CSCI1820 HW3

Due: Tuesday, Feb 25 11:59pm

Choose to complete either the Biology **or** Computational problem 0.

This homework is scored out of 100 (or more with extra-credit). For all problems, you will only receive full credit if you document how you obtained your solution; in most cases, commented code is the best way to do this. In some cases a brief description is adequate. The other common cause of points subtracted is failing to show the data or mathematics which justify your statements. Please submit all source code, compiled binaries, and a README describing how to run the code for problems that require an algorithm implementation.

Problem 0 BIOLOGY: Mathematica alignments (30)

We have compiled some of the DNA and amino acid sequences of the BRCA1 homologs from Homework 1. You can now find these sequences on the homework resources.

You will also find the code for two Mathematica programs on the website (and a tutorial). One program is for DNA global alignment and the other is for protein global alignment. Mathematica can be run within the department with the command `/local/bin/mathematica`, or downloaded from CIS `cis.brown.edu`. In order to run the Mathematica code, either Evaluate Notebook or select all (Cntl-A) and click Evaluation→Evaluate Cells. Attach the `xls` and your modified code to the assignment.

After installing Mathematica, open up these two files. Anything between `(*` and `*)` is a comment. You will notice that one of the first comments says to enter two sequences. You must copy and paste DNA/protein sequences from the files provided on the website between the quotation marks in order for the program to align them. Note that the sequences you enter cannot have any spaces or line breaks.

If you scan through the code, you will notice various sections for finding the max, creating tables with zeros, defining the scoring matrix, converting letters to numbers, running the DP algorithm, and performing the traceback. The scoring matrix is of particular importance. Note that in the protein code, you can set the value of indels and in the DNA code, you can set the value of matches, mismatches, and indels. You might try playing around with various scoring matrices and seeing the difference in the result.

- To complete this problem you will fill out the Excel file located on the website with a comparison of one gene's DNA and amino acid sequence to that of the other homologs. You can choose whichever gene you like to begin. You will need to run the Mathematica code with the appropriate sequences and record the score, number of matches, number of mismatches, and number of indels. Also record the same for the reverse compliment DNA (Part II). Let the score be for match = 5, mismatch = -3, indel = -1.

- Part II: One of the features of BLAST is that it also searches the reverse compliment of one of the two search strings. The reverse compliment of a DNA sequence is reversed and switched, letter for letter, with the nucleotide bases with an equal number of hydrogen bonds (ie. $ATCG \rightarrow CGAT$). One method of implementation is to reverse the string and then replace characters. However, to demonstrate that you understand the dynamic programming algorithm, modify the Mathematica code to search the reverse compliment strand by making changes to the scoring matrix and array indices at key places. As a consequence, the algorithm should output the correct score number for aligning a sequence against the compliment of the other (without using string reversal, etc.). *Extra-credit:* Modify the code to perform a forward and reverse alignment, implement the traceback properly in the reverse case, and display both results neatly.

Problem 0 COMPUTATIONAL: Local alignment with inversions (30)

It is easily seen that we can not only align two DNA sequences, but that we can also align a sequence to the reverse compliment of the other. Biologically, the reverse compliment is functionally equivalent to the forward sequence due to complementarity of DNA. In some genomes, we may have optimal alignments in which segments alternate between the forward and reverse state. In other words, an alignment of two species may agree in the forward direction until some coordinate at which the orientation switches, and then switches back at some later point (etc.) This is especially the case in viral genomes where foreign DNA vectors are inserted with random orientation.

Please read the Waterman paper for a formal statement of the local alignments with inversion problem and the algorithm, and implement it. The presentation of the algorithm with inversions begins on page 525. An inversion of DNA is the reverse compliment, so, the sequence is reversed and the bases are complimented ($A \rightarrow T$, $T \rightarrow A$, $C \rightarrow G$, and $G \rightarrow C$). The paper describes the algorithm with linear gap function (affine gap - gap open + gap extension) but we will accept either affine gap or the standard gap penalty implementations. Similar to gaps, a penalty of γ is included for the opening of an inversion. **Write a program that performs local alignment with inversions. Your algorithm should take as input two fasta files and output their alignment with inversions.** You need only implement the recursive algorithm given on page 526 and copied below. Also, **give a brief and informal explanation of what the U, V, W, and Z matrices are storing.** Implementations in Mathematica will receive a pastiche pie award!

Problem 1 Contamination and Dynamic Programming (30)

For this problem, you will need to download a multiple contaminated sequence file and a multiple sequence file of cloning vectors.

A *contaminated sequence* is one that does not faithfully represent the genetic information from the biological source organism because it contains one or more sequence segments of foreign origin. A common source of contamination is when a sequence of interest is inserted within another sequence, called a *cloning vector*, which allows biologists to easily clone, propagate, and manipulate it. Failure to remove the vector sequence is often the source of contamination.

For computational students, determine *which* of the sequences is contaminated with *which* of the vectors. You must code your *own* dynamic programming approach as the solution, and attach it with your answers. You may assume that a consecutive substring of 10 "match" bases from the library that is found in a sequence means that the sequence has been contaminated. Due to sequencing errors, one mismatch in an otherwise-perfect 11-mer alignment is also evidence of contamination. (Hint: Score with a threshold for 11 bases, 10 of which are identities and 1 of which is mismatched.)

For biology students, please describe how you would solve this problem including as much of the following as possible:

1. description of the method
2. pseudo-code of the algorithm
3. time-complexity analysis
4. any scoring matrices used

Problem 2 The Car and Goat Revisited (30)

Consider that the car and goat problem discussed in class. We have uploaded the solution to this problem on the homework page of the website. It was crucially important to carefully define the mathematical assumptions before solving the problem. This problem may have different probabilistic properties if the quantities of doors, cars, and goats are varied. Mathematically determine whether you should switch doors, and explain why:

1. Case: 5 doors, 3 goats, 2 cars (10 pts)
2. Write a general proof for $(m + n)$ doors, m goats, n cars (15 pts). Be as mathematically rigorous as possible.

Problem 3 The Pharma Drug Problem (10)

Consider three Pharmas (pharmaceutical companies) A, B, C. They each sent a new drug to FDA for approval, and this is public knowledge. The news from the FDA is that two of them will not be approved.

Pharma A asks the FDA to tell them the identity of a Pharma other than A whose drug is not going to be approved.

FDA refuses and their PR representative explains that "the probabilty for A not to get its drug approved is $\frac{2}{3}$. So your probablity to get approved is $\frac{1}{3}$. If we tell you that Pharma B's drug, say, is not approved, then you would be one of the two Pharmas whose drug fate is unknown, and your probability of not getting approved would consequently decrease to $\frac{1}{2}$. So your probability to get approved will be now much bigger: $\frac{1}{2}$. Since we cannot be unfair to the other Pharmas, we cannot reveal this information."

Is the FDA PR Rep correct in her reasoning?