Lab 3B: Bioinformatics, part II

**Objectives**

1) Expand your skills with the resources at NCBI and other public databases.

2) Review and connect lecture material on genomics with database resources.

3) Carry out a small project to investigate a protein family using database resources.

**Lab Questions**

1) How are genome-sequencing projects and databases connected?

2) What can I learn about my assigned protein family using database resources?

**Introduction**

Last week, in part I of this lab, you got some basic practice with database search tools, working with accession data, and comparing and aligning sequences. This week we will build on those skills and review lecture material on genome sequencing. Most of the material in this exercise comes from Chapter 1 of your Campbell and Heyer text. Make sure you read this chapter before lab and bring the text with you to class.

Then, each student will be assigned a different protein family. You will use the bioinformatics tools you have learned in these exercises to research and report on your protein family.

You may complete this exercise and Part I with a lab partner if you would like. No one should work in a group of more than two people. Turn in assignments from part II at the end of the class period. Then, each student should write his or her own lab report for protein family assignment. Your lab report is due Wednesday 24 March in lecture.

**Procedure**

Sign onto your lab group workspace. In this part of the exercise, you will use the NCBI web site and the web site for your Campbell and Heyer text. You may want to bookmark these web sites so that you can return to them later. Open each site in a different browser window (or use tabbed windows).

http://www.aw-bc.com/geneticsplace/

1. Sequence data and E-values: Read the text (before class!) and work through Discovery Questions 1-5.

2. EST databases: Read about ESTs, and BLAST the mystery sequence from your text web site. Answer Discovery Questions 10-12. The EST statistics are on the UniGene home page.

3. ORF prediction: Read about ORFs, and BLAST the ORF sequence from your text web site. Answer discovery questions 13-17. The ORF finder is at http://www.ncbi.nlm.nih.gov/gorf/gorf.html

When you are aligning the two mRNA sequences, why is the identity restricted to these three regions that comprise a small part of the total transcript length? What does this restricted area of transcript identity tell you about your answer to 17a? When you answer 17d, keep in mind that the COX1 and COX2 proteins are the molecular targets for all non-steroidal anti-inflammatory drugs, including aspirin. What can you tell from
the degree of sequence conservation near GAPFS and the fact that this is the aspirin binding site?

4. Use the Kyte-Doolittle hydropathy analysis (linked from your text web site - not NCBI - to determine the hydropathy profile of leptin. Give specific evidence for how these hydropathy profiles support our understanding of the hormone leptin and its receptor.

5. Do a conserved domain search (back at NCBI) with the Accession number P06784, instead of the one in your book. What is the difference between this search and the BLAST searches you performed last week on STE7?

6. Conservation and Phylogeny: Read about isocitrate dehydrogenase (IDH), and practice using COG and Swissprot to retrieve information about protein families. Go to the Phylogenetic Trees Method from your course web site, construct the phylogenetic trees for IDH, and answer Discovery Questions 24-30. Then, do a PSI-BLAST with the fly period sequence from your text web site. What is the accession number for the human ortholog? What does your text say about the combination of math and biology expertise?

7. Read Section 1.2, including Box 1.2 (before class!) Information from the summary statements is bound to be on the quiz. Answer Discovery Questions 35-38.

Figure 1: Who are these men and what do they have to do with genomics?