Introduction to Bioinformatics – AS 250.265
Assignment 10

The sequencing of the human genome represented a landmark effort in science, but many of its potential uses are still locked up because of our poor understanding of cell and protein biochemistry. To date, we lack the ability to predict the positions of introns with high sensitivity and specificity, and even when protein products are predict, we cannot always identify their structure and function computationally. This final assignment will give you a chance to apply the tools you have learned in this class in identifying a gene in a hypothetical sequence of nucleotides. Through this, you should appreciate how difficult it is to accurately identify genes automatically in much longer sequences like the human genome.

This assignment is worth 100 points toward the homework portion of the class. Electronic portions of the assignment should be submitted through the course website with appropriate filenames (e.g. `hw10-q2a.xls`). The written portions of the assignment may be submitted by hand or uploaded as a Word document or PDF file. The due date for this assignment is May 5th at the start of class.

Question 1: Annotating a Mystery Gene (60 points)

On the course website, you will find a file `sequence.fasta` that contains a hypothetical DNA sequence, representing a protein with its surrounding sequence. Your task for this part of the assignment is to identify the protein using whatever methods you can (excepting extortion or otherwise cheating!), and then to answer the questions below. Because this assignment is highly open-ended, 45 points will be based on your description of how you identified the gene (along with any programs you wrote and your supporting evidence). Thus, it is important for you to document your work.

The following facts about the protein sequence are given:

- There are no substitutions in the gene's amino acid sequence, and there are no mutations in the DNA sequence.
- The protein is translated in the forward direction from the given DNA sequence (and not the reverse complement)
- The protein is from *Homo sapiens*
- The protein is between 350 and 360 amino acids long
- The protein has RefSeq entries in both gene and protein
- In addition to upstream and downstream introns, there are two internal introns.

In searching for this gene, you are encouraged to use all of the tools and programs that you have written for the course. You are also welcome to download, use, and modify the solution programs if you don't wish to use your own. Some of the tools you may find useful are:

- SIXFRAME, from the Biology workbench
- BLASTP, BLASTN on the NCBI Website
- Entrez, to narrow your list of searches

- ALIGN, from the Biology workbench, to compare your results
- RepeatMasker, at any of the locations listed in your book (Table 16-7, page 550)

Here are a few hints that you may want to consider as you work on this project:

- The hydropathy program could be readily modified to calculate GC content in a particular window. A larger window size is likely more appropriate here.
- As early as possible, try to narrow the list of possible proteins using a rough estimate of where the introns are. Then, you can use this list to flush out the location using global sequence alignment.

a. Describe the method you used to identify the gene from the sequence data given. How did you determine the intron boundaries? What were the various stages of your solution sequence? Be detailed, as though you were documenting your work for a lab notebook or journal. (45 points)

b. What is the NCBI accession number for your protein? What is the protein sequence? (5 points)

c. Explain the function of this protein, to the best of your ability. (5 points)

d. What are the sequences of each of the four introns in your protein? (5 points)

Question 2: Comparing the IHGSC and Celera Genomes (30 points)

On the Entrez genome web page, it is possible to search for entire chromosomes. In this part of the assignment, we will roughly compare the International Human Genome Sequence Consortium (IHGSC) and Celera sequences.

a. Create a scatter plot comparing the sizes of each human chromosome in the IHGSC and Celera genomes. The IHGSC length should be along the x-axis, and the Celera length should be along the y-axis, and there should be twenty-three data points on your plot. Be sure to label your axes and give your plot a title.

   Using Excel's linear regression functionality fit a line to your data. Display the equation and R^2-statistic on the plot.

   Submit your plot in Microsoft Excel format to the course web page. (10 points)

b. What overall trend do you observe from the plot? (5 points)

c. How do the methodologies of the two genome projects differ? (5 points)

d. Using what you know of the methodologies, what might explain the systematic trend you observed in part (b)? (10 points)

Question 3: Final Project Progress (10 points)

The final project, for this class, is due at 12:01 am, on the morning of Monday, May 15. As everyone has opted either to identify a novel gene or characterize an unknown gene, this part of the assignment is designed to ensure you are on-track for completing the assignment on time, since no extensions will be given for the final project.

For this part of the assignment submit either (a) the amino acid sequence of the novel gene you plan to identify along with the organism, or (b) the closest homologue of the sequence you were given in class. (10 points)

Of course, you may choose not to complete this part of the assignment, but keep in mind that staying on top of your project is important, as it counts for 20% of your final grade.

Starting now, a "final project" option has been made available on the course homework submission web page for you to upload your final project report. If you choose not to upload your assignment, you may submit it by hand, but keep in mind that Jenkins Hall is typically locked on the weekend. You will have to use your key card to get in on Sunday, and then you should slip the project packet under the door to 201 Jenkins Hall.