

## CS195L HW3

*Due: Thursday, Feb 24 2:30pm*

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.

### Problem Reading

- Introduction to the biology of DNA sequencing. On the website we link you to three tutorials on DNA sequencing. Included is a general overview of the 'old school' methods of DNA sequencing, and also a short video tutorial of one of the newer methods of 'next generation' sequencing.
- Doolittle, Russell. Similar Amino Acid Sequences: Chance or Common Ancestry? (pdf on site)
- Dijkstra, Edsger. A note on two problems in connexion with graphs. The first problem described is minimum spanning tree (MST); the second is the famous Dijkstra's algorithm. Note that the solution to the second – shortest path – problem doesn't necessarily follow the branches of the MST, although the solutions look similar at first.

### Problem Mathematica (20)

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. One program is for DNA global alignment and the other is for protein global alignment.

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.

Mathematica can be run within the department with the command `/cfarm/bin/mathematica`, or downloaded from CIS [cis.brown.edu](http://cis.brown.edu). In order to run the Mathematica code, select all (Cntl-A) and click Evaluation→Evaluate Cells. Attach the `xls` and your modified code to the assignment.

## Problem Statistics (20)

For this problem, you will be collecting statistics about the BRCA1 homolog DNA sequences in the homework resources.

Determine each of the following for the entire length of **one** of the BRCA1 homolog sequences linked above:

- What percentage of the nucleotides are As? Cs? Gs? Ts?
- What percentage of the gene sequence consists of two consecutive As? Cs? Gs? Ts?  
(Note that such substrings may overlap, so the sequence "CAAAG" has two substrings of "AA.")
- What percentage of the gene sequence consists of three consecutive As? Cs? Gs? Ts?
- What is the longest inexact repeat, allowing for at most one mismatch?  
(You may find it useful to reuse code you wrote for Problem 1 of Homework 2.)
- What is the longest inexact repeat, allowing for at most two mismatches?

The percentage is defined as  $100 * \frac{\text{base pairs covered by problem description}}{\text{length of sequence}}$ .

## Problem 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.

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.)

## Problem The Car and Goat Revisited (30)

Consider that the car and goat problem discussed in class 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.