Regular Section - A Total of 100 points

Problem 1. GenBank (10 points)

The NCBI website at http://www.ncbi.nlm.nih.gov gives access to the GenBank database of nucleotide sequences (make sure to choose "nucleotide" in the search pull down menu).

Your task is to gather information on the photoactive yellow protein (PYP), an interesting protein that is the subject of many experimental and computational studies. If you search the nucleotide database for PYP (by typing PYP at the text box), you will obtain 26 results (at the time this hw was put together).

You want to narrow the set of results. One way to do so is to click on "Limits" just below the box where you chose the Nucleotide database. Limit the search to "gene name" by choosing Gene Name on the Field menu, and hit the "Go" button. Answer the following questions:

(a) How many entries do you obtain now?

(b) Click on entry U17017 to look at the Genbank record of this entry. How long is the nucleotide sequence?

(c) From which organism was this nucleotide sequence obtained?

(d) What is the percentage distribution of each of the four nucleotides - A, T, C, G - in the sequence? Do you notice any bias towards a particular nucleotide?

Problem 2. Swissprot (15 points)

Now you will gather information about PYP from Swissprot, a database of amino-acid sequences that can be found at http://us.expasy.org/sprot/. At this website, type PYP and click the Search button. Click on entry PYP_HALHA (P16113). Answer the following questions:

(a) How many references are cited for this protein?

(b) How long is the amino-acid sequence?

This Swissprot cross-references with sequence databases, 3D structural databases, and family and domain databases. Keep the page for this record opened because you will use it to reference records in other databases related to PYP_HALHA (P16113). One of the databases to which Swissprot links is Pfam (Protein Families), a database of multiple alignments. Pfam accession numbers begin with the letters PF, followed by five numbers (e.g. PF12345).

(c) What is the Pfam accession number for this entry?

(d) Click on "Graphical View" alongside the Pfam accession number. What domain do you obtain for the PYP protein? You should also be able to confirm that the number of amino acids is the same as your answer to a question above.

(e) If you click on the PAS domain, you obtain more information on PAS domains. What is the biological function of proteins that belong to the PAS domain? What are the structural characteristics of this domain?

(f) If you go back to the Swissprot page associated with the PYP_HALHA (P16113) entry, you should be able to see a list of all PDB entries for PYP. How many PDB entries do you see? How many of these are listed as resolved through X-ray crystallography and how many with NMR?

(g) Which entry has the best X-ray resolution? Clicking on that entry should allow you to jump to the PDB. You should be able to see that the authors that submitted this PDB entry are Genick, U.K., Soltis, S.M., Kuhn, P., Canestrelli, I.L., and Getzoff, E.D.

Problem 3. The Protein Data Bank (10 points)

Another way to obtain structures of PYP protein is by simply typing PYP on the PDB database, which can be found at http://www.rcsb.org/pdb. Click Search. This gives too many entries. You can get to the same entry as in Problem 2(g) by simply typing the PDB entry id on the textbox at the website of the database. Instead, here you will focus on another deposition by the Getzoff lab, with PDB entry 2phy. Search the PDB database with this entry.

The menu on the left allows to download the coordinate file associated with the entry. Click on Download Files, and choose PDB Text. The lines that show the coordinates of the atoms in the deposited structure start with the ATOM keyword.

(a) How many atoms are there in the file?

(b) What kind of atom is atom number 899?

(c) What are the x, y, z coordinates of this atom?

Problem 4. Visualizing a Protein Structure with VMD (15 points)

You will now visualize the 2PHY.pdb file you downloaded in Problem 3 with vmd. Load the file in vmd and display it using the NewCartoon representation, using Structure to color it.

(a) How many secondary structures do you see?

(b) How are the secondary structure elements colored?

(c) You can now render the display using snapshot. This will create a .tga file in your machine. Print this file and submit it with the homework. You may want to change the color of the background from black to white to save ink. This can be done under the Graphics option of the menu, under Background.

Problem 5. ENSEMBL (10 points)

For this problem, you will use the ENSEMBL database at http://www.ensembl.org/. Answer the following questions:

(a) How many species do you see as available on the ENSEMBLE home page?

(b) Search for anything with zinc finger, which is an important structural motif in proteins. Find the first human GENE and go to the corresponding page. You will see that the results are grouped in several ways. Focus on GENE index and follow the first link. What is the mouse gene id associated with the first link?

(c) What is the genomic location for this mouse gene?

(d) How many cDNA transcripts do you see? What does cDNA refer to?

(e) Click on Population Comparison and then Comparison Image. What do the resulting pages show? How many variations do you see? What is the meaning of these variations?

Problem 6. Searching a Nucleotide Sequence Database (10 points)

Go to http://bioinformatics.rit.edu/workshop2003/database/dbLab.htm (if the rit link is broken, try http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/BLAST/q_jurassicparkDNA.html), and search for the paragraph on the JurassicPark DinoDNA (the paragraph on which Dr. Henry Wu explains the principles behind the assembly of dino DNA). A small fragment of this DNA is available at the end of this paragraph. This fragment is copied here for your convenience, as well.

>DinoDNA "Dinosaur DNA" from Crichton JURASSIC PARK p. 103 nt 1-1200
GCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAAATCGACGC

GGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAAGCTCCCTCG
TGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGGAAGCGTGGC
TGCTCACGCTGTACCTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTG
CCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAA
AGTAGGACAGGTGCCGGCAGCGCTCTGGGTCATTTTCGGCGAGGACCGCTTTCGCTGGAG
ATCGGCCTGTCGCTTGCGGTATTCGGAATCTTGCACGCCCTCGCTCAAGCCTTCGTCACT
CCAAACGTTTCGGCGAGAAGCAGGCCATTATCGCCGGCATGGCGGCCGACGCGCTGGGCT
GGCGTTCGCGACGCGAGGCTGGATGGCCTTCCCCATTATGATTCTTCTCGCTTCCGGCGG
CCCGCGTTGCAGGCCATGCTGTCCAGGCAGGTAGATGACGACCATCAGGGACAGCTTCAA
CGGCTCTTACCAGCCTAACTTCGATCACTGGACCGCTGATCGTCACGGCGATTTATGCCG
CACATGGACGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAA
CAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTTCCCCCTGGAA
GCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTCCGCCTTTCTCCCTTCGGG
CTTTCTCAATGCTCACGCTGTAGGTATCTCAGTTCGGTGTAGGTCGTTCGCTCCAAGCTG
ACGAACCCCCCGTTCAGCCCGACCGCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCA
ACACGACTTAACGGGTTGGCATGGATTGTAGGCGCCGCCCTATACCTTGTCTGCCTCCCC
GCGGTGCATGGAGCCGGGCCACCTCGACCTGAATGGAAGCCGGCGGCACCTCGCTAACGG
CCAAGAATTGGAGCCAATCAATTCTTGCGGAGAACTGTGAATGCGCAAACCAACCCTTGG
CCATCGCGTCCGCCATCTCCAGCAGCCGCACGCGGCGCATCTCGGGCAGCGTTGGGTCCT

Go to the NCBI Blast home page at http://www.ncbi.nlm.nih.gov/BLAST/. Go to the link that says nucleotide BLAST (blastn) and paste the DinoDNA DNA fragment into the text box. Make sure to choose the Expressed Sequence Tag Database to search against. Hit the Blast button. You should get about 197 Blast hits.

Do you notice anything particular about most of the hits? What does this mean? Any ideas on what species Dr. Wu used to make up the dino DNA fragment?

Problem 7. Tracing Needleman-Wunsch (10 points)

Recall the recursion used to fill the dynamic programming matrix in Needleman-Wunsch:

$$S_{i,j} = \max \begin{cases} S_{i-1,j-1} + s(x_i, y_j) \\ S_{i-1,j} + g \\ S_{i,j-1} + g \end{cases}$$

Fill in the dynamic programming matrix and find an optimal global alignment for the DNA sequences GAATTC and GATTA. Use a score of $s(x_i, y_j) = +2$ for a match and -1 for a mismatch, and a gap penalty of 2 (that is, $g = -2$).

Problem 8. Semi-global Alignment (10 points)

Suppose we do not want trailing gaps. That is, we do not want gaps at the beginning or the end of an optimal alignment. How would you modify Needleman-Wunsch so as not to allow such gaps?

Problem 9. Tandem Repeats (10 points)

Tandem repeats refer to sequences with repeted units (subsequences) that are usually short. For instance, a tandem repeat in human DNA comprises hundreds of copies of a 6-nucleotide repeat TTAGGG.

Consider the following problem: We are given a short sequence p such as TTAGGG and we want to find repetitions of this sequence in a long DNA sequence s. However, we want to find imperfect repetitions. That is, we allow a few mismatches and so want to find "imperfect" tandem repeats of p in s. Design and describe a dynamic programming method to find such imperfect repeats. Hint: Use wrap-around dynamic programming where your space and time costs remain in $O(|p| \cdot |s|)$.