In-class Assignment for Week 3

Due on April 21st, 2009

Part I

Exercise 1: Download the file mystery_sequence.txt from WebCT.

1. Perform a **blastn** and **tblastx** using the entire sequence. Set it up with the following parameters:
   Nr/nt database
   Bacterial genetic code (for the tblasx)

2. Perform a tblastx with the same settings as in 1 for:
   a. The first 40 base pairs
   b. The first 300 base pairs

3. Answer the following questions:

   a. What was the difference between the blastn and tblastx searches?

   b. Why were the shorter lengths chosen?

   c. For the full-length BLAST what organisms were closely related? What taxa do these belong to?

   d. Did this sequence share homology (similarity) with any known and functionally annotated genes?

   e. What was the difference between the tblastx searches performed by the full-length sequence vs. 300 bp sequence vs. 40 bp sequence?

Exercise 2: For this exercise you need the dinoDNA.text file from WebCT.

In the book "The Lost World" by Michael Crichton he used , Dr. Mark Boguski, at the NIH's National Center for Biotechnology Information (*NCBI*) as a consultant in order to create a DNA sequence that might be believable as that of a dinosaur. Mark chose a living organism that is closely related to the dinosaurs. In addition Mark mixed in some frog DNA just as was described in the first book, "Jurassic Park" by the scientist who describes how the dinosaurs were genetically engineered. While Mark was working on this for Mr. Crichton he decided to play a bit of a trick on him and he embedded a hidden message in the DNA sequence. Lets take a look at the sequence in the book and try to find the hidden message.

1. Download the dinoDNA.txt file from WebCT
2. Go to the BLAST website and start a new **blastn** search
3. Make sure to use the nr/nt database (i.e. search all of genbank)
4. Answer the following questions based on the BLAST results:
    a. What are the top hits of this BLAST search?
    b. What organism do you think Mark used for his dinoDNA sequence.
5. Now let's look for Mark's hidden message, to do this we need to start a new translated blastx search (**tblastx**)
6. This time use the SwissProt protein database which can be selected from the drop down menu
7. Now answer these questions:
    a. Are the top hits for the tblastx search the same as the ones you saw for blastn? Why might this be the case?
    b. What is the hidden message that Mark put in the sequence?

Part II

Introduction

We have just talked about how periodicity in DNA sequences can be detected. We discussed the difference between coding and non coding regions and how periodicity can be used to determine what type of region a specific DNA sequence is. Now we will do an exercise illustrating these concepts.

1 Locate a Sequence

Using your knowledge of Genbank from last week find the record for Homo sapiens oculocutaneous albinism II. Make very sure that the sequence you find is 3154 base pairs in length.

Save this Genbank record to a file on your desktop.

2 Load into MATLAB and Parse

Now that we have a Genbank record we wish to investigate we need to load this into MATLAB. To do this make use of the **genbankread** command. If you are not familiar with this command type **help genbankread** into the MATLAB prompt to learn how the command works.

Now that the record is loaded view the help file for the command **featuresparse** in MATLAB. This command lets you pull out certain parts of the Genbank record and save them into seperate variables.

Use the **featuresparse** command to seperate the exons from the entire sequence. Additionally, it would be helpful to view the indicies of where the exons are located which can also be done with the **featuresparse** command. Use these indicies to seperate out the introns from the sequences as well.

You should now have a list of intron regions and a list of exon regions saved into seperate variables in the MATLAB workspace.

3 Binary Indicator Sequences

Pick one intron region and one exon region from the list you have, preferably one of the longer regions.

Now transform the intron and exon into their 4 binary indicator sequences, u_A, u_C, u_T, u_G. There are a number of ways to do this in MATLAB including the use of a loop. One way to investigate that is very simply is the following (y1 is the sequence you are parsing):

bin1c=(y1=='C')

This will parse through the sequence, y1 and mark with a 1 each time the letter C is present. All other positions will be marked as a 0.

4 FFT and Power Spectrum

Now that we have the BIS for each an intron and an exon we can look at the power spectrum of each of these sequences. To do this we will take the FFT of each binary indicator sequence and then take the sum of the squares of all BIS for the intron. We will perform the same thing again for the exon.

Investigate the **fft** command in MATLAB to perform the FFT. Please remember that for the FFT to be most efficient we need to choose a length that is a power of 2. Additionally, for full resolution and no data truncation we must make the FFT length the same or greater than the length of the sequence we are processing.

5 Comparisions

Plot the Power Spectrum vs. Frequence for the intron region and the exon region on one plot with a legend. How are these two power spectrums different? Do you see periodicity in either region?

6 One Final Step

As an matter of investigation use **featuresparse** to seperate the entire sequence from the Genbank record. Perform the Binary Indicator Sequence and FFT and Power Spectrum steps on the full sequence and plot it. How is this different from the plots of an exon and intron region seperately?