## Introduction to Bioinformatics – AS 250.265
## Lab 1

Yesterday in class we discussed the use of the National Center for Biotechnology Information (NCBI) website to obtain information about genetic data and publications. The goal of this lab is to give you practical experience using the NCBI interface: how to navigate the website, how to perform basic and advanced searches. You will need to draw on what you learned in lecture yesterday and today, as well as the on-line help available to you through the website itself. In the second part of this lab, we will explore another website, the San Diego Supercomputing Center's "Biology Workbench." This web site provides many of the same features the NCBI website has, plus it provides a simple program for performing pairwise alignments, which we discussed today. Becoming experienced in using these sites will help you in the future as you try to identify novel genes or perform your own research.

This lab will constitute the bulk of the written part of your assignment for the coming week. As there is no programming involved, you may either complete the assignment and turn it in by hand or submit it electronically under the Assignment 2 inbox on the course website. Your answers from this lab (to the questions highlighted in gray) will count 40 points to your assignment.

### Part 1: NCBI Entrez and Searching Biological Databases

*Triose Phosphate Isomerase*

We are going to investigate the human triose phosphate (or triosephosphate) isomerase 1 gene. This gene is responsible for the reaction that converts dihydroxyacetone phosphate to glyceraldehyde-3-phosphate in glycolysis. For those of you who have not taken biochemistry yet, glycolysis is the pathway in cells where a simple sugar (glucose) is transformed into two pyruvate molecules, which are then used to generate energy for the cell. Much is known about this gene, and when it is deficient, severe problems can occur. A loss of function mutation in this gene would be lethal.

First, visit the NCBI website (http://ncbi.nlm.nih.gov/) and visit the "All Databases" page.

1. What would be a good search query to use for this gene that would specify both the name of the gene as well as the organism? For your answer here, don't worry about the difference between "triose phosphate" vs. "triosephosphate;" however, you may have to use both searches for the rest of the lab.

Use this search query in the Gene, Protein, and Nucleotide sections of Entrez. You should observe different results in each, although they will contain much similar information.

2. What is the RefSeq accession number for this gene in the mRNA form? For the protein form?

3. On what chromosome is this gene found?

4. How many amino acids are in the protein chain? What are the first five?

One of the useful abilities of Entrez is to cross reference recent publications that relate to this gene. A recent paper published implicates the triose phosphate isomerase protein in the disease Lupus.

5. Who were the three authors of this paper? What is this paper's unique PubMed ID?

*UniGene and ESTs*

The book describes the UniGene part of Entrez as a database containing expressed sequence tags (ESTs) clustered around individual genes. An expressed sequence tag is cDNA that has been stored of actual, expressed mRNA in a given cell. Visit the website for UniGene (it's found on the main Entrez page). In 2003, there were 26 organisms that were included in the UniGene project.

6. How many organisms are currently represented in the UniGene project?

Point your browser to the UniGene project for *Homo sapiens*. Each gene in the database is represented by a list of ESTs corresponding to that gene. By piecing together the ESTs, one can obtain a view of the entire sequence for a particular gene. At the end of the page, histogram is given that counts the number of ESTs per gene.

7. What is the general trend of the histogram?

8. How many ESTs are there for Human triose phosphate isomerase in UniGene?

*Additional Querying in Entrez*

In question 4, you determined the number of amino acids in triose phosphate isomerase. Suppose that you would like to find the total number of RefSeq proteins in humans that have that same number of amino acids.

9. What query would you use to find this out? What is the final answer?

10. How many papers have been published by researchers at Johns Hopkins on any form of "triose phosphate isomerase?" (Use the three word query here.) Given that a principle investigator is generally listed last in a list of authors, which investigator at Hopkins has published the most papers on this protein?

**Part 2: Logging in and Using BioWorkbench**

As we progress throughout the course, we will see that many of the tools for bioinformatics are spread across a broad range of disparate web sites with different interfaces and even in different countries. Much of the difficulty in accomplishing a task can simply be finding the information at one site and transforming it so it can be read at another. The San Diego Supercomputing Center (SDSC) has provided a free utility that in many cases simplifies the workflow issues that often arise in Bioinformatics. Their site, the Biology Workbench, consolidates many bioinformatics tasks by implementing them at one site with a common interface. Although the site requires you to create an account, use of the applications is totally free of charge (another plus, as some competing sites charge thousands of dollars per year to do essentially the same thing)

In this part of the lab, we will create a BioWorkbench account and obtain the sequences we will need for the final part of the lab.

First, browse to the SDSC BioWorkbench (BW) site and create an account. The site's URL is http://workbench.sdsc.edu/, or you can select it from a link on the course website. The username and account choice is up to you, but make sure you enter an appropriate email address. Don't worry about spam from this organization: in the five years I've been registered with BW, I've not received a single email from them once my account was created.

When you first log in, you will be given several options. BW works under a session paradigm. A session contains a collection of protein or DNA sequences that you have collected. As you collect many sequences, each for different projects, having multiple sessions allows you to organize the sequences into independent lists. As a user, you may create several sessions, or you may use the default session. For now, we will ignore the session tools, but if you find you enjoy this site and are accumulating lots of sequences, you may wish to consider exploring them.

The tools we will be interested in are those for proteins and nucleic acids. If you click on the nucleic acids button, you will notice a list of programs appears. Some of these allow you to manipulate your session—add sequences, edit sequences, delete sequences, etc. However, many of the options are useful bioinformatics tools to perform on sequences in your session. Below the list of tools is the list of nucleic acids in your session, which is currently empty.

*Adding a Nucleotide Sequence Using Ndjinn*

Whereas the search tool for the NCBI website is called Entrez, the search tool for BW is called Ndjinn (pronounced engine). Select the "Ndjinn – Multiple Database Search" option and click "Run." You will be presented with a menu of options that will be useful for selecting sequences from within any number of databases. The basic option, at the top, is simply a keyword search, which works in a similar way to the Entrez keyword search (except fields cannot be specified with brackets). Below the keyword search box, you can select the databases that you would like to search. The sheer number of databases should give you a high regard for the amount of genetic data out there. For now, check "GenBank RefSeq," and search for Retinol Binding Protein. Your search may take a minute or so, but in the results you should be able to find human retinol binding protein 4. Select this entry, along with the entry for human RBP1, and click "Import Sequences." This will copy these mRNA sequences into your session.

To test the search capability of Ndjinn, open up the NCBI Nucleotide database and find the accession number for the RBP4 mRNA (i.e. NM_XXXXXX).

11. If you search Ndjinn RefSeq for the Entrez accession number of RPB4, is the appropriate record in the search results you obtain? Is the appropriate record the only search result you obtain?

Scroll through the list of programs available for nucleotide sequences. You will see some familiar tools: one tool allows you to calculate the melting temperature of a DNA sequence (you implemented this using the Wallace rules last week). Another tool allows you to compute the reverse-complement of a sequence (you will implement this in your assignment this week).

Try out the application for displaying the database record of a sequence. Make sure that the RPB4 sequence is selected, then select the "View Database Records" option, and click "Run." After selecting how you would like your records formatted (use the default), you will see something that looks quite similar to the record for RPB4 in Entrez Nucleotide. A disadvantage of BW, however, is that it cannot integrate as well into the NCBI database: while the references are given, there are no clickable links into PubMed or PubMed Central.

Select the RBP4 mRNA and then run the program "SIXFRAME." This utility (which you will also implement in your assignment this week) takes an mRNA sequence and translates it into a protein sequence, which is then imported into your list of available protein sequences. You will be able to select which genetic code to use and which reading frame to translate. For our purposes, we will only use the standard code (though it's good to know there are others). For now, translate all six frames, and use the default options.

12. Why are there six possible translations for just one sequence of RNA?

As you view the results, you will see all six frames, and at the end you will see the longest *open reading frame* (ORF). An ORF is a sequence of DNA bracketed by a start codon (typically Met – AUG) and a stop codon (either ochre – UAA, amber – UAG, or umber – UGA). Select the longest reading frame an import it. Then, do the same thing with the RBP1 RNA sequence you found.

*Protein Tools*

The other way to import protein sequences is through GenBank directly. The protein tools section also has its own version of Ndjinn.

13. What are the Entrez accession number for the protein sequence of human apolipoprotein D and bovine (cow) beta-lactoglobulin?

Using the accession numbers you gave above, look up and import the protein RefSeq sequences for human apoliprotein D (precursor), β-lactoglobulin, and RBP4 (import RBP4 directly from GenBank, it will appear along with your translated sequence from SIXFRAME). When you have imported the sequences, compare the GenBank protein sequence with the longest open reading frame of the translated mRNA sequence using the “View Protein Sequences” option.

14. Are the protein sequences identical?

**Part 3: Simple Pairwise Sequence Alignments**

As discussed in class today, it is often profitable to perform alignments between protein sequences. As a warm-up for this process, we will perform some simple alignments of our own. From a human perspective, it is intuitive to align two sequences with a high degree of sequence identity. For example, suppose we have two proteins.

Protein A GGHKLAPVWQEDDNAIATLNCV

Protein B GHRLAPVQEDENAIATLYFNCV

One can, with a little inference, identify the optimal global sequence alignment between these two proteins. Minimizing the number of gaps, we have:

```
Protein A: GGHKLAPVWQEDDNAIATL--NCV
            ||.|||| ||| ||||||  |||
Protein B: -GHRLAPV-QEDFNAIATLYFNCV
```

In the above example, aligned residues are those that are matched with each other in the sequence. For example, the very first G-G pair is aligned, but also the R-K pair as well as the D-F pair are aligned. G-G (and all residues with a |) are marked as identical. Aligned residues that are biochemically similar are marked with a period, and those residues are aren’t similar but are nonetheless aligned have no marker. The residues that are given as gaps (-) are those residues which are not aligned.

15. Given the definition of sequence identity from class, what is the percent sequence identity in the above alignment? What is the percent sequence similarity?

To get a feel for sequence alignment, let’s work out another example. Consider the following two protein sequences.

Protein A MDFELVNDINGSVLQLGEVPR

Protein B MFEIVNDINGFFSVLNLGEVP

16. Manually align the two protein sequences above.

17. Given that L is similar to I and V, N is similar to Q, and D is similar to E, what is the percent sequence identity? What is the percent sequence similarity?

As you can probably tell, lowering the sequence identity would rapidly make alignment a more and more difficult problem. Computers are able to perform much better than humans would in this case, but even they are limited, and it becomes less clear whether two proteins are homologous when their aligned sequence identity slips beneath a given level. Sequences with only 10%-20% sequence identity are said to be in the “Twilight Zone,” since statistically speaking, it is very difficult to determine homology solely on the alignment results with only 10% sequence identity. Once sequence identity falls below 10%, inferences of homology are statistically meaningless. This region is called the “midnight zone.”

### *Global Alignment of Lipocalins*

Now that we’ve performed some manual alignments, let’s see what the computer will do. Select both your GenBank translated sequence for RBP4 and your Entrez Protein sequence from your protein session, and run the program “ALIGN.” For now, we will use all the default options. Looking at the results, you should see 100% sequence identity (in the upper left of the alignment result), and the overall alignment in a format very similar to what we introduced above.

That’s all well and good—the computer can align two identical sequences. Now perform the following alignments and answer the questions about them.

18. Compare the protein sequences of Apoliprotein D and RBP4. What is the percent sequence identity between these proteins? In your estimate, are these proteins homologous?

19. Compare the protein sequences of RBP4 and β-lactoglobulin. What is the percent sequence identity between these proteins? In your estimate, are these proteins homologous?

20. Compare the protein sequences of RBP4 and RBP1. What is the percent sequence identity between these proteins? In your estimate, are these proteins homologous?

When you have completed the lab, submit your answers as part of Assignment 2, either by hand at the start of class next Friday, or submit them through the course website.