Biol 47800/59500 Homework 5

1. Retrieve the protein sequence of the E. coli strain MG1655 RecA protein from UniProt or Entrez and then submit the sequence to the University of Virginia protein FASTA server (*http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm=select&pgm=fap*).

The protein sequence should be 353 residues long.

- A Genbank GI number, or the sequence itself in FASTA format, may be pasted into the sequence entry window. Please record the ID of the sequence you use in your query in your report.
- Select the protein database described as the Human/Refseq proteins (section C), and
- check the box that says "Show Histogram" (Other search options) , but otherwise use the default search parameters provided by the program,.

Answer the following questions:

a. What scoring matrix and gap penalties were used as default values by the FASTA program?

b. Identify the name and gi (GenBank index) of the highest scoring sequence.

c. Look at the alignment for the highest scoring sequence. How many standard deviations above the mean is this score? Note that what is shown in the result is a scaled Z score, z'. z´ is a normalized score, calculated as z´ = 50 + 10Z, where Z is the raw Z score. This raw Z score represents the number of standard deviations that a given score, S, is from the mean, calculated by z = (S - m)/s, where m is the mean and s is the standard deviation.

d. Look at the lowest scoring alignment that is reported. This should be a sequence that is unrelated to the query. What is the (unscaled) Z-score?

e. How many database sequences were searched? What is the probability of an alignment between unrelated sequences achieving a Z score as high as the alignment in d.?

f. By looking at the scores and E values from this search, what is the approximate value of z´ that corresponds to an expect value of 0.3? How many sequences reached this high a score? Explain why this value would or would not be an appropriate cutoff value for significance.

g. Is the comparison of Rad51B (NP_598193) with RecA protein significant and why? How could the significance be further tested?

h. What does the match in f suggest about the structure and function of RAD51B?

i. What is the least significant E-value reported in this search? Explain why this score is or is not significant.

2. For protein database searches, the BLASTP algorithm first makes a list of three-letter words in the query sequence and then scores these words for matches with themselves and with all other possible words using the BLOSUM62 scoring matrix. The 50 highest-scoring matches are kept. Database sequences are then scanned for matches to these high-scoring words, and if such are found, a local alignment is made with the query sequence by dynamic programming. Use the BLOSUM62 scoring matrix in Figure 4.4A, page 83, in the text. Note that the matrix values are in half-bit units.

a. Suppose that the three-letter word HFA is in the query sequence, what is the log odds score of a match of HFA with itself?

b. Scan through the table and find the highest-scoring match with H (say amino acid X_1). What is X_{H1} and what would be the score be for HFA matching X_{H1}FA in the database sequence?

c. Scan again and find the worst-scoring match with H (amino acid X_{H2}). What is the score for a match of HFA with X_{H2}FA?

d. Repeat the last two questions for the second and third letters in HFA, i.e. what are the scores for HFA compare to HX_{F1}A , HX_{F2}A, HFX_{A1}, and HFX_{A2}.

e. How many possible matches are there with HFA? (BLASTP uses approximately the best 50.)

f. About how many neighborhood words (not number of windows) will be searched for in the database sequences in a BLASTP search, starting with a query sequence that is 300 amino acids long?