

Laboratory Worksheet, Friday, Jan. 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. 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?

Now repeat the same exercise, but using the acronym “CFTR” to plug into one of NCBI or SwissProt as a start. These sites will accept a variety of nomenclatures, though you sometimes have to give it a hint as to which you are using.

You do not have to turn in written versions of what you find, but you should be developing a lab notebook style which enables you to find the results of such searches at a later date. This may be a file, e.g., as opposed to something written.

II. 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.

III. Related Seminars. A very relevant seminar will be given next Thursday in the BI seminar series:

Thursday, January 29, 2003

4:00 - 5:00 PM

2903 LRC (2nd Floor of Taubman Medical Library)

**“Prediction of protein structure
and function on a genomic scale.”**

Jeffrey Skolnick, Ph.D.

Director, Center For Excellence in Bioinformatics

University at Buffalo, Buffalo