HOMEWORK #11
DNA Sequence Analysis

Below are hints and links to sites which will be beneficial in answering the questions associated with this assignment. It will probably be beneficial to have several browser windows open while doing this assignment. On most browsers right clicking the links will give you an option to open in a new window. (Homework assignment as pdf file.)

Before beginning you will probably want to open the sequences into a word processor or text editor or separate browser window. This will allow you to cut and paste the sequences into the applications. Some applications require sequences to be in FASTA format (first line designated by >gene_name) and other applications will just require the sequences. Link to text file of the sequences in fasta format.

A compilation of tools for sequence analysis can be found at http://us.expasy.org/tools/

Question 1

Are these 3 clones likely to be of the same gene? Indicate which fragments overlap and their order.

An approach to this question is to align the sequences and look for exact (or nearly exact) matches. This will show the overlap (or lack of) between the sequences. Most alignment programs are not designed to do this specific task though and will give ambiguous results. A site for pair-wise alignments that does allow for such analysis is http://www.ch.embnet.org/software/LALIGN_form.html. Carry out three (x/y, x/z, y/z) pair-wise alignments and analyze the results and determine which sequences have long stretches of nearly identical sequences. This will allow you to determine how the 3 fragments are related to each other and their order with respect to each other.

Specific instructions. Paste the two sequences (do not include the name) into the two boxes and type the names into the indicated boxes. Check the 'global without end-gap penalty' button. (Local alignments will only show areas of overlap and may be a little more difficult to interpret.) Choose '1' subalignment. Choose 'DNA' scoring matrix. Sequence format is plain text. When ready, hit the perform align button. Results usually are returned immediately. Print out the alignments (or copy and paste them into a word processor) and analyze results. For 2 of the pairs you should see that the 3' end of one sequence is highly homologous to the 5' end of the other sequence. For the third pair, which will be the 2 sequences on the end, there will be very little homology (only short regions with a lot of gaps). From this information you know the order of the 3 clones and will be able to combine the 3 sequences into a single sequence.

Question 2

Are there any ambiguities in the sequences? If so, how would you resolve this ambiguity?

In the alignment(s) generated above look for regions of non-identity within the homologous regions.

Although not completely valid, without additional evidence you can choose the sequence with the ambiguity closest to the 5'-end as being correct since sequencing becomes less accurate the

further you go from the 5'-end.

Question 3

Do the compiled sequences encode for the entire protein? Why or why not?

You will need to analyze the compiled sequence for reading frames. In other words take the 3 sequences and assemble them into one sequence and correct any ambiguities. Either compile the sequence using a word processor or text editor or click here for a text file of the compiled sequence in fasta format.

A good site for analyzing reading frames is http://us.expasy.org/tools/dna.html. Paste in the compiled sequence and hit the translate button. The results will show translations of all 6 reading frames with start Met and stop codons highlight. Examine the sequences and click on the link with the most likely reading frame. (Hint: look for long stretches of amino acids without stop codons.) You can generate a protein sequence by clicking on potential start sites in the sequence. There are also links to other protein analysis tools as well as links to BLAST searches in which the newly generated sequence can be analyzed.

Question 4

Identify the protein or the class of proteins encoded by this sequence.

This will require doing either a BLASTP (protein) search using the protein sequence or a BLASTX (translated) search using the nucleotide sequence. A BLASTX search may be helpful if there are doubts about the reading frame or if there still may be errors in the nucleotide sequence.

To do a blast search, go to http://www.ncbi.nlm.nih.gov/BLAST/ and click on either 'Standard protein-protein blast [blastp]' or 'Nucleotide query - Protein db [blastx]' depending on the type of search you are doing. Paste the sequence in FASTA format into the window and click on the BLAST! button. If desired, the default settings can be changed. It will generally take a few minutes for the results to be returned. The results will be a list sequences in the databases with high homology to your sequence and alignments to the these sequences. There are also links to more information on the homologous sequences which may also included publications.