Introduction to Bioinformatics – AS 250.265
Laboratory Assignment 4

A strong motivation in the study of proteins has been the generalization of experimental results and incorporation into large scale servers that can be applied to proteomics problems. Hundreds of these sites now exist, and while we cannot explore all of them, we will examine a few in this lab. Specifically, we will use three different online services—BioCyc, ExPASy, and 2D-PAGE—to locate information about several different genes of interest.

In proteomics, there are generally two types of online services available: those that predict properties of proteins based on previous experimental results, and those that collect large amounts of data on previously performed protein studies. We will examine both classes of sites today, and as you work through the lab you should be mindful of the uncertainty in the data you collect. In general, those sites that collect data will be more reliable than those sites that attempt to make predictions, but even collected experimental data may be incorrect. Building your intuition with respect to what is reliable and what is not will help you, both as you work on your final project as well as if you ever do this type of research independently.

This lab will constitute the first part of your assignment for the coming week. Your answers from this lab will count as 40 points toward assignment number seven. You may submit your answers in class on April 7th, or you may upload them as part of your solutions for assignment seven on the course webpage.

You may notice that this lab requires you to be somewhat more independent than previous labs. This is because, in the real world, you will not have a lab handout to inform you about how to use the given web page. When you get to a point where you don't know exactly what to do, you may have to search the documentation on the website for clues on how to continue.

Part 1: The Chemistry and Domains of Phosphofructokinase

The protein 1-phosphofructokinase (PFK) is a protein involved in glycolysis. In this section, we will use several proteomics prediction servers and information storage databases to study this protein.

Enzyme Information: EcoCyc

Visit the EcoCyc website (`http://www.ecocyc.org/`). This website contains an exhaustive catalog of information for nearly every known pathway in the bacteria *E. coli*.

a. What reaction does PFK catalyze in *E. coli*? Ignoring ATP/ADP molecules, draw the chemical structures of the reactants and products. You can find all of this information without leaving the EcoCyc web page. (3 points)

b. What cofactor(s) does the enzyme require in order to catalyze the reaction? (2 points)

The Conserved Domains Database

The PFK protein structure is known, and valuable information is available about it on other websites as well. One of these websites is the conserved domain database (CDD) at the NCBI site. CDD is a manually curated database that stores the position specific scoring matrices corresponding to various known protein domains.
Visit the CDD website by selecting it from the list of databases in NCBI Entrez. (Recall that NCBI's website is `http://ncbi.nlm.nih.gov/`.) Search for PFK, and select the accession number `cd00363` from the list of results. Examine the page closely: the information stored on this site is entirely different from that stored on EcoCyc. Instead of detailed functional information, one has information about the sequences that are typical to the domain itself.

Toward the top of the page (after the summary) you will observe that the main PFK domain has been organized into a hierarchy. The domain entry you are looking at now is further subdivided into three groups. You can click on the subgroups to get more information on the sequences stored there. At the left, you can view information about this entry, including the number of proteins (rows) used to define the domain.

c. What do the three elements in the domain hierarchy correspond to? How many members are in each? (3 points)

Below the information on hierarchy is specific information corresponding to features in the protein. There is only one feature displayed currently, the active site, which performs the chemistry determined in part (a). Below the description of features is the multiple sequence alignment that defines the domain. Multiple sequence alignments like this are used to define the BLOSUM matrices as well as the position specific scoring matrices (PSSMs) you used when performing PSI-BLAST searches. You should be able to identify several positions in the multiple sequence alignment where a residue is perfectly conserved (that is, it never changes across many different proteins). Some, but not all, of these perfectly conserved residues correspond to the protein active site, as indicated by the pound symbols (#) at the top of the alignment.

The statistics for this domain indicate that 16 proteins are used to construct the PSSM, but by default only 10 are displayed. Select all the proteins by selecting "Display all 16 rows" from the row display option.

There are two aspects of the multiple sequence alignment that may not be immediately obvious. In the next two questions, you will figure out what the colors and the brackets mean in this representation.

d. Change the "color bits" option to "identity," and re-format the multiple sequence alignment. Slowly work your way back down to 2.0 color bits, going one step at a time. Explain what you think the colors red and blue mean on the multiple sequence alignment. (3 points)

e. Change the format of the plot to several of the other options available. What does the formalism "`.[4].`" mean? (3 points)

Depending on the "color bits" setting, it is possible to derive your own consensus sequences for active site residues simply by examining the proteins. Recall that RBP4 has a GXW motif near the retinol binding site that is useful in its function. In general, motifs are represented by PSSMs or (more popularly) by PHI-BLAST-like patterns.

Setting the number of color bits to three, we may make the assumption in this case that red residues are those that are conserved whereas blue residues are not conserved (i.e. they are wildcards).

f. Given these assumptions, write down a PHI-BLAST pattern that describes the PFK consensus sequence from positions 371-389 in the multiple sequence alignment. (3 points)

The InterPro Server, Gene Ontology, and PROSITE

To round off our investigation of PFK, let's visit the InterPro server, maintained by EMBL/EBI at `http://www.ebi.ac.uk/interpro/`. This site, like the other two, stores curated information pertaining to this gene that may be useful to us as experimenters investigating its function and structure.

Search the InterPro website for PFK. The information page on PFK contains cross references to many sites, including Pfam, SCOP, and CATH. The page also displays summary information about the domain itself. Unlike some of the other site, this page allows you to view the representative proteins by taxonomy (species classification) as well as other InterPro entries that share members with the PFK domain entry—in this case, all of the other entries contain subsets of the main PFK class that you are looking at. The philosophy behind this page is to display the cross references to other sites, similar to the NCBI web pages.

g. What are the gene ontology (GO) entries for process, function, and component? Is this protein usually found in the nucleus, mitochondria, or cytosol? (2 points)

h. Give one PDB ID for PFK. (2 points)

The InterPro also has links to PROSITE patterns of motifs in a given domain. Recall that motifs are short patterns of residues similar to what you determined above in part (f). In this case, motifs are supposed to be specific to the family of domains, which are (by definition) homologous. PROSITE patterns need not apply across homologous structures, however.

Click on the PROSITE documentation link on the main InterPro web page for PFK. You will see the PROSITE pattern corresponding to the motif you identified above in part (f).

i. Why do you think the two patterns are different? Which would you be more inclined to trust? (3 points)

From the exercises we have done so far, you should be able to appreciate how some simple online utilities can give you a great deal of insight into how proteins function. Although such information could be accessed through PubMed as well, the sites we have examined have sought to make understanding the function of proteins simpler than sifting through a long (and possibly difficult to understand) paper.

Part 2: Using the BioCyc BLAST Server

So far we have not visited any predictive web servers; however, it is easy to see how BLAST could be combined with all of these services to predict the function of a novel gene we are studying. In this next part of the lab, we will investigate that exact problem, once again using the BioCyc server.

Using the NCBI website, obtain the protein sequence for human retinol binding protein 4 (RBP4). By now, this should be very familiar to you. Suppose for a moment that the function of this sequence was unknown to us, and we wanted to gain insight on it possible function. By providing the ability to use BLAST to search its database, the BioCyc server gives us a valuable tool for identifying this protein's function.

From the BioCyc server main page (`http://www.biocyc.org`), select Database Search, and then choose the BLAST option. Perform a BLASTP search of the *E. coli* K-12 strain using the sequence of RBP4.

j. Knowing that our sequence is from humans and the organism is a bacterium, what would be the most appropriate scoring matrix to use for this BLAST search? (2 points)

In the BLAST results, you should observe one significant match. Unfortunately, the BioCyc server does not allow you to immediately access the protein function website by simply clicking on the accession number. You will have to copy and paste the accession number of your significant hit back in appropriate box on the database search web page. Remember to select the appropriate organism when you perform this second search—the *E. coli* K-12 strain.

k. What is the function of the protein you have identified? Assuming you did not know the function of RBP4, what would be the function you would predict based on your BLAST search result? How correct would this inference be? (3 points)

The BioCyc web server has other organisms stored, as well, but one of the drawbacks is that presently there are no higher organisms stored on this site. Therefore, when making inferences about pathways in humans or other mammals, you must exercise care.

Part 3: Predicted and Experimental 2D-PAGE

In this final section of the lab, we will explore a useful tool for examining the charge and molecular weight properties of proteins in the cell. We will do this using an online database of 2D polyacrylamide gels (2D PAGE). Recall from class that 2D PAGE is a technique that first separates proteins from a cell extract on the basis of their pI, that is, the pH at which the protein has neutral net charge. The, these "isoelectrically focused" proteins are saturated with SDS and run on a standard gel—sorting by molecular weight and shape.

We will compare some of the predictions made by the EBI pI/MW server to the observed pI and molecular weight on 2D PAGE gels The pI/MW server, because it is based on fairly well understood physical properties based only on the primary sequence, is one of the most reliable predictive servers available for proteomics.

First, visit the NCBI website and obtain the sequence to human triose phosphate isomerase 1 (TPI1, we used this gene previously). Visit the pI/MW server by navigating to the `http://www.expasy.org/` site and finding pI/MW in the right hand column (it's small, so look closely). Use the server to determine the predicted molecular weight and pI for human TPI1—write this down, as we will use it shortly.

Now, visit the SWISS-2DPAGE database, also available as a link from the ExPASy web page. Search the database for the human TPI1 gene—you should get one hit. Read through the documentation on the TPI1 page to determine what information is displayed on the gel. Note that the documentation uses a yellow box, in actuality the server displays the results in red. When you are finished with the documentation, click on the gel at the top of the list of results, from erytrholeukemia cells.

l. What does the solid red box mean? What does the dotted red box mean? What does the dot mean? (3 points)

On the original page listing the gels, compare your original predicted pI/MW to some of the values listed in the 2D PAGE gels.

m. What were your original predicted values of pI/MW? How do they compare with the values determined by the 2D PAGE experiments? Which value, pI or MW seems to vary more? (3 points)

Finally scroll down toward the bottom of the list of gels. There you will see several gels with multiple spots corresponding to TPI. These spots are often called "isoelectric isoforms." Your final task in the lab is to think hard about why a protein may have different molecular weights or pI's. Understanding this will allow you to see one of the major uses of the 2D PAGE technology.

n. Why might a protein exist in the cell with multiple molecular weights? Why might the protein exist in the cell with multiple pI's? (4 points)

It is hoped that this lab gives you experience using some of the popular databases containing information about protein folds and protein properties. In continuing your bioinformatics education, however, you shouldn't stop with these three websites. Google provides a useful resource in identifying other sites that may calculate protein properties that you are interested in. In addition, the journal *Nucleic Acids Research* (`http://nar.oxfordjournals.org/`) publishes a journal article every January focusing on new proteomics and bioinformatics databases. If you continue to do bioinformatics in your professional career, this journal is definitely worth a look.