. There are at last six freely available programs for viewing this sort of file. The principal file format is extension .pdb (from Protein Data Base format). The older viewers were *RasMol* and *Protein Explorer*. One that is used in the structural portion of the NCBI is called *Cn3D*. SwissProt also provides one, called *Deep View*. Another is *Chime*. Finally, a recent one is *iMol* for the Mac OS X operating system which our lab machines are running. The others will each require a parallel installation of the necessary graphics application. *iMol* has the advantage that the necessary graphics package is OpenGL which is built in to OS X. We may be able to download an image disk for *iMol* today and use it in local directories just for today to view some first examples. The disk image is not very large. Do not do this until I confirm this, however, since I have to check with the local IT people whether we could all make temporary local directory installations for the morning.

Often, an individual site will offer an interactive online view, but to use this you will need to be running the locally approved viewer. In general, though, you should be able to download a .pdb file to your computer and view with your viewer.

You may want to consider installing one of these viewers on your personal machine, if you have one you are using at home or away from the labs. I will see to it that the compatible ones are installed in the math labs and in the BICC in Med Sci II.