

Laboratory Worksheet, Monday, Oct. 23.

I. Familiarity With Web Resources. Go to the Web resources page from the class page sidebar. There are three fundamental ones you have to familiarize yourself with right away. These are the links labeled Entrez (NCBI at NIH), ExPasy and PDB (Protein Data Base). Basically, these are the extreme points in Web resources. NCBI is the largest and most complete depository of sequence and similar information. It is the default first setting for tracking down information from a query sequence. ExPasy is the site for SwissProt and other tools. SwissProt is a smaller database, but curated, so the reliability or coherence of the data is higher there. Finally, PDB is specifically a database primarily for structural information about proteins. There are often crosslinks between these, though unfortunately, they use different accession number systems.

This week I want you to just walk through these three locations following two query items. The first is the sequence data (in FastA format) which you will find as the file 548_unknown.dna.1 in the 548 Resources directory. (You may have to be careful about the “decorations” on this data.) Go to NCBI first, link your way through to BLAST. Notice that you will have to choose between nucleic acid versus protein submission data to the server. Submit the sequence, then follow up the links to the best “hits”. Link to the individual files associated with each match. Notice the parameters associated with each reported hit. From the file you should, in this favorable case, find a synopsis of what is known about the sequence. Find at least one paper where you could learn more about the nature of this sequence. It will be a gene coding for a protein. Go to PDB to see whether there is a structure file for this protein. This will be harder to guarantee. As an alternative, go to SwissProt and pull up their file on this protein. Is there substitute information about the structure of this protein there? What kind, and how is it obtained?

The protein CFTR from last week was a membrane ion channel protein. These are generally notoriously difficult to X-ray. One such protein that has been solved via X-ray crystallography 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 last 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 EctRNAdata in the 548 directory. (You should pull that file out and save it alongside your ECORRD file.)

vskip 2mm

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/bl2seq/bl2.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. Beginning an Exercise. There are simple but illustrative calculations behind a paper of Karlin and Mrázek on predicting highly expressed genes. The paper is available from the 548 resource directory, where it is listed under KMjbac.pdf. You should be looking at this paper, especially the first two or three pages. I want you to reproduce these calculations reported in this paper, at least in a few cases. The first step will be to write pseudocode for the calculations described in the paper in those first two or three pages. This means lay out the steps of where you get the data, what kind of data you will be handling, and what you will be producing. What do you have to do to the data to perform the calculations you need? This should be an outline and should take about a page or two to set down. Finally, where do you see difficulties in programming (that is, for your programming or for general adequacy of tools at hand)? Email me if you feel a need for some Perl background beyond the tutorial materials posted.

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

III. Beginning Perl. If you have need for it, we can begin the Perl tutorial this afternoon. This can be found in a subdirectory of the 548 Resources directory.