Introduction to Bioinformatics – AS 250.265
Assignment 4

Of all the algorithms in bioinformatics today, BLAST is one of the most important. It is therefore important that you understand how to use this algorithm and when it is appropriate to apply it. This assignment will test your ability to use the BLAST package available at the NCBI website, and it will also expose you to the types of problems that are best solved using BLAST.

While there is no programming portion to this assignment, you will have to reason analytically about how the BLAST algorithm is working for your solution to question 1. In your answers to all the questions, please be as concise as possible. At several points you will be asked to justify your selection of certain BLAST parameters—think hard about why you choose the options you are choosing, particularly if you deviate from one of the default options.

This assignment is due March 10, 2006 at the start of class. It may be submitted by hand or uploaded as a Word document or text file to the course web page.

Question 1: BLAST-off (25 points)

In this question you hand-simulate the BLAST algorithm on a simple query sequence and target database. Imagine that you live on the planet Ludicon IV where proteins are made up of only three amino acids. These amino acids are athyphine (A), bandoic acid (B), and cryzinine (C). These residues are joined in much the same way as they are on earth, and your recent discovery of Stephen Altshul's work from a stray satellite broadcast has led you to investigate the similarities of your own proteins. After some effort, you devise the following scoring matrix for your planet's biosystem:

	A	**B**	**C**
A	3	1	-1
B	1	5	-3
C	-1	-3	4

As is the case so frequently, the Ludicon IV government controls much of the science funding that you will need to investigate protein relationships. As is also the case so frequently, the Ludicon IV government is not made up of the brightest minds on the planet. You have been asked to demonstrate how the BLASTP algorithm works to your local representative. In this problem, you will construct your own outline of how the BLAST algorithm works.

a. Your search query for your presentation will be BBACA. Using a word length of three, generate the list of continuous words for your query sequence. (2 points)

b. Write down every possible three-letter word that can be constructed from the protein alphabet. Next to each word, write the ungapped alignment score each of the words you derived for part (a). (5 points)

c. Using a threshold value of T=10, highlight (or star, or mark) those alignments in your table above whose score is greater than or equal to T. The rows with at least one column highlighted will be searched in your database. (5 points)

Once you have outlined exactly how phase one of the BLAST algorithm works, you plan to discuss briefly what happens in phase two of the algorithm.

d. In your own words, describe how the high scoring alignments from part (c) are used in phase two of the BLAST algorithm. (5 points)

In your presentation, you plan to use a simple example for illustrating how phase three of the algorithm works. You decide to follow the extension of the word BBC in the following protein sequence: ACCCBBCBA.

e. What would be a reasonable extension (without gaps) of the word BBC? Explain why you stopped extending when you did on the left and right side. For this question, you are not limited to the method of extension described in class—the desired answer is much simpler. (3 points)

Finally, you want to convince your local representative that the BLAST algorithm is statistically rigorous.

f. Explain in words the purpose of the E value. Limit your response to no more than three sentences. (5 points)

Question 2: Probability Distribution Functions (15 points)

Back on planet earth, you have been trying to master darts. Standing at the dartboard, you decide one night to examine how well your dart throwing by measuring the distance from each dart to the bulls-eye. After 100 dart throws, you construct a histogram and find that it roughly fits the following functional form, where n is the number of darts and d is the distance in inches from the center of the dartboard.

$$n(d) = 5e^{-2d}$$

For the following questions, remember that there is no such thing as a negative distance.

a. As written, this function is not a probability density function. What multiplicative factor will scale *n(d)* such that it is normalized? In other words, find the factor A such that the product of A times the complete integral of *n(d)* is equal to one. (5 points)

b. With the understanding that it is undefined for negative distances, what is the functional form of the cumulative density function? (3 points)

c. What is the probability that d is less than or equal to 4 inches? (2 points)

d. What is the probability that d is greater than four inches? (2 points)

e. What is the probability that d is less than six inches? (2 points)

f. What is the probability that d is between four and six inches? (1 point)

Question 3: Expect Values, Scores, and Bit Scores (15 points)

a. Starting with equations 4.5 and 4.6 in your textbook, derive equation 4.7. (2 points)

b. Suppose you have performed a local alignment between two sequences, one of length 26 and the other of length 128. Given that the raw score for this alignment is 118, and that the K and λ parameters are 0.15 and 0.25, respectively, determine the E and P values for this alignment. In your calculation, do not correct for the effective search space (in the book's, terms, use L=0). Is this alignment statistically significant? (5 points)

c. Convert the score from part (b) into a bit score. (3 points)

d. In one sentence, explain the utility of a bit score versus a raw score. (5 points)

Question 4: Using BLAST (20 points)

On the assignment web page, you will find three protein/DNA sequences that you will use for this assignment, labeled A, B, and C. The sequences are stored in what is called "FASTA" format, a format that is easily portable across operating systems and very simple. A FASTA file is simply a file whose first line optionally begins with a greater than sign ">" followed by a description of the sequence. The rest of the lines contain the sequence itself, and blank lines are ignored. For most sequence web servers, including those at the NCBI, uploading a FASTA file directly is fine—you do not need to remove the comment line. The web server will then worry about handing the line breaks and merging your sequence into one string once it receives your data.

While not particularly relevant for this assignment, most NCBI searches also allow you to view a sequence in FASTA format that can then be cut and pasted in to a file. This provides a useful way to save sequences for later use, and it's more convenient than trying to cut and paste the result of a GenPept record. Look for this option as a drop down menu near the top of Entrez Gene and Protein entries.

All of the sequences describe some human protein (RefSeq) and have been modified in such a way as to make them unidentifiable to even a close inspection. You will need to use the BLAST set of programs to identify the accession number from which the sequences were derived. Given the information below, list the values you chose for the following options, and give a brief one-sentence explanation why you chose the options you did: program type (BLASTP, BLASTN, etc.), Entrez filters, scoring matrix.

a. Sequence A is a protein sequence that is expected to have 80% sequence identity from the original protein sequence. (5 points)

b. Sequence B is a protein sequence that is expected to have 30% sequence identity from the original protein sequence. (5 points)

c. Sequence C is a DNA sequence that is expected to have very little sequence identity to the original DNA sequence, but 40% sequence identity to the translated protein sequence. (10 points)

Question 5: PHI-BLAST (10 points)

The following questions require you to construct patterns that you can enter into PHI-BLAST. You should not have to use the web server.

a. The sequence Glycine – {Any Residue} – {Any Residue} – Glycine has been indicated in some helix dimerization motifs in membrane proteins. Construct a PHI-BLAST pattern for this sequence. (5 points)

b. Construct a PHI-BLAST pattern to represent 3 turns of an amphipathic α-helix. An amphipathic helix is a helix where one side is hydrophobic and the rest is hydrophilic. Specifically, if a helix has approximately 4 residues, an amphipathic helix would have a hydrophobic residue at every fourth position. For the purposes of this problem, assume that the hydrophobic residues are leucine, isoleucine, valine, phenylalanine, and alanine. Also assume that any residue can be in the other three positions. (5 points)

Question 6: PSI-BLAST (15 points)

In this problem you will observe the increased sensitivity that PSI-BLAST provides over standard BLASTP. First, use standard BLAST to search the non-redundant database for the human retinol binding protein 4 sequence in mice (organism *Mus musculus*). Then, perform a PSI-BLAST of RPB4 over all organisms in the non-redundant database. For the first BLAST, use all the default parameters, but for PSI-BLAST, use an expect value cutoff of 0.0001). After 10 iterations of PSI-BLAST, you should be able to re-format your search to give you the position-specific scoring matrix (PSSM) rather than the alignment results.

Cut and paste this PSSM into the PSSM field of the standard BLAST search, and re-search the mouse genes with the PSSM in BLASTP (again, using all standard options except for the PSSM, which replaces the query sequence you used originally.)

a. How many final alignments go into making your PSSM after ten iterations? Briefly explain why having a PSSM derived from many alignments can result in a more sensitive BLAST search? (10 points)

b. Describe some of the differences between using the original human RPB4 sequence and the refine PSSM to search for lipocalins in mice. How many significant matches were identified using standard BLAST (use $E < 0.05$)? How many significant matches were identified using the PSSM matrix? (5 points)